Statistics II: Statistical Inference

Prerequisites: Statistics I; Probability I and II

References:
1. Casella, G. and Berger, R. *Statistical Inference*
2. Bickel, D. and Doksum, K. *Mathematical Statistics*
3. Hogg and Craig. *Introduction to Mathematical Statistics*

Grading (Tentative): 40 marks for assignments; 20 marks for class test; 40 marks for final exam

Lectures: Online lectures will be on Zoom; Time: Monday, Wednesday, Friday, 3:15 – 4:15 pm

Lecture Notes: Lecture notes will be posted on Moodle. Attempt will be made to post recording of lectures on Moodle too

Assignments: Assignments will also be posted on Moodle. Answers to these must be submitted by uploading to Moodle.

Contact e-mail: *mohan.delampady@gmail.com*

What is statistical inference? Why is a probability model needed? Inference based on modeling data or observations using probability models is of interest. Consider this example.

Example 1. In fisheries and ecology, one uses capture-recapture methods to estimate the size of a population. Let N = total size; this is unknown. First N_1 of the individuals in this population are caught, tagged and then released. In the next step n out of N are caught again. If n_1 out of n have tags on them, what is an estimate for N? Let X = number of tagged fish out of n. Assuming that the tagged individuals had mixed well with the others in the entire population before sampling and that all individuals in the population have the same chance of being caught, we can write down a probability model for X:

$$P(X = x|N) = \frac{\binom{N_1}{x}\binom{N-N_1}{n-n_1}}{\binom{N}{n}}, x = 0, 1, \ldots n; x \leq N_1; n - x \leq N - N_1.$$

Thus, a statistical model is an approximate but simple theoretical framework to work with.

Statistical inference is mathematical but not mathematics itself, because statistics is inductive reasoning, not deductive as in mathematics. In mathematics, one states certain axioms and then proves (or deduces) certain conclusions, such as theorems. In statistics, one uses observations or data, which are instances of events or occurences. From these one generalizes to other situations (which is induction). The data may be compatible with many models which are just mathematical structures for the generation of data. So, this is a one-to-many problem. Consider an experiment where one tosses a coin 10 times and observes it coming up heads 8 times. What is $\theta = P($ coin comes up heads on any toss $)$? Any $0 < \theta < 1$ can produce the outcome $X = 8$ in $n = 10$ tosses.

It is important to have a theory which can show that statistical inference is valid, consistent and leads to correct conclusions under appropriate assumptions. This is provided by the mathematical theory of probability.

Since statistical inference involves picking a model from a collection of models, it is necessary to study the properties of such classes of models. Optimality results of statistical procedures will be established later for some such classes.

First a few definitions needed in the discussion.

As mentioned previously, a probability model is a structure assumed to model

the realization of random observables or data. Simple models involve a few unknown quantities which are used to index or label the distributions in the family of models. These labels are called *parameters*. The set of all parameter values in a family is called the parameter space. Usually, parameters are associated with important features of the distribution, such as mean and variance.

If $\mathcal{P}$ denotes the family of probability models under consideration, then $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, θ is the unknown parameter and Θ is the parameter space.

Example 2. X = number of times a coin comes up heads when tossed 10 times, $\theta = P($ coin comes up heads on any toss $)$. Then $X \sim \text{Binomial}(10, \theta)$. So,
$\mathcal{P} = \{$ all Binomial(10, θ), $0 < \theta < 1\}$. Note $\Theta = (0, 1) \subset \mathcal{R}^1$.

Example 3. Suppose X denotes the length of time required for a randomly chosen person to recover from common cold, and we model it as $X \sim N(\mu, \sigma^2)$. Then $\mathcal{P} = \{$ all $N(\mu, \sigma^2), -\infty < \mu < \infty, \sigma^2 > 0.\}$. Here $\Theta = \{(\mu, \sigma^2), \mu \in \mathcal{R}^1, \sigma^2 \in \mathcal{R}^+\} = \mathcal{R}^1 \times \mathcal{R}^+ \subset \mathcal{R}^2$. Note that recovery time cannot be negative, but this approximation is reasonable if almost all the probability lies on the positive region.

The idea of a parameter is for it to specify the distribution.

Identifiability. For any θ_1 and θ_2 in Θ, whenever $\theta_1 \neq \theta_2$, we must have $P_{\theta_1} \neq P_{\theta_2}$.

Example 4. Suppose N is the number of tigers in a reserve forest, and we assume $N|\lambda \sim \text{Poisson}(\lambda)$. Let S be the number of tigers sighted by a team of investigators in a study here. Since this involves detection, probability of which is usually less than 1, we can assume, $S|(N = n), p \sim \text{Binomial}(n, p)$ and therefore, $S|\lambda, p \sim \text{Poisson}(\lambda p)$. (Show this as an exercise.) If $S_1, S_2, \ldots, S_k$ are i.i.d Poisson(λp) can we make inferences about both λ and p? The model for S, $\{\text{Poisson}(\lambda p), \lambda > 0, 0 < p < 1\}$, is not identifiable. Here $\theta = (\lambda, p)$ and $P_\theta = \text{Poisson}(\lambda p)$. Take $\theta_1 = (10, 0.4)$ and $\theta_2 = (20, 0.2)$. Then $\theta_1 \neq \theta_2$, but $P_{\theta_1} = P_{\theta_2}$.

Symmetric distribution. Z is distributed symmetrically about 0 if Z and $-Z$ have the same distribution. If Z has density f_Z, then note, in the presence of symmetry, $f_Z(z) = f_Z(-z)$ for all z. X is symmetric about μ if $X - \mu$ is symmetric about 0, or $X - \mu$ and $-(X - \mu)$ have the same distribution. If X has density f_X, then we need $f_X(\mu + x) = f_X(\mu - x)$ for all x. ($f_X(\mu + x) = f_Z(\mu + x - \mu) = f_Z(x) = f_Z(-x)$; $f_X(\mu - x) = f_Z(\mu - x - \mu) = f_Z(-x) = f_Z(x)$)

A statistical model can arise in two different ways, and hence may have two different interpretations.
(i) Measurement error model: Suppose we want to determine the length μ of a table by measuring it using a tape measure. Then any measurement X can be represented as $X = \mu + \epsilon$, where ϵ stands for the deviation from the true length μ due to measurement error. If $\epsilon \sim N(0, \sigma^2)$, then $X \sim N(\mu, \sigma^2)$. σ provides a measure of how large a typical deviation from μ can be.
(ii) Sampling from a population: Consider collecting a random sample (independent and identically distributed or i.i.d.) of observations from a population to determine certain features such as height, weight or family income. Now there is variation within the population. Therefore, if we model an observation X as $X \sim N(\mu, \sigma^2)$, then μ represents the population average and σ measures the deviation or spread of the population around μ.

The statistical inferential procedure will be the same, irrespective of how the model is arrived at. i.e., it is for μ and σ^2 of $N(\mu, \sigma^2)$ whether the observation came from (i) or (ii).

Location-Scale Families Consider $U \sim U(-1, 1)$ with density $f_U(u) = \frac{1}{2} I_{(-1,1)}(u)$, and let $X = \mu + U$. Then $X \sim U(\mu - 1, \mu + 1)$ with density $f_X(x) = \frac{1}{2} I_{(\mu-1,\mu+1)}(x) = \frac{1}{2} I_{(-1,1)}(x - \mu) = f_Z(x - \mu)$. In other words, the location of X is a translation by μ of the location of U. The family of distributions for X indexed by μ is called a location family with location parameter μ. Note that μ is location for X if $X - \mu$ has a distribution which is free of μ.

Similarly, if $Z \sim N(0, 1)$ with density $f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$, then $X = \sigma Z$, $\sigma > 0$, then $X \sim N(0, \sigma^2)$ with density $f_X(x) =$
$\frac{1}{\sqrt{2\pi}\sigma} \exp(-(x - \mu)^2/(2\sigma^2)) = \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} \exp(-((x - \mu)/\sigma)^2/2) = \frac{1}{\sigma} f_Z(\frac{x-\mu}{\sigma})$.
In this case, X is scaled by σ, and the family of distributions for X indexed by σ is called a scale family with scale parameter σ. It is important to note that σ is scale for X if X/σ has a distribution which is free of σ. Combining location and scale gives the location-scale family.

Definition Let X be a real-valued random variable, with density

$$f(x|\mu,\sigma) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right),$$

where g is also a density function, $-\infty < \mu < \infty$, $\sigma > 0$. The parameters μ and σ are called location and scale parameters.

With X as above, $Z = (X-\mu)/\sigma$ has density g. The normal $N(\mu, \sigma^2)$ is a location-scale family with Z being the standard normal, $N(0,1)$. Exponential is a scale family with $\mu = 0$, $\sigma = \theta$. We can make it a location-scale family if we set

$$f(x|\mu,\sigma) \quad = \quad \begin{cases} \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right) & \text{for } x > \mu; \\ 0 & \text{otherwise.} \end{cases}$$

Bernoulli, binomial, and Poisson are not location-scale families.

Example. Let X have uniform distribution over (θ_1, θ_2) so that

$$f(x|\theta_1,\theta_2) \quad = \quad \begin{cases} \frac{1}{\theta_2-\theta_1} & \text{if } \theta_1 < x < \theta_2; \\ 0 & \text{otherwise.} \end{cases}$$

This is also a location-scale family, with a reparameterization.

Example. The Cauchy distribution specified by the density

$$f(x|\mu,\sigma) = \frac{1}{\pi}\frac{\sigma}{\sigma^2 + (x-\mu)^2}, \quad -\infty < x < \infty$$

is a location-scale family. It has several interesting properties. As $|x| \to \infty$, it tends to zero but at a much slower rate than the normal.

One can verify that $E(|X|^r) = \infty$ for $r = 1, 2, \ldots$ under any μ, σ. So Cauchy has no finite moment. However, Figure 1.1 shows remarkable similarity between the normal and Cauchy, except near the tails. The Cauchy density is much flatter at the tails than the normal, which means x's that deviate quite a bit from μ will appear in data from time to time. Such deviations from μ would be unusual under a normal model and so may be treated as outliers by a data analyst. It provides an important counter-example to the *law of large numbers* or *central limit theorem* when one has infinite moments. It also plays an important role in robustness studies.

Result. Any location-scale family of models is closed under location-scale transformations.

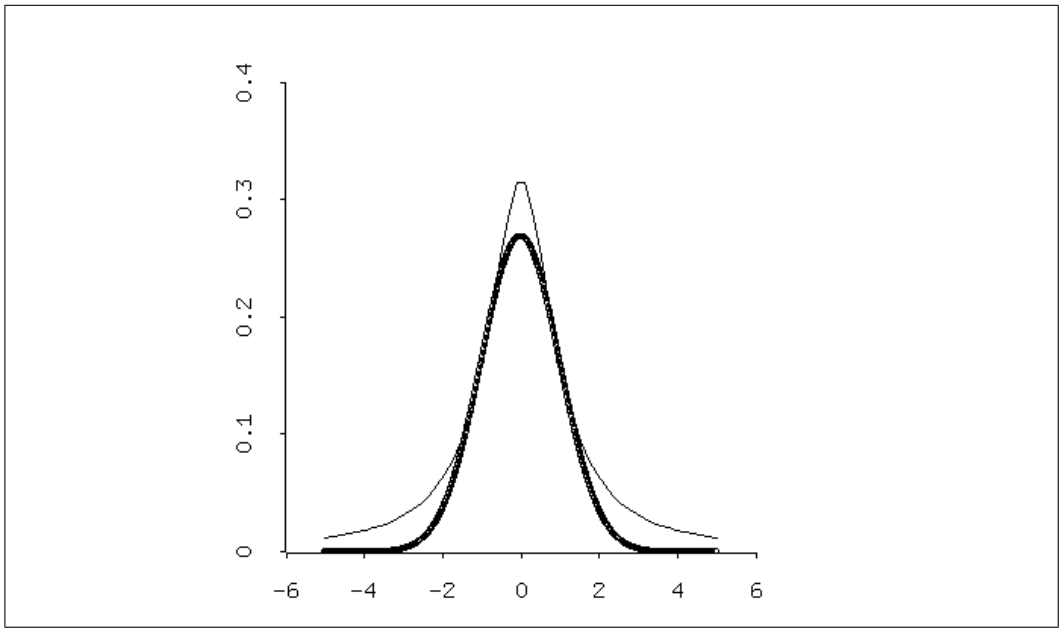


Figure 1: Densities of Cauchy(0, 1) and normal(0, 2.19).

Proof. Let $\mathcal{F}$ be a location-scale family. Then,

$$\mathcal{F} = \left\{ f(x|\mu,\sigma) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right), -\infty < \mu < \infty, \sigma > 0 \right\},$$

where g is a density function. Let $Y = aX + b$. Take $a > 0$ for convenience. Then $X = (Y-b)/a$, so $dx = dy/a$, and hence

$$f_Y(y) = \frac{1}{a\sigma} g\left(\frac{\frac{y-b}{a} - \mu}{\sigma}\right) = \frac{1}{a\sigma} g\left(\frac{y-(a\mu+b)}{a\sigma}\right).$$

Therefore, $f_Y \in \mathcal{F}$.

Location-scale family is a special case of a Group family, a family of models which is closed under a group of transformations.

In our discussion we will confine ourselves to *parametric models*, in which case the parameter space is a nice subset of $\mathcal{R}^k$ for some k. Nonparametric models deal with much larger classes of models, such as all symmetric distributions or all symmetric unimodal distributions. Nonparametric methods (such as histograms) are quite different from what we discuss here. Even among parametric models, we restrict ourselves to models which satisfy the regularity conditions.

Regular models. Either (i) all P_θ are continuous with density $f(.|\theta)$, or (ii) all P_θ are discrete with mass function $p(x|\theta)$ and there exists a countable set $\mathcal{S} = \{x_1, x_2, \ldots\}$ independent of θ such that $\sum_{i=1}^{\infty} p(x_i|\theta) = 1$.

For example, the following models are not regular under our definition.

Example. Let X be the weight of a randomly chosen product from a population, and assume $X \sim N(\theta, 10^2)$. However the weighing device cannot show weights above a certain level c, so it fixes (censors) such weights at c. Then the observed weight, Y has the distribution, $Y = \begin{cases} X & \text{if } X < c; \\ c & \text{if } X \geq c. \end{cases}$. Then Y has a continuous part as well as a point mass (at c).

Example. Let X be discrete with pmf $p(x|\theta) = \begin{cases} 1/2 & \text{if } x = \theta; \\ 1/2 & \text{if } x = \theta + 1, \end{cases}$ where $\theta \in \mathcal{R}$. Here $\mathcal{S}$ is not countable.

Sufficient Statistics

Statistical inference is our objective, i.e., making inferences about unknown parameters in the model. The first step in this direction is data reduction or compression – condensing all the useful information and removing all the irrelevant pieces. This will allow modeling only the informative parts of the data. For example, suppose the mean yield of fruit in a farm is of interest, and a random sample of trees is investigated for this purpose. Note that the sample data may look different depending on the order in which one records the yields from the trees in the sample, but this order of observations is not relevant for inferential purposes.

Let $\mathbf{X}$ denote sample data or the list of observations. Then any real or vector-valued function $T(\mathbf{X})$ is called a statistic. Examples are:
$T(X_1, \ldots, X_n) = \bar{X}$,
$T(X_1, \ldots, X_n) = \sum_{i=1}^{n} (X_i - \bar{X})^2$,
$T(X_1, \ldots, X_n) = (\bar{X}, \sum_{i=1}^{n} (X_i - \bar{X})^2)$.
Intuitively, a statistic $T(\mathbf{X})$ is sufficient if it contains all the useful information about the quantities of interest. We will suppress the boldface for vectors unless there is ambiguity.

Definition. A statistic $T(X)$ is called sufficient for a parameter θ, or sufficient for a family of distributions P_θ indexed by θ if the conditional distribution of X given $T(X) = t$ does not involve θ. i.e.,
$P_\theta(a < X \leq b | T(X) = t)$ is independent of θ for all a, b if X is a random variable.

Note. When $T(\mathbf{X})$ is sufficient, if you know its value, you don't care what $\mathbf{X}$ is anymore.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson(λ).
Claim: $S = \sum_{i=1}^n X_i$ is sufficient for λ.
Note that

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\lambda) = \prod_{i=1}^n \exp(-\lambda)\frac{\lambda^{x_i}}{x_i!} = \exp(-n\lambda)\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}, \text{ and}$$

$$f_S(s|\lambda) = \exp(-n\lambda)\frac{(n\lambda)^s}{s!}.$$

Therefore,

$$\begin{aligned} f_{(X_1,\ldots,X_n)|S=s}(x_1, \ldots, x_n|\lambda) &= \frac{P_\lambda(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n, S = s)}{P_\lambda(S = s)} \\ &= \frac{P_\lambda(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = s - \sum_{i=1}^{n-1} x_i)}{P_\lambda(S = s)} \quad \text{if } \textstyle\sum_{i=1}^n x_i = s \\ &= \frac{\exp(-n\lambda)\ \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}}{\exp(-n\lambda)\ \frac{(n\lambda)^s}{s!}} \quad \text{if } \textstyle\sum_{i=1}^n x_i = s \\ &= \frac{s!}{\prod_{i=1}^n x_i!}\left(\frac{1}{n}\right)^s \quad \text{if } \textstyle\sum_{i=1}^n x_i = s\ , \end{aligned}$$

which is free of λ. In fact, the conditional distribution above is Multinomial($s; \frac{1}{n}, \ldots, \frac{1}{n}$).

One needs to guess T for using the above definition. Instead, there is the following useful and important result.

Factorization Theorem (Neyman-Fisher). Let $f(\mathbf{x}|\theta)$ be the density of $\mathbf{X}$ under the probability model $P_\theta, \theta \in \Theta$. Then, if the model is regular, a statistic $T(\mathbf{X})$ is sufficient for θ iff there exists a function $g(t, \theta)$ and a function $h(\mathbf{x})$ such that

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x}).$$

i.e., one is able to factor f into two parts, one involving θ and data through T, and the other involving data only. If $\mathbf{x} \in \mathcal{R}^n$, then we have
$T : \mathcal{R}^n \to I \subseteq \mathcal{R}^k, k \leq n$,
$g : I \times \Theta \to \mathcal{R}^+$,
$h : \mathcal{R}^n \to \mathcal{R}^+$. g and h are not unique.

Proof. For the discrete case only. The continuous case is similar, but a rigorous proof requires measure theoretical arguments.

Let $\mathcal{S} = \{x_1, x_2, \ldots\}$ be the (countable) sample space, the set of all possible values of $\mathbf{X}$. Let $t_i = T(x_i)$. Then $T = T(X)$ is discrete and
$\sum_{i=1}^{\infty} f_T(t_i|\theta) = P_\theta(T = t_i, 1 \leq i < \infty) = 1$ for all θ.
if part: There exist g and h such that $f(\mathbf{x}|\theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$. Need to show that T is sufficient. i.e., show that $P_\theta[X = x_j|T(X) = t_i]$ does not involve θ for each i and j. Since $P_\theta[X = x_j|T(X) = t_i]$ is defined when $P_\theta[T = t_i] > 0$ only, it is enough to show that $P_\theta[X = x_j|T(X) = t_i]$ is independent of θ when $\theta \in \Omega_i = \{\theta : P_\theta[T = t_i] > 0\}$, $i = 1, 2, \ldots$ Now, note

$$P_\theta[T = t_i] = \sum_{\{x:T(x)=t_i\}} f(x|\theta) = g(t_i, \theta) \sum_{\{x:T(x)=t_i\}} h(x).$$

If $\theta \in \Omega_i$, then

$$\begin{aligned} P_\theta[X = x_j|T(X) = t_i] &= \frac{P_\theta[X = x_j, T(X) = t_i]}{P_\theta[T = t_i]} \\ &= \begin{cases} \frac{f(x_j|\theta)}{P_\theta[T=t_i]} & \text{if } T(x_j) = t_i; \\ 0 & \text{if } T(x_j) \neq t_i. \end{cases} \end{aligned}$$

Note that $P_\theta[X = x_j, T(X) = t_i] = 0$ if $T(x_j) \neq t_i$. Using the fact that $f(\mathbf{x}|\theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$, whenever $T(x_j) = t_i$, we get

$$\begin{aligned} \frac{f(x_j|\theta)}{P_\theta[T = t_i]} &= \frac{g(t_i, \theta)h(x_j)}{g(t_i, \theta)\sum_{\{x:T(x)=t_i\}} h(x)} \\ &= \frac{h(x_j)}{\sum_{\{x:T(x)=t_i\}} h(x)}, \end{aligned}$$

which is independent of θ. Thus $T = T(X)$ is sufficient for θ.

only if: Now we have that $T = T(X)$ is sufficient for θ. We want to find functions, g and h such that $f(\mathbf{x}|\theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$. Define $g(t_i, \theta) = P_\theta[T = t_i]$ and $h(x) = P_\theta[X = x|T = T(x)]$. Then $h(x)$ is independent of θ by definition (of sufficiency). Also,

$$\begin{aligned} f(x|\theta) &= P_\theta[X = x] = P_\theta[X = x, T = T(x)] \\ &= P_\theta[X = x|T = T(x)]P_\theta[T = T(x)] \\ &= h(x)g(T(x), \theta), \end{aligned}$$

noting that $P_\theta[T = T(x)]$ is a function of $T(x)$ and θ only.

Example. Suppose $X_1, \ldots, X_n$ is a random sample (i.i.d.) from Poisson(λ), $\lambda > 0$. Is $T(X) = \sum_{i=1}^n X_i$ sufficient for λ? Note that

$$\begin{aligned} f(x_1, \ldots, x_n|\lambda) &= \prod_{i=1}^n \exp(-\lambda)\frac{\lambda^{x_i}}{x_i!} = \exp(-n\lambda)\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \\ &= \left(\exp(-n\lambda)\lambda^{\sum_{i=1}^n x_i}\right)\frac{1}{\prod_{i=1}^n x_i!}. \end{aligned}$$

Take $g(t, \lambda) = \exp(-n\lambda)\lambda^t$ and $h(x_1, \ldots, x_n) = \frac{1}{\prod_{i=1}^n x_i!}$ to satisfy the factorization theorem.

Example. Let $X_1, \ldots, X_n$ be a random sample (i.i.d.) from $N(\mu, \sigma^2)$. What are jointly sufficient for μ and σ^2? We have,

$$\begin{aligned} f(x_1, \ldots, x_n|\mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\ &= (2\pi)^{-n/2}\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right) \\ &= (2\pi)^{-n/2}\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^n x_i^2 + n\mu^2 - 2\mu\sum_{i=1}^n x_i\right\}\right) \\ &= \sigma^{-n}\exp\left(-\frac{n\mu^2}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i\right)(2\pi)^{-n/2}. \end{aligned}$$

Note that $T(X_1, \ldots, X_n) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is sufficient since we can take $g((t_1, t_2), (\mu, \sigma^2)) = \sigma^{-n}\exp\left(-\frac{n\mu^2}{2\sigma^2} - \frac{1}{2\sigma^2}t_2 + \frac{\mu}{\sigma^2}t_1\right)$ and $h(x) = (2\pi)^{-n/2}$. Also, the map $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2) \to (\bar{X}, \sum_{i=1}^n (X_i - \bar{X})^2)$ is one-one, so $(\bar{X}, S^2)$ is another set of sufficient statistics.

Example. Let $X_1, \ldots, X_n$ be a random sample from the population with density, $f(x|\theta) = \frac{1}{2}\exp(-|x - \theta|)$, $-\infty < \theta < \infty$. Then $f(x_1, \ldots, x_n|\theta) = \frac{1}{2^n}\exp(-\sum_{i=1}^n |x_i - \theta|)$. What is sufficient for θ? Not much reduction of data is possible here, except for noting that the joint density can be written as $f(x_1, \ldots, x_n|\theta) = \frac{1}{2^n}\exp(-\sum_{i=1}^n |x_{(i)} - \theta|)$, where $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ are the ordered observations. Therefore $T(X_1, \ldots, X_n) = (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ is sufficient. Order statistics provide data reduction whenever data is a random sample from a continuous distribution. Some models permit further reduction.

Interpretation of sufficiency. Observing $\mathbf{X} = (X_1, \ldots, X_n)$ is equivalent to observing $T(\mathbf{X})$ as far as information on θ is concerned. Given $T(\mathbf{X})$ (the

distribution of which depends on θ), one can generate $\mathbf{X}' = (X_1', \ldots, X_n')$ from $P(\mathbf{X}|T)$ (which does not require θ, being uninformative). Then the probability distributions of $\mathbf{X}$ and $\mathbf{X}'$ are the same. Note that if two random quantities have the same probability distribution then they contain the same amount of information.

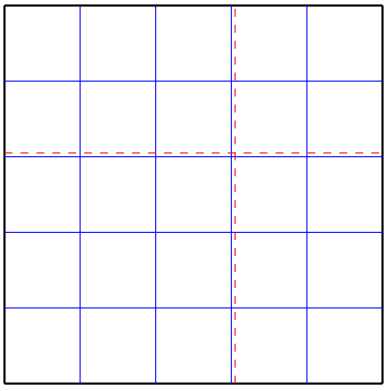
Example. $X_1, \ldots, X_n$ i.i.d Poisson(λ). Then $S = \sum_{i=1}^n X_i$ is sufficient. If $S = s$ is given from Poisson($n\lambda$), then simply generate $(X_1', \ldots, X_n')$ from Multinomial($s; \frac{1}{n}, \ldots, \frac{1}{n}$). It is clear that the joint distribution of $(X_1', \ldots, X_n')$ is the same as that of $(X_1, \ldots, X_n)$, which is i.i.d Poisson(λ).

Definition. Two statistics S_1 and S_2 are said to be equivalent if $S_1(x) = S_1(y)$ iff $S_2(x) = S_2(y)$.

Note that, if S_1 and S_2 are equivalent, then
(i) they give the same partition of the sample space,
(ii) they provide the same reduction,
(iii) they provide the same information.

Example. $S_1(X_1, \ldots, X_n) = \bar{X}$, $S_2(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$. $S_1(x_1, \ldots, x_n) = S_1(y_1, \ldots, y_n)$ iff $\frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{n}\sum_{i=1}^n y_i$ iff $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ iff $S_2(x_1, \ldots, x_n) = S_2(y_1, \ldots, y_n)$.

Data is a realization of the random observable $\mathbf{X}$. It is a point in the sample space. The values of $\mathbf{X}$ form the finest partition of the sample space. Any statistic $T(\mathbf{X})$ also gives a partition of the sample space. For example, $(X_1, \ldots, X_n)$ are points in $\mathcal{R}^n$. $T_1(X_1, \ldots, X_n) = (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ partitions $\mathcal{R}^n$ into sets where the points are permutations of each other. This is coarser than $\mathcal{R}^n$. $T_2(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i$ partitions $\mathcal{R}^n$ into sets where the points have the same average. This partition is coarser than the one provided by T_1 since permutations do not change the average.



In the figure above, the dashed, red partition is coarser than the blue parti-

tion.

Sufficient statistics gives a partition (of the sample space) which retains all the information about the parameters. Therefore, maximum possible reduction of data, or the coarsest possible partition of the sample space is desirable. How does one choose sufficient statistics?

Example. $X_1, \ldots, X_n$ i.i.d $N(0, \sigma^2)$. Then

$$f(x_1, \ldots, x_n|\lambda) = (2\pi)^{-n/2}\sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right).$$

Note from this joint density that
(i) $T_1(X_1, \ldots, X_n) = (X_1, \ldots, X_n)$ is sufficient for σ^2;
(ii) $T_2(X_1, \ldots, X_n) = (X_1^2, \ldots, X_n^2)$ is sufficient for σ^2;
(iii) $T_3(X_1, \ldots, X_n) = (X_1^2 + \cdots + X_m^2, X_{m+1}^2 + \cdots + X_n^2)$ is sufficient for σ^2;
(iv) $T_4(X_1, \ldots, X_n) = (X_1^2 + \cdots + X_n^2)$ is sufficient for σ^2.

Observe that $T_4 = g_1(T_3) = g_1(g_2(T_2)) = g_1(g_2(g_3(T_1)))$. i.e., if T is sufficient and $T = H(U)$, then U is also sufficient. Knowledge of U implies knowledge of T and hence permits reconstruction of the original data. T provides greater reduction or coarser partition unless H is one-one.

Minimal sufficiency. $T = T(X)$ is minimal sufficient if it provides the maximal amount of data reduction. i.e., for any sufficient statistics $U = U(X)$, there exists a function H such that $T = H(U)$.

Minimal sufficiency. $T = T(X)$ is minimal sufficient if it provides the maximal amount of data reduction. i.e., for any sufficient statistics $U = U(X)$, there exists a function H such that $T = H(U)$.

Usually, one can find minimal sufficient statistics applying the factorization theorem and inspection. However, there are some techniques to find them also.

Theorem. $\mathcal{P}$ is a family of probability models with common support and $\mathcal{P}_0 \subset \mathcal{P}$. If T is minimal sufficient for $\mathcal{P}_0$ and sufficient for $\mathcal{P}$, then it is minimal sufficient for $\mathcal{P}$ also.

Proof. Let U be any sufficient statistic for $\mathcal{P}$. Then it is sufficient for $\mathcal{P}_0$. But T is minimal sufficient for $\mathcal{P}_0$. Therefore $T = H(U)$. Now consider $\mathcal{P}$. T is minimal sufficient for $\mathcal{P}$ and for any other sufficient statistic U, $T = H(U)$. Therefore, T is minimal sufficient.

Theorem. Let $f(x|\theta)$ be pmf or pdf of X. Suppose there exists a function $T(x)$ such that, for two sample points x and y the ratio $f(x|\theta)/f(y|\theta)$ is a constant function of θ iff $T(x) = T(y)$. Then $T(X)$ is minimal sufficient for θ.

Example. $X_1, \ldots, X_n$ i.i.d Poisson(λ). Then

$$f(x_1, \ldots, x_n|\lambda) = \exp(-n\lambda)\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}.$$

Therefore,

$$\frac{f(x_1, \ldots, x_n|\lambda)}{f(y_1, \ldots, y_n|\lambda)} = \lambda^{(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i)}\frac{\prod_{i=1}^n y_i!}{\prod_{i=1}^n x_i!}$$

is a constant function of $\lambda > 0$ iff $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. Therefore, $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ is actually minimal sufficient for λ.

Proof (of Theorem). Assume $f(x|\theta) > 0$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$. Suppose there exists T such that $f(x|\theta)/f(y|\theta)$ is a constant function of θ iff $T(x) = T(y)$. We show then that $T(X)$ is minimal sufficient for θ. First, let us show that it is sufficient. The map T is
$T : \mathcal{X} \to \mathcal{T} = \{t : t = T(x) \text{ for some } x \in \mathcal{X}\}$. Let $A_t = \{x \in \mathcal{X} : T(x) = t\}$. Then $\{A_t\}_{t\in\mathcal{T}}$ is a partition of $\mathcal{X}$. For each A_t, fix one element $x_t \in A_t$. For any $x \in \mathcal{X}$, we have that $x \in A_{T(x)}$, and hence $x_{T(x)}$ is the fixed element which belongs to the same partitioning set as x does. $T(x) = T(x_{T(x)})$ since x and $x_{T(x)}$ belong to $A_{T(x)}$. Hence $f(x|\theta)/f(x_{T(x)}|\theta)$ is a constant function of θ. Then $h(x) = f(x|\theta)/f(x_{T(x)}|\theta)$ is independent of θ and $h : \mathcal{X} \to \mathcal{R}^+$.

Define g by $g(t,\theta) = f(x_t|\theta)$ and $g : \mathcal{T} \times \Theta \to \mathcal{R}^+$. Then

$$f(x|\theta) = \frac{f(x|\theta)}{f(x_{T(x)}|\theta)} f(x_{T(x)}|\theta) = h(x)g(T(x),\theta).$$

Therefore, $T(X)$ is sufficient for θ. Let $T'(X)$ be any other sufficient statistics. Then there exist g' and h' such that $f(x|\theta) = g'(T'(x),\theta)h'(x)$. Let x and y be any two sample points such that $T'(x) = T'(y)$. Then

$$\frac{f(x|\theta)}{f(y|\theta)} = \frac{g'(T'(x),\theta)h'(x)}{g'(T'(y),\theta)h'(y)} = \frac{h'(x)}{h'(y)},$$

which is independent of θ. We already have that $T(x) = T(y)$ whenever $f(x|\theta)/f(y|\theta)$ is a constant function of θ. Therefore, $T'(x) = T'(y)$ implies $T(x) = T(y)$. This means that T is coarser than T' or $T(x) = q(T'(x))$ for some function q. Therefore T is minimal sufficient.

Example. Let $X_1, \ldots, X_n$ be i.i.d Exp(θ), $\theta > 0$ with density $f(x|\theta) = \theta \exp(-\theta x)$ for $x > 0$. Then

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n|\theta) = \prod_{i=1}^{n}(\theta \exp(-\theta x_i)) = \theta^n \exp(-\theta \sum_{i=1}^{n} x_i), x_i > 0, \theta > 0.$$

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{\theta^n \exp(-\theta \sum_{i=1}^n x_i)}{\theta^n \exp(-\theta \sum_{i=1}^n y_i)} = \exp(-\theta\{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i\})$$

is a constant function of θ in the interval $\theta > 0$ iff $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. Therefore, $\sum_{i=1}^n X_i$ is minimal sufficient for θ.

Example. Suppose $X - \theta \sim \text{Exp}(\theta)$, $-\infty < \theta < \infty$. Then

$$f(x|\theta) = \begin{cases} \exp(-(x-\theta)) & x > \theta; \\ 0 & \text{otherwise.} \end{cases}$$

Then $\Theta = \mathcal{R}$ and the common support is $\mathcal{X} = \mathcal{R}$ also. Consider a random sample, $X_1, \ldots, X_n$ from this distribution. Then

$$\begin{aligned} f(x_1,\ldots,x_n|\theta) &= \begin{cases} \exp(-(\sum_{i=1}^n x_i - n\theta)) & x_i > \theta \text{ for all } i; \\ 0 & \text{otherwise.} \end{cases} \\ &= \begin{cases} \exp(-(\sum_{i=1}^n x_i - n\theta)) & x_{(1)} > \theta \ ; \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Since $f(x_1, \ldots, x_n|\theta) = \exp(-\sum_{i=1}^n x_i) \exp(n\theta) I(x_{(1)} > \theta)$, $X_{(1)}$ is sufficient. Further,

$$\begin{aligned}
\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} &= \frac{\exp(-\sum_{i=1}^n x_i) \exp(n\theta) I(x_{(1)} > \theta)}{\exp(-\sum_{i=1}^n y_i) \exp(n\theta) I(y_{(1)} > \theta)} \\
&= \begin{cases} \exp(-(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i)) & \text{if } \theta < \min\{x_{(1)}, y_{(1)}\}; \\ 0 & \text{if } x_{(1)} < \theta < y_{(1)}; \\ \infty & \text{if } y_{(1)} < \theta < x_{(1)}; \\ \text{undefined elsewhere.} & \end{cases}
\end{aligned}$$

This is a constant function of θ iff $x_{(1)} = y_{(1)}$. Therefore, $X_{(1)}$ is minimal also.

Exponential Families

Definition. $\{P_\theta, \theta \in \Theta\}$ with density (pdf or pmf) is a single-parameter exponential family if there exist real valued functions $c(\theta)$, $d(\theta)$ on Θ and $T(\mathbf{x})$ and $S(\mathbf{x})$ on $\mathcal{R}^n$ and a set $A \subset \mathcal{R}^n$ such that

$$f(\mathbf{x}|\theta) = \exp\left[c(\theta)T(\mathbf{x}) + d(\theta) + S(\mathbf{x})\right] I_A(\mathbf{x}),$$

where A must not depend on θ.

Example. $X \sim$ Poisson(λ), $\lambda > 0$. Then $f(x|\lambda) = \exp(-\lambda)\frac{\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$. Take $A = \{0, 1, 2, \ldots\}$ and write the density as

$$f(x|\lambda) = \exp\left(x\log(\lambda) - \lambda - \log(x!)\right) I_A(x).$$

Choosing $c(\lambda) = \log(\lambda)$, $d(\lambda) = -\lambda$, $T(x) = x$ and $S(x) = -\log(x!)$ shows that Poisson(λ), $\lambda > 0$ is a single-parameter exponential family of distributions.

Example. $X \sim U(0, \theta)$, $\theta > 0$. Then

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta; \\ 0 & \text{otherwise.} \end{cases}$$

Is $U(0, \theta)$, $\theta > 0$ exponential family? Note that $f(x|\theta) = \exp(-\log(\theta))I_A(x)$, where the support of the density, $A = (0, \theta)$ depends on θ. It is not possible to express the density in the required exponential form on a common support of the entire family, so $U(0, \theta)$, $\theta > 0$ is not exponential family.

Example. $X \sim N(\theta, 1)$. Then

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x-\theta)^2) I_A(x), \quad A = (-\infty, \infty),$$

which can be written as

$$\begin{aligned} f(x|\theta) &= \exp\left(-\frac{1}{2}[x^2 + \theta^2 - 2\theta x] - \log(\sqrt{2\pi})\right) I_A(x) \\ &= \exp\left(\theta x - \frac{\theta^2}{2} - [\frac{x^2}{2} + \log(\sqrt{2\pi})]\right) I_A(x). \end{aligned}$$

Choosing $c(\theta) = \theta$, $d(\theta) = -\theta^2/2$, $T(x) = x$ and $S(x) = -(x^2/2 + \log(\sqrt{2\pi}))$ we can show that $N(\theta, 1)$, $-\infty < \theta < \infty$ is a single-parameter exponential family.

Consider $X_1, \ldots, X_m$ i.i.d P_θ with density $f(x|\theta)$. Suppose $\{P_\theta, \theta \in \Theta\}$ is exponential family. i.e., $f(x|\theta) = \exp(c(\theta)T(x) + d(\theta) + S(x))I_A(x)$. Then

$$\begin{aligned} f_{X_1,\ldots,X_m}(x_1, \ldots, x_m|\theta) &= \prod_{i=1}^{m} \exp(c(\theta)T(x_i) + d(\theta) + S(x_i))I_A(x_i) \\ &= \exp(c(\theta)\sum_{i=1}^{m} T(x_i) + md(\theta) + \sum_{i=1}^{m} S(x_i))I_{A^m}(x_1, \ldots, x_m). \end{aligned}$$

Therefore, $(X_1, \ldots, X_m)$ has distribution belonging to a single-parameter exponential family.

Result. If $\{P_\theta, \theta \in \Theta\}$ is a single-parameter exponential family with density $f(x|\theta) = \exp(c(\theta)T(x) + d(\theta) + S(x))I_A(x)$, then $T(X)$ is sufficient for θ. (Actually minimal sufficient if Θ contains an open interval, as shown later.)

Simply note that

$$f(x|\theta) = \exp(c(\theta)T(x) + d(\theta)) \exp(S(x))I_A(x).$$

Thus we have $g(t, \theta) = \exp(c(\theta)t + d(\theta))$ and $h(x) = \exp(S(x))I_A(x)$. Combining this result with above we get the following.

Corollary. If $X_1, \ldots, X_m$ are i.i.d P_θ with density $f(x|\theta) = \exp(c(\theta)T(x) + d(\theta) + S(x))I_A(x)$, the $\sum_{i=1}^{m} T(X_i)$ is sufficient for θ.

Example. $X \sim$ Bernoulli(θ). Then

$$\begin{aligned} f(x|\theta) &= \theta^x(1-\theta)^{1-x} I_{\{0,1\}}(x) \\ &= \exp\left(x \log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right) I_{\{0,1\}}(x), \end{aligned}$$

so $T(X) = X$ is sufficient for θ. Now consider $X_1, \ldots, X_n$ i.i.d Bernoulli(θ). Then using the Corollary (or otherwise) observe that $\sum_{i=1}^{n} T(X_i) = \sum_{i=1}^{n} X_i$ is sufficient for θ.

Theorem. Let $\{P_\theta, \theta \in \Theta\}$ be a one-parameter exponential family with density $f(x|\theta) = \exp(c(\theta)T(x) + d(\theta) + S(x))I_A(x)$. Suppose that either P_θ is discrete, or $T(X)$ has a continuous distribution. Then the family of distributions $\{Q_\theta\}$ for $T(X)$ is also a one-parameter exponential family, and has density $q(t|\theta) = \exp(c(\theta)t + d(\theta) + S^*(t))I_{A^*}(t)$.

Proof. Discrete case:

$$\begin{aligned} q(t|\theta) &= P_\theta\left(T(X)=t\right) = \sum_{\{x:T(x)=t\}} f(x|\theta) \\ &= \sum_{\{x:T(x)=t\}} \exp(c(\theta)T(x)+d(\theta)+S(x))I_A(x) \\ &= \exp(c(\theta)t+d(\theta)) \left\{ \sum_{\{x\in A:T(x)=t\}} \exp(S(x)) \right\} I_{A^*}(t), \end{aligned}$$

where $A^* = \{t : t = T(x), x \in A\}$. Now define

$$S^*(t) = \begin{cases} \log \sum_{\{x\in A:T(x)=t\}} \exp(S(x)) & \text{if } t \in A^*; \\ 0 & \text{otherwise.} \end{cases}$$

The continuous case is similar.

One-parameter exponential family in natural form.

In the usual form, we have the density as: $f(x|\theta) = \exp(c(\theta)T(x) + d(\theta) + S(x))I_A(x)$. Define $\eta = c(\theta)$ for $\theta \in \Theta$. Let $\Gamma = \{\eta : \eta = c(\theta), \theta \in \Theta\}$. Then we get

$$f^*(x|\eta) = \exp(\eta T(x) + d_0(\eta) + S(x))I_A(x),$$

where $d_0(\eta) = d(c^{-1}(\eta))$ if c is one-one. Otherwise, since we must have

$$\begin{aligned} 1 &= \int_A f^*(x|\eta)\,dx = \int_A \exp(\eta T(x) + d_0(\eta) + S(x))\,dx \\ &= \exp(d_0(\eta)) \int_A \exp(\eta T(x) + S(x))\,dx, \end{aligned}$$

$d_0(\eta) = \log\left(\int_A \exp(\eta T(x) + S(x))\,dx\right)^{-1}$ or $\log\left(\sum_A \exp(\eta T(x) + S(x))\,dx\right)^{-1}$. Let $H = \{\eta : |d_0(\eta)| < \infty\}$. Whenever $\theta \in \Theta$, note that

$$\int_A \exp(c(\theta)T(x) + S(x))\,dx = \int_A \exp(\eta T(x) + S(x))\,dx < \infty$$

since

$$1 = \int_A f(x|\theta\,dx = \exp(d(\theta)) \int_A \exp(c(\theta)T(x) + S(x))\,dx.$$

Therefore, whenever $\theta \in \Theta$, $|d_0(\eta)| < \infty$ and $\eta \in H$.

$$f^*(x|\eta) = \exp(\eta T(x) + d_0(\eta) + S(x))I_A(x),$$

$\eta \in H$ is called exponential family in natural form. H can be shown to be an interval.

Theorem. If X has density

$$f(x|\eta) = \exp(\eta T(x) + d_0(\eta) + S(x))I_A(x),$$

and η is an interior point of H (i.e., $(\eta - \epsilon, \eta + \epsilon) \subset H$), the mgf of $T(X)$ exists and is

$$\psi(s) = E(\exp(sT(X)) = \exp(d_0(\eta) - d_0(s + \eta))$$

for s in some neighbourhood of 0. Also,

$$\begin{aligned} E\left[T(X)\right] &= -\frac{d}{d\eta} d_0(\eta), \\ Var\left[T(X)\right] &= -\frac{d^2}{d\eta^2} d_0(\eta). \end{aligned}$$

Proof. Note that

$$\begin{aligned} E(\exp(sT(X)) &= \int_A \exp(sT(x) + \eta T(x) + d_0(\eta) + S(x))\, dx \\ &= \exp(d_0(\eta)) \int_A \exp((s+\eta)T(x) + S(x))\, dx. \end{aligned}$$

Since η is an interior point of H, $s + \eta \in H$ if s is small enough. Therefore, $\int_A \exp((s+\eta)T(x) + S(x))\, dx < \infty$. But then $f(x|s+\eta)$ is also a density. Thus,

$$\exp(d_0(s+\eta)) \int_A \exp((s+\eta)T(x) + S(x))\, dx = 1$$

or

$$\int_A \exp((s+\eta)T(x) + S(x))\, dx = \exp(-d_0(s+\eta)).$$

Therefore,

$$\begin{aligned} E\left[T(X)\right] &= \frac{d}{ds} E\left[\exp(sT(X)\right]|_{s=0} \\ &= \exp(d_0(\eta) - d_0(s+\eta))\left(-d_0'(s+\eta)\right)|_{s=0} = -d_0'(\eta), \end{aligned}$$

$$\begin{aligned} & E\left[T^2(X)\right] \\ &= \frac{d^2}{ds^2} E\left[\exp(sT(X)\right]|_{s=0} \\ &= \frac{d}{ds}\left\{\frac{d}{ds} E\left[\exp(sT(X)\right]\right\}|_{s=0} \\ &= \frac{d}{ds}\left\{-\exp(d_0(\eta) - d_0(s+\eta)) d_0'(s+\eta)\right\}|_{s=0} \\ &= \left\{\exp(d_0(\eta) - d_0(s+\eta))\left(d_0'(s+\eta)\right)^2 - \exp(d_0(\eta) - d_0(s+\eta)) d_0''(s+\eta)\right\}|_{s=0} \\ &= \left(d_0'(\eta)\right)^2 - d_0''(\eta). \end{aligned}$$

Therefore,

$$\begin{aligned} Var\left[T(X)\right] &= E\left[T^2(X)\right] - E^2\left[T(X)\right] \\ &= -d_0''(\eta) + \left(d_0'(\eta)\right)^2 - \left(-d_0'(\eta)\right)^2 \\ &= -d_0''(\eta). \end{aligned}$$

Example. $X \sim$ Binomial(n, p), $0 < p < 1$, n fixed. Then $E(X) = np$, $Var(X) = np(1-p)$ and $E(\exp(sX)) = [p\exp(s) + (1-p)]^n$. Derive these using the result above. You will note that these formulas are not especially useful for such purposes. They are useful for deriving certain theoretical results instead.

k-parameter exponential family

A family of distributions, $\{P_\theta, \theta \in \Theta\}$ with density $f(\mathbf{x}|\theta)$ is called a k-parameter exponential family if there exist real-valued functions $c_1(\theta), \ldots, c_k(\theta)$ and $d(\theta)$, real-valued functions $T_1(\mathbf{x}), \ldots, T_k(\mathbf{x})$ and $S(\mathbf{x})$ on $\mathcal{R}^n$, and $A \subset \mathcal{R}^n$ such that

$$f(\mathbf{x}|\theta) = \left\{ \exp\left(\sum_{j=1}^{k} c_j(\theta) T_j(\mathbf{x}) + d(\theta) + S(\mathbf{x}) \right) \right\} I_A(\mathbf{x}).$$

By the Factorization Theorem, $(T_1(\mathbf{X}), \ldots, T_k(\mathbf{X}))$ is sufficient for θ. Note that, in a k-parameter exponential family, $(T_1, \ldots, T_k)$ is the k-dimensional sufficient statistics for θ. The parameter here is θ, and not $(c_1(\theta), \ldots, c_k(\theta))$.

Example. $X \sim N(\mu, \sigma^2)$. Then

$$\begin{aligned} f(x|\mu,\sigma^2) &= (2\pi)^{-1/2}\sigma^{-1} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) I_{(-\infty,\infty)}(x) \\ &= (2\pi)^{-1/2}\sigma^{-1} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} \right) I_{(-\infty,\infty)}(x) \\ &= \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 + -\frac{\mu^2}{2\sigma^2} - \log(\sigma) - \frac{1}{2}\log(2\pi) \right) I_{(-\infty,\infty)}(x). \end{aligned}$$

We can take $T_1(x) = x$, $T_2(x) = x^2$, $c_1(\mu,\sigma^2) = \frac{\mu}{\sigma^2}$, $c_2(\mu,\sigma^2) = -\frac{1}{2\sigma^2}$, $d(\mu,\sigma^2) = -\log(\sigma) - \frac{\mu^2}{2\sigma^2}$, $S(x) = -\frac{1}{2}\log(2\pi)$, $A = \mathcal{R}$ to see that it is a 2-parameter exponential family. Now consider $X_1, \ldots, X_m$ i.i.d from $N(\mu, \sigma^2)$. Then $(\sum_{i=1}^m X_i, \sum_{i=1}^m X_i^2)$ is sufficient for (μ, σ^2).

Note that in a k-parameter exponential family, θ need not be k-dimensional. For example, consider $N(\theta, \theta^2)$, which is a 2-parameter exponential family, but the parameter is $\theta \in \mathcal{R}^1$.

Ancillary Statistics

There are various results in classical statistics that show a sufficient statistic contains all the information about θ in the data $\mathbf{X}$. At the other end is a statistic whose distribution does not depend on θ and so contains no information about θ. Such a statistic is called *ancillary*.

Definition. Let $\mathbf{X} \sim P_\theta$. A statistic $S(\mathbf{X})$ whose distribution does not depend on the parameter θ is called an ancillary statistic.

Alone, ancillary statistic contains no information about the parameter. However, combination of ancillaries may be informative. For example, consider

(X, Y) which is bivariate normal, with both means equal to 0, both variances equal to 1, and covariance of ρ. Then both X and Y are ancillary by themselves, but together they are informative about ρ. Ancillary statistics are easy to exhibit if $X_1, \ldots, X_n$ are i.i.d. with a location-scale family of densities.

Example. $X_1, \ldots, X_n$ are i.i.d. $N(\theta, 1)$, $-\infty < \theta < \infty$. Then $\bar{X}$ is minimal sufficient. (Show this directly by checking when $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ is free of θ. A different method will be given later.) Now note that $S(X_1, \ldots, X_n) = \sum_{i=1}^n (X_i - \bar{X})^2$ is ancillary. Either because, $S^2 \sim \chi^2_{n-1}$ which is free of θ, or because $X_i - \bar{X} = (X_i - \theta) - (\bar{X} - \theta) = Z_i - \bar{Z}$ where $Z_i = X_i - \theta$. Since θ is location parameter for X_i, distribution of Z_i is free of θ. Similarly, if $X_1, \ldots, X_n$ are i.i.d. $N(0, \sigma^2)$, then $V^2 = \sum_{i=1}^n X_i^2$ is sufficient and $T = \bar{X}/V$ is ancillary. Either, note that

$$nT^2 = \frac{n\bar{X}^2}{n\bar{X}^2 + \sum_{i=1}^n (X_i - \bar{X})^2} \sim \text{Beta}(\frac{1}{2}, \frac{n-1}{2}),$$

which is free of σ, or that

$$T = \frac{\bar{X}}{V} = \frac{\bar{X}/\sigma}{V/\sigma} = \frac{\bar{Z}}{V_Z},$$

where $Z_i = X_i/\sigma$ and $V_Z^2 = \sum_{i=1}^n Z_i^2$; Z_i is free of σ since it is a scale parameter of X_i.

In fact, here is a general result. Let $X_1, \ldots, X_n$ be i.i.d from a location-scale distribution with location μ and scale σ. Then, for any four integers a, b, c, and d (between 1 and n), the ratio

$$\frac{X_{(a)} - X_{(b)}}{X_{(c)} - X_{(d)}} = \frac{Z_{(a)} - Z_{(b)}}{Z_{(c)} - Z_{(d)}}$$

is ancillary because the right-hand side is expressed in terms of order statistics of Z_i's where $Z_i = (X_i - \mu)/\sigma$, $i = 1, \ldots, n$ are i.i.d. with a distribution free of μ and σ.

Example. Let $X_1, \ldots, X_n$ be i.i.d $U(\theta, \theta + 1)$, $-\infty < \theta < \infty$. Then

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\theta) &= \begin{cases} 1 & \text{if } \theta < x_{(1)} < \cdots < x_{(n)} < \theta + 1; \\ 0 & \text{otherwise.} \end{cases} \\ &= \begin{cases} 1 & \text{if } x_{(n)} - 1 < \theta < x_{(1)}; \\ 0 & \text{otherwise.} \end{cases}, \end{aligned}$$

implying that $(X_{(1)}, X_{(n)})$ is sufficient for θ. For two sample points $\mathbf{x}$ and $\mathbf{y}$ (they must satisfy $x_{(1)} < x_{(n)} < x_{(1)} + 1$ and similar property for $\mathbf{y}$), consider the ratio $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$. This is a constant equal to 1 if $x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)}$. If these equalities do not hold, then there will exist θ for which $f(\mathbf{x}|\theta) > 0$ and $f(\mathbf{y}|\theta) = 0$ and some other θ for which $f(\mathbf{x}|\theta) = 0$ and $f(\mathbf{y}|\theta) > 0$. Then the ratio above will not be a constant function of θ. Therefore, $(X_{(1)}, X_{(n)})$ is minimal sufficient for θ. Then $((X_{(1)}+X_{(n)})/2, X_{(n)}-X_{(1)})$ which is a one-one function is also minimal sufficient. (Note they are equivalent statistics and provide the same partition of the sample space.) Now note that $R = X_{(n)} - X_{(1)} = (X_{(n)} - \theta) - (X_{(1)} - \theta) = Z_{(n)} - Z_{(1)}$, where $Z_i = X_i - \theta \sim U(0, 1)$. Thus we see that R is ancillary even though it is part of the minimal sufficient statistics. Note from the following that $R \sim \text{Beta}(n-1, 2)$, which shows once again that it is free of θ.

$$
\begin{aligned}
&P(X_{(1)} > x_{(1)}, X_{(n)} \le x_{(n)}) = P(x_{(1)} < X_i \le x_{(n)} \ \forall\ i) \\
&\quad = \quad (x_{(n)} - x_{(1)})^n \text{ if } \theta < x_{(1)} < x_{(n)} < \theta + 1; \quad \text{so} \\
&F_{X_{(1)},X_{(n)}}(x_{(1)}, x_{(n)}) = P(X_{(1)} \le x_{(1)}, X_{(n)} \le x_{(n)}) \\
&\quad = \quad P(X_{(n)} \le x_{(n)}) - P(X_{(1)} > x_{(1)}, X_{(n)} \le x_{(n)}) \\
&\quad = \quad g(x_{(n)}) - (x_{(n)} - x_{(1)})^n \ \text{ if } \ \ \theta < x_{(1)} < x_{(n)} < \theta + 1.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
f_{X_{(1)},X_{(n)}}(x_{(1)}, x_{(n)}) &= \frac{\partial^2}{\partial x_{(1)} \partial x_{(n)}} F_{X_{(1)},X_{(n)}}(x_{(1)}, x_{(n)}) \\
&= n(n-1)(x_{(n)} - x_{(1)})^{n-2} \ \text{ if } \ \ \theta < x_{(1)} < x_{(n)} < \theta + 1.
\end{aligned}
$$

Taking $R = X_{(n)} - X_{(1)}$, $M = (X_{(1)} + X_{(n)})/2$, we get $X_{(1)} = (2M - R)/2$ and $X_{(n)} = (2M + R)/2$, with the Jacobian of the transformation equal to 1 (since $\begin{vmatrix} -\frac{1}{2} & 1 \\ \frac{1}{2} & 1 \end{vmatrix} = -1$), and hence

$$
\begin{aligned}
f_{R,M}(r, m) &= \begin{cases} n(n-1)r^{n-2} & \text{if } \ \ 0 < r < 1,\ \theta + \frac{r}{2} < m < \theta + 1 - \frac{r}{2}; \\ 0 & \text{otherwise, and} \end{cases} \\
f_R(r) &= \int_{\theta + \frac{r}{2}}^{\theta + 1 - \frac{r}{2}} n(n-1)r^{n-2}\, dm = n(n-1)r^{n-2}(1-r), 0 < r < 1.
\end{aligned}
$$

Let us state this as a general result.

Result. Let $X_1, \ldots, X_n$ be i.i.d from a location parameter family with cdf $F_X(x|\theta) = F_0(x - \theta)$, $-\infty < \theta < \infty$. Then $R = X_{(n)} - X_{(1)}$ is ancillary.

Proof. Let $Z_i = X_i - \theta$. Then Z_i has location 0 and cdf $F_Z(z) = F_0(z)$. Further,

$$\begin{aligned} F_R(r|\theta) &= P_\theta(R \leq r) = P(X_{(n)} - X_{(1)} \leq r) \\ &= P((Z_{(n)} + \theta) - (Z_{(1)} + \theta) \leq r) = P(Z_{(n)} - Z_{(1)} \leq r), \end{aligned}$$

which is free of θ.

Result. Let $X_1, \ldots, X_n$ be i.i.d from a scale parameter family with cdf $F_X(x|\sigma) = F_1(x/\sigma)$, $\sigma > 0$. Then any statistic, $h\left(\frac{X_1}{X_n}, \cdots, \frac{X_{n-1}}{X_n}\right)$ is ancillary.

Proof. Let $Z_i = X_i - \sigma$. Then Z_i has scale 1 and cdf $F_Z(z) = F_1(z)$. Note that

$$\begin{aligned} h\left(\frac{X_1}{X_n}, \cdots, \frac{X_{n-1}}{X_n}\right) &= h\left(\frac{X_1/\sigma}{X_n/\sigma}, \cdots, \frac{X_{n-1}/\sigma}{X_n/\sigma}\right) \\ &= h\left(\frac{Z_1}{Z_n}, \cdots, \frac{Z_{n-1}}{Z_n}\right), \end{aligned}$$

which is free of σ.

Estimation and Optimality

Estimation of features of interest of many populations is very important in many areas. Average or median family income, average yield of an agricultural crop, proportion of eligible voters who favour a certain candidate and so on are some examples. In many fields this is done without making use of probabibility models for the data. With the use of appropriate models, efficient procedures can be developed for this. The first thing to realize then is that it becomes a model fitting problem where unknown parameters of the model are to be determined. This is parametric estimation. In other words, we assume that the data $\mathbf{X}$ comes from the model $\{P_\theta, \theta \in \Theta\}$. The first project is *model fitting*, which means we want to fit the best model to the data: choose $\theta \in \Theta$ which best describes the realization of $\mathbf{X}$. This is also known as point estimation to distinguish it from other procedures. The setup is as follows.

Point Estimation. Consider $X_1, \ldots, X_n$ i.i.d from P_θ. Estimate θ or $q(\theta)$. This is the simplest setup, and we will also consider $\mathbf{X} \sim P_\theta$ when it is not necessarily a random sample or i.i.d.

1. Method of moments. Let the population moments be
$\mu_r(\theta) = E_\theta(X^r)$, $r = 1, 2, \ldots$ and the sample moments be
$\hat{\mu}_r(\theta) = \frac{1}{n}\sum_{j=1}^n X_j^r$, $r = 1, 2, \ldots$. Suppose
$q(\theta) = g(\mu_1(\theta), \ldots, \mu_k(\theta))$ for $k \geq 1$, and g is continuous. Then the method of moments estimate of $q(\theta)$ is
$\widehat{q(\theta)} = g(\hat{\mu}_1(\theta), \ldots, \hat{\mu}_k(\theta))$.

Example. $\sigma^2 = Var_\theta(X) = E_\theta(X^2) - \{E_\theta(X)\}^2 = \mu_2(\theta) - \mu_1^2(\theta) = g(\mu_1(\theta), \mu_2(\theta))$ where $g(x, y) = y - x^2$ is continuous in (x, y). The method of moments estimate of $Var_\theta(X)$ is

$$\begin{aligned} \hat{\sigma}^2 &= \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{j=1}^n X_j^2 - \left(\frac{1}{n}\sum_{j=1}^n X_j\right)^2 \\ &= \frac{1}{n}\left(\sum_{j=1}^n X_j^2 - n\bar{X}^2\right) = \frac{1}{n}\sum_{j=1}^n (X_j - \bar{X})^2, \end{aligned}$$

which is the sample variance with divisor n.

Example. $X_1, \ldots, X_n$ i.i.d Poisson(λ). Here $\theta = \lambda > 0$. Then $\mu_1(\theta) = \lambda$ and $\mu_2(\theta) = \lambda + \lambda^2$. Two different method of moments are readily available for λ. The one using only the first moment gives $\hat{\lambda}_1 = \hat{\mu}_1(\theta) = \bar{X}$, and

another using the first two gives $\hat{\lambda}_2 = \hat{\mu}_2(\theta) - \hat{\mu}_1^2(\theta) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2$. Normally one would use the first one, unless there was a need to check how good the Poisson model would be for the given data. Note that the mean and the variance are equal for the Poisson model. In many applications over-dispersion (i.e., variance larger than mean) is common suggesting other possibilities such as the negative binomial model.

It can be readily seen that the method of moments is basically a substitution method, where population moments are substituted by the corresponding sample moments. The idea of fitting a model is not stressed there. The idea of model fitting forms an important basis for the following method.

2. Maximum likelihood estimation

This requires consideration of a concept of fundamental importance called the *likelihood function*.

Likelihood function. Let $\mathbf{x}$ be the observed data; $\{P_\theta, \theta \in \Theta\}$ with density $f(\mathbf{x}|\theta)$ is the model under consideration for model fitting. Then the function $L(\theta, \mathbf{x}) = f(\mathbf{x}|\theta)$, regarded as a function of θ for fixed $\mathbf{x}$ is called the likelihood function. Often $\mathbf{x}$ is suppressed and f is taken as the likelihood function and written $L(\theta)$.

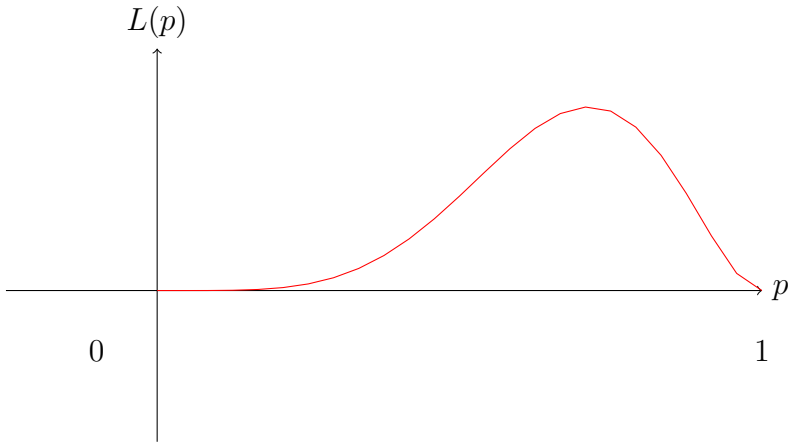
Interpretation of the likelihood function as relevant for inference about θ is the following. The data, $\mathbf{x}$, has been observed already, so θ is the only unknown. Then it makes sense to assume (according to a principle called *likelihood principle*), that all information about θ is contained in $L(\theta)$ for the observed $\mathbf{x}$. Since $f(\mathbf{x}|\theta)$ measures how likely $\mathbf{x}$ is if θ is the true parameter, observing $\mathbf{x}$ must then provide information through $L(\theta)$, on how to regard θ as the true parameter. The likelihood function is not unique in that for any $c(\mathbf{x}) > 0$ that may depend on $\mathbf{x}$ but not on θ, $c(\mathbf{x})f(\mathbf{x}|\theta)$ is also a likelihood function. What is unique are the likelihood ratios $L(\theta_\mathbf{2})/L(\theta_\mathbf{1})$, which indicate how plausible is $\theta_\mathbf{2}$, relative to $\theta_\mathbf{1}$, in the light of the given data $\mathbf{x}$. In particular, if the ratio is large, we have a lot of confidence in $\theta_\mathbf{2}$ relative to $\theta_\mathbf{1}$ and the reverse situation holds if the ratio is small.

Maximum likelihood estimate (MLE). MLE of θ is $\hat{\theta} = \hat{\theta}(x)$ where $L(\hat{\theta}, x) = \max_{\theta\in\Theta} L(\theta, x)$ if the maximum exists.
(i) MLE may not exist, or may not be unique; (ii) if $\hat{\theta}$ is the MLE of θ, then $q(\hat{\theta})$ is taken to be the MLE of $q(\theta)$.

Example. Let $X_1, \ldots, X_n$ be i.i.d Bernoulli(p). Then

$$L(p) = f(x_1, \ldots, x_n|p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

The MLE of p is $\hat{p} = \sum_{i=1}^{n} x_i/n$ as can be seen from the graph of $L(p)$ and using calculus:



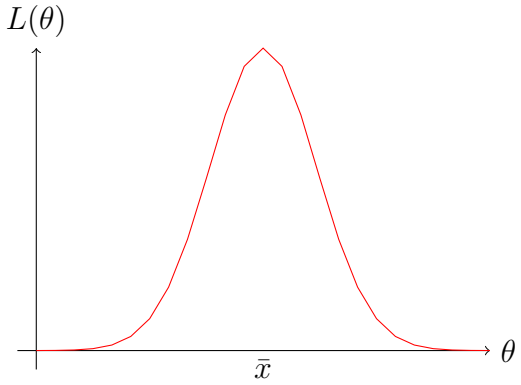
Result. MLE depends on $\mathbf{x}$ only through the sufficient statistics $T(\mathbf{x})$.

Proof. $L(\theta, x) = f(x|\theta) = g(T(x), \theta)h(x)$. Therefore, we have $L(\hat{\theta}(x), x) = \max_\theta g(T(x), \theta)h(x)$. Since $h(x) > 0$, we must have $L(\hat{\theta}(x), x) = h(x) \max_\theta g(T(x), \theta)$, where the maximization is on the part that invoves x through $T(x)$ only.

Example. Let $X_1, \ldots, X_n$ be i.i.d $N(\theta, 1)$. What is the MLE of θ? Sufficient statistic is $\bar{X} \sim N(\theta, 1/n)$. Then

$$L(\theta, x_1, \ldots, x_n) \propto f(\bar{x}|\theta) \propto \exp\left(-\frac{n}{2}(\bar{x} - \theta)^2\right),$$

which is maximized by $\hat{\theta}(x_1, \ldots, x_n) = \bar{x}$.



If the sample size n is large, usually the likelihood function has a sharp peak as shown in the following figure. This peak is at the maximum likelihood

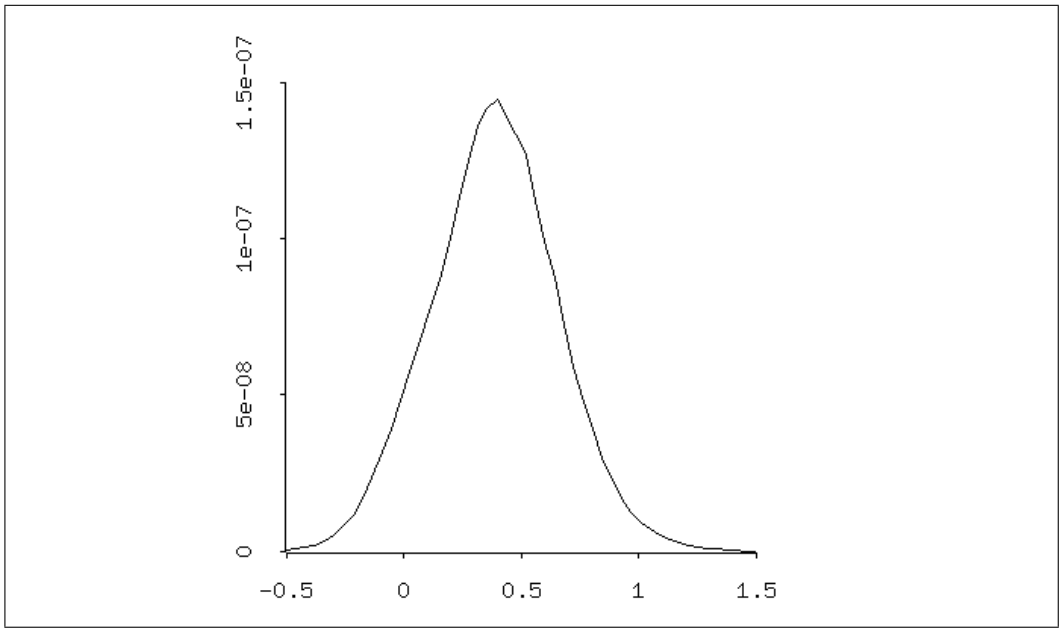


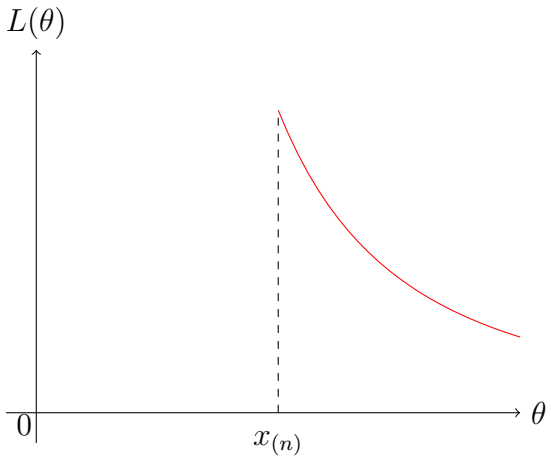
Figure 1: $L(\theta)$ for the double exponential model when data is normal mixture.

estimate (MLE) $\hat{\theta}$. In situations like this, one feels $\hat{\theta}$ is very plausible as an estimate of θ relative to any other points outside a small interval around $\hat{\theta}$. One would then expect $\hat{\theta}$ to be a good estimate of the unknown θ, at least in the sense of being close to it in some way.

Example. Let $X_1, \ldots, X_n$ be i.i.d $U[0, \theta]$, $\theta > 0$. What is the MLE of θ?

$$L(\theta, x_1, \ldots, x_n) = \frac{1}{\theta^n} I\{x_i \leq \theta, i = 1, 2, \ldots, n\} = \frac{1}{\theta^n} I\left\{\max_i x_i \leq \theta\right\}$$

$X_{(n)} = \max_i X_i$ is sufficient (minimal) for θ and $L(\theta)$ is as shown:



Therefore, $\hat{\theta}(x_1, \ldots, x_n) = x_{(n)}$.

Likelihood equations.

Define $\mathcal{L}(\theta, x) = \log L(\theta, x)$ as the log-likelihood function of θ. Suppose Θ is an open set and $\mathcal{L}$ is differentiable in θ for each fixed x. Then, if the MLE $\hat{\theta}(x)$ exists, it satisfies the likelihood equations:

$$\frac{\partial}{\partial \theta_j} \mathcal{L}(\theta, x) = 0, \ j = 1, \ldots, p.$$

This follows from the fact that $\hat{\theta}$ maximizes $L(\theta, x)$, and hence maximizes $\mathcal{L}(\theta, x) = \log L(\theta, x)$ also. Since $\mathcal{L}$ is differentiable, $\hat{\theta}$ is a zero of its derivative.

If $X_1, \ldots, X_n$ are independent and X_i has density $f_i(x|\theta)$, then
$L(\theta, x_1, \ldots, x_n) = \prod_{i=1}^n f_i(x_i|\theta)$, $\mathcal{L}(\theta, x_1, \ldots, x_n) = \sum_{i=1}^n \log f_i(x_i|\theta)$,
and hence the likelihood equations are given by

$$\frac{\partial}{\partial \theta} \mathcal{L}(\theta, x_1, \ldots, x_n) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_i(x_i|\theta) = 0.$$

Example. $X \sim \text{Binomial}(n, \theta)$, $0 < \theta < 1$. Then

$$\begin{aligned} L(\theta, x) &= \binom{n}{x} \theta^x (1-\theta)^{n-x}, \\ \mathcal{L}(\theta, x) &= x \log(\theta) + (n-x) \log(1-\theta) + c(x) \\ \frac{\partial}{\partial \theta} \mathcal{L}(\theta, x) &= \frac{x}{\theta} - \frac{n-x}{1-\theta}. \end{aligned}$$

$\frac{\partial}{\partial \theta} \mathcal{L}(\theta, x) = 0$ has only one solution $\hat{\theta} = x/n$ which is a maximum since $\frac{\partial^2}{\partial \theta^2} \mathcal{L}(\theta, x) < 0$.

Theorem. Let $\{P_\theta, \theta \in \Theta\}$ be a one-parameter exponential family with density $f(x|\theta) = \exp(c(\theta)T(x) + d(\theta) + S(x))I_A(x)$, and let C be the interior of $\{c(\theta), \theta \in \Theta\}$. Suppose $\theta \to c(\theta)$ is one-one. If the equation $E_\theta(T(X)) = T(x)$ has a solution $\hat{\theta}(x)$ for which $c(\hat{\theta}(x)) \in C$, then $\hat{\theta}(x)$ is the unique MLE of θ.

Proof. Since $\theta \to c(\theta)$ is one-one, maximizing the likelihood over θ is the same as maximizing over $\eta = c(\theta)$. Hence consider the natural parametrization:

$$\begin{aligned} f(x|\eta) &= \exp\left(\eta T(x) + d_0(\eta) + S(x)\right) I_A(x), \ \eta \in H, \\ \mathcal{L}(\eta, x) &= \eta T(x) + d_0(\eta) + S(x) \ \text{ if } \ x \in A, \\ \frac{\partial}{\partial \eta} \mathcal{L}(\eta, x) &= T(x) + d_0'(\eta), \\ \frac{\partial^2}{\partial \eta^2} \mathcal{L}(\eta, x) &= d_0''(\eta), \end{aligned}$$

For η which is an interior point of H, we have, $-d_0'(\eta) = E_\eta(T(X))$ and $-d_0''(\eta) = Var(T(X)) > 0$. Therefore, we get,

$$\frac{\partial}{\partial \eta} \mathcal{L}(\eta, x) = T(x) - E_\eta(T(X)) = 0$$

implying that $E_\eta(T(X)) = T(x)$. Now $\frac{\partial^2}{\partial \eta^2} \mathcal{L}(\eta, x) < 0$ so that $\mathcal{L}$ is strictly concave. Thus we get a unique maximum at $\hat{\eta}(x)$ for which $E_\eta(T(X))|_{\eta=\hat{\eta}(x)} = T(x)$. The same argument goes through for k-parameter exponential family, but one needs to work with the covariance matrix.

Example. Let $X_1, \ldots, X_n$ be i.i.d $N(\mu, \sigma^2)$, $n \geq 2$. This is a 2-parameter exponential family with

$$f(x_1, \ldots, x_n) = \exp\left(\frac{\mu}{\sigma^2} \sum_{i=1}^n x_i - \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 - \frac{n\mu^2}{2\sigma^2} - n \log(\sigma) - \frac{n}{2} \log(2\pi) \right),$$

so that $c_1(\theta) = \frac{\mu}{\sigma^2}$, $c_2(\theta) = -\frac{1}{2\sigma^2}$, $T_1(\mathbf{x}) = \sum_{i=1}^n x_i$, $T_2(\mathbf{x}) = \sum_{i=1}^n x_i^2$. Also, $(\mu, \sigma^2) \to (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ is one-one. Solve: $E_\theta(T(\mathbf{X})) = T(\mathbf{x})$. i.e., solve

$$\begin{aligned} E_\theta(T_1(\mathbf{X})) = E_\theta(\sum_{i=1}^n X_i) = n\mu &= \sum_{i=1}^n x_i, \\ E_\theta(T_2(\mathbf{X})) = E_\theta(\sum_{i=1}^n X_i^2) = n(\mu^2 + \sigma^2) &= \sum_{i=1}^n x_i^2, \end{aligned}$$

yielding, $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$. They are MLE if $(\hat{\mu}, \hat{\sigma}^2)$ is an interior point. i.e., $\hat{\sigma}^2 > 0$.

What if $n = 1$? Then, $f(x|\mu, \sigma^2) \propto \sigma^{-1} \exp(-\frac{1}{2\sigma^2}(x - \mu)^2) = L((\mu, \sigma^2), x)$ which is unbounded as $\sigma \to 0$. To see this, consider $\hat{\mu} = x$ and $L(\hat{\mu}, \sigma^2) = 1/\sigma$. MLE do not exist in this case.

Some more examples where some methods of estimation work whereas others do not.

Example. ϵ-contamination models. Consider the model with cdf:

$$F(x|\theta) = 0.9\Phi(\frac{x-\mu}{\sigma}) + 0.1\Phi(x-\mu),$$

where Φ is the standard normal cdf. In this distribution, $X \sim N(\mu, \sigma^2)$ with chance 90% and with 10% chance it is $N(\mu, 1)$. Then

$$f(x|\theta) = 0.9\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma}) + 0.1\phi(x-\mu),$$

where ϕ is the standard normal pdf. Suppose we have a random sample, $X_1, \ldots, X_n$ from F_θ, where $\Theta = \{\theta = (\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$. What is MLE of θ?

$$\begin{aligned} L(\theta, x_1, \ldots, x_n) &= f(x_1, \ldots, x_n|\theta) \\ &= \prod_{i=1}^{n} \left[0.9\frac{1}{\sigma}\phi(\frac{x_i-\mu}{\sigma}) + 0.1\phi(x_i-\mu)\right] \\ &= (2\pi)^{-n/2} \prod_{i=1}^{n} \left[0.9\frac{1}{\sigma}\exp(-\frac{(x_i-\mu)^2}{2\sigma^2}) + 0.1\exp(-\frac{(x_i-\mu)^2}{2})\right]. \end{aligned}$$

What is $\max_\theta L(\theta, x_1, \ldots, x_n)$? Consider $\hat{\mu} = x_{(1)}$ and

$$\begin{aligned} &L((\hat{\mu}, \sigma^2), x_1, \ldots, x_n) \\ &= (2\pi)^{-n/2} \left[0.9\frac{1}{\sigma}\exp(-\frac{(x_{(1)}-\mu)^2}{2\sigma^2}) + 0.1\exp(-\frac{(x_{(1)}-\mu)^2}{2})\right] \\ &\quad \times \prod_{i=2}^{n} \left[0.9\frac{1}{\sigma}\exp(-\frac{(x_{(i)}-\mu)^2}{2\sigma^2}) + 0.1\exp(-\frac{(x_{(i)}-\mu)^2}{2})\right] \\ &= (2\pi)^{-n/2} \left[\frac{0.9}{\sigma} + 0.1\right] \\ &\quad \times \prod_{i=2}^{n} \left[0.9\frac{1}{\sigma}\exp(-\frac{(x_{(i)}-x_{(1)})^2}{2\sigma^2}) + 0.1\exp(-\frac{(x_{(i)}-x_{(1)})^2}{2})\right]. \end{aligned}$$

Note that as $\sigma \to 0$, $\frac{1}{\sigma}\exp(-\frac{(x_{(i)}-x_{(1)})^2}{2\sigma^2}) \to 0$ for $i \geq 2$. Therefore,

$$\begin{aligned} &\lim_{\sigma\to 0} L((\hat{\mu}, \sigma^2), x_1, \ldots, x_n) \\ &= (2\pi)^{-n/2} \lim_{\sigma\to 0} \left[\frac{0.9}{\sigma} + 0.1\right] \prod_{i=2}^{n} \left[0.1\exp(-\frac{(x_{(i)}-x_{(1)})^2}{2})\right] = \infty. \end{aligned}$$

Therefore, MLE of (μ, σ^2) does not exist. However method of moments estimate can be derived.

An example where MLE exists but method of moments do not.

Example. Let $X_1, \ldots, X_n$ be i.i.d Cauchy(θ) with density
$f(x|\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$, $-\infty < x < \infty$; $-\infty < \theta < \infty$.
If $n = 1$, then $L(\theta, x_1) = \frac{1}{\pi}\frac{1}{1+(x_1-\theta)^2}$, so $\hat{\theta} = x_1$ is the MLE. If $n = 2$,

$$\begin{aligned} L(\theta, x_1, x_2) &= \frac{1}{\pi^2}\frac{1}{1+(x_1-\theta)^2}\frac{1}{1+(x_2-\theta)^2}, \\ \mathcal{L}(\theta, x_1, x_2) &= \text{constant} - \log(1+(x_1-\theta)^2) - \log(1+(x_2-\theta)^2), \\ \frac{\partial}{\partial\theta}\mathcal{L}(\theta, x_1, x_2) &= \frac{2(x_1-\theta)}{1+(x_1-\theta)^2)} + \frac{2(x_2-\theta)}{(1+(x_2-\theta)^2)}. \end{aligned}$$

Therefore, $\frac{\partial}{\partial\theta}\mathcal{L} = 0$ iff

$$\begin{aligned} \frac{(x_1-\theta)[1+(x_2-\theta)^2] + (x_2-\theta)[1+(x_1-\theta)^2]}{[1+(x_1-\theta)^2][1+(x_2-\theta)^2]} &= 0 \text{ iff} \\ g(\theta) \equiv (x_1-\theta) + (x_1-\theta)(x_2-\theta)^2 + (x_2-\theta) + (x_2-\theta)(x_1-\theta)^2 &= 0. \end{aligned}$$

Since $\hat{\theta}_1 = (x_1+x_2)/2$ satisfies $x_1 - \hat{\theta}_1 = (x_1-x_2)/2 = -(x_2-x_1)/2 = -(x_2-\hat{\theta}_1)$, we have that $\hat{\theta}_1$ is a root of $g(\theta)$ or a solution of $\frac{\partial}{\partial\theta}\mathcal{L} = 0$. Now note that

$$g(\theta) = (\theta - \hat{\theta}_1)\left(-2\theta^2 + 2(x_1+x_2)\theta - 2(1+x_1x_2)\right).$$

Therefore the other two roots, $\hat{\theta}_2$ and $\hat{\theta}_3$ are

$$\begin{aligned} & \frac{x_1+x_2}{2} \pm \frac{1}{2}\sqrt{(x_1+x_2)^2 - 4(1+x_1x_2)} \\ = & \frac{x_1+x_2}{2} \pm \frac{1}{2}\sqrt{x_1^2+x_2^2+2x_1x_2-4x_1x_2-4} \\ = & \frac{x_1+x_2}{2} \pm \frac{1}{2}\sqrt{(x_1-x_2)^2-4}. \end{aligned}$$

Case 1. $(x_1-x_2)^2 < 4$. $\hat{\theta}_1$ is the only real root. Check that this is the unique MLE.
Case 2. $(x_1-x_2)^2 = 4$. Only one root. Again, check that this is MLE.
Case 3. $(x_1-x_2)^2 > 4$. There are 3 real roots now. Check that $\hat{\theta}_1$ is a minimum, $\hat{\theta}_2$ and $\hat{\theta}_3$ are both MLE since $L(\hat{\theta}_2) = L(\hat{\theta}_3)$. Since Cauchy does not possess any moments, method of moments estimates are not available.

Example. Consider a random sample from $U[0,\theta]$. Then $X_{(n)}$ is the MLE, which is a function of the minimal sufficient statistic, whereas the method of moments estimate is $2\bar{X}$ which is not.

Example. Consider a random sample $X_1,\ldots,X_n$ from Gamma(α,λ). This is a 2-parameter exponential family, so it is easy to write down the likelihood equations. However, they cannot be solved explicitly since they involve $\Gamma(\alpha)$. One is confronted with a computational issue here. Newton's method can be used in some of these situations. We need $\hat{\theta}(x)$ such that $\frac{\partial \mathcal{L}(\theta,x)}{\partial\theta}\Big|_{\theta=\hat{\theta}(x)} = 0$. Let $g(\theta) = \frac{\partial \mathcal{L}(\theta,x)}{\partial\theta}$. We know that $g(\hat{\theta}) = 0$. Let $\tilde{\theta}$ be an approximation for $\hat{\theta}$. Then, assuming g is a smooth function,

$$\begin{aligned} 0 &= g(\hat{\theta}) = g(\tilde{\theta}) + (\hat{\theta}-\tilde{\theta})g'(\tilde{\theta}) + \frac{(\hat{\theta}-\tilde{\theta})^2}{2}g''(\theta^*) \\ &\approx g(\tilde{\theta}) + (\hat{\theta}-\tilde{\theta})g'(\tilde{\theta}), \end{aligned}$$

where θ^* lies between $\hat{\theta}$ and $\tilde{\theta}$, and the last term is ignored. Therefore, $\hat{\theta}-\tilde{\theta} \approx \frac{-g(\tilde{\theta})}{g'(\tilde{\theta})}$, or

$$\hat{\theta} = \tilde{\theta} - \frac{g(\tilde{\theta})}{g'(\tilde{\theta})} = \tilde{\theta} - \frac{\frac{\partial \mathcal{L}(\theta,x)}{\partial\theta}}{\frac{\partial^2 \mathcal{L}(\theta,x)}{\partial\theta^2}}\Big|_{\theta=\tilde{\theta}}$$

is an iterative procedure to locate $\hat{\theta}$.

Truncated data. Observations below or above a certain level cannot be measured. For example, due to limitations of the measuring device,
X = blood alcohol level in a test is recorded if and only if $X > a$. Consider $X_1,\ldots,X_n$ i.i.d from this distribution. Suppose Y = untruncated blood alcohol level $\sim N(\mu,\sigma^2)$. Find MLE of (μ,σ^2) using $X_1,\ldots,X_n$. Note that

$$P(X > x|\mu,\sigma^2) = P(Y > x|Y > a,\mu,\sigma^2) = \frac{1-\Phi(\frac{x-\mu}{\sigma})}{1-\Phi(\frac{a-\mu}{\sigma})}, x > a.$$

Therefore,

$$f_X(x|\mu,\sigma^2) = \frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{1-\Phi(\frac{a-\mu}{\sigma})}, x > a.$$

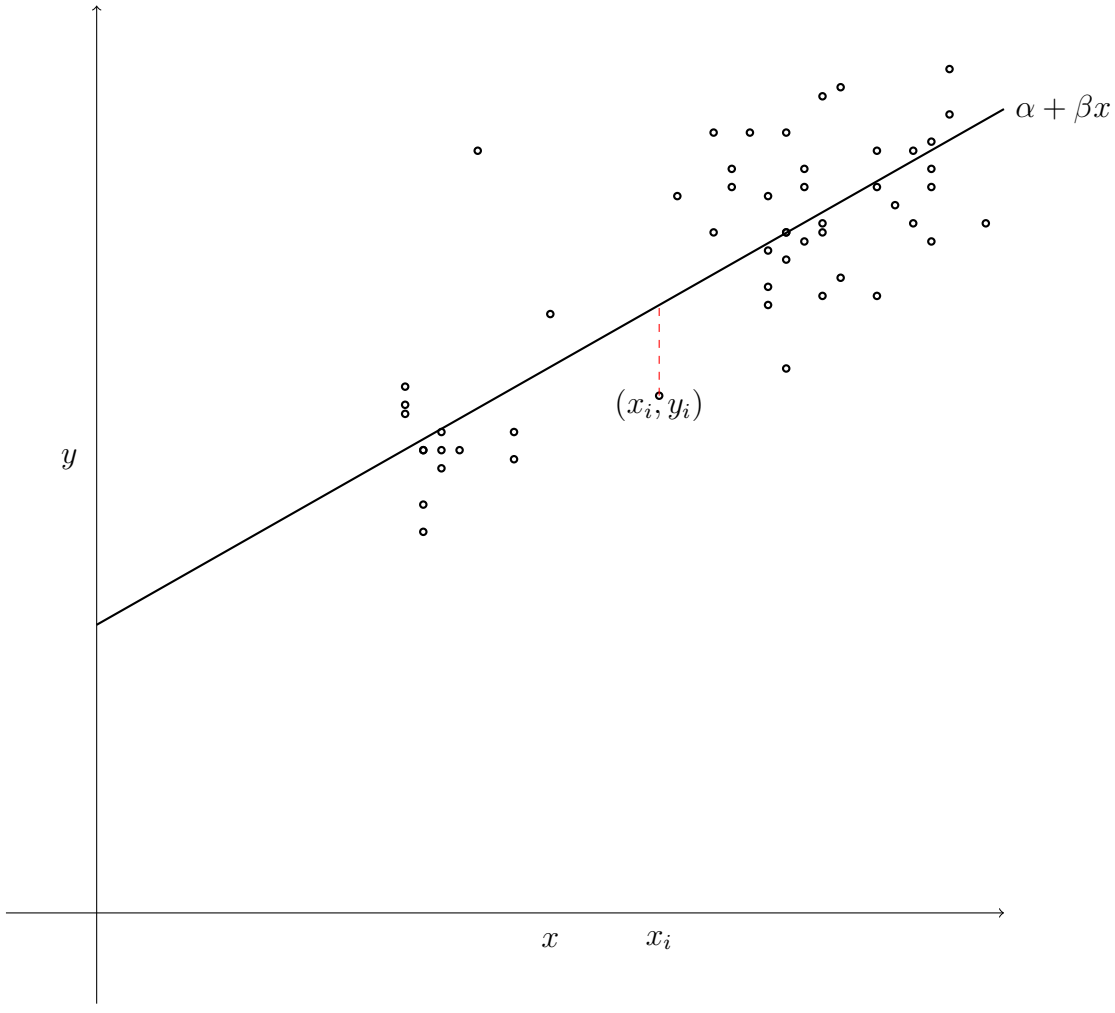
$$\begin{aligned} &L((\mu,\sigma^2),x_1,\ldots,x_n) \\ &= \sigma^{-n}\left[1-\Phi(\frac{a-\mu}{\sigma})\right]^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2\right)I_{(a,\infty)^n}(\mathbf{x}), \\ &\mathcal{L}((\mu,\sigma^2),x_1,\ldots,x_n) \\ &= -n\left\{\log(\sigma)+\log\left[1-\Phi(\frac{a-\mu}{\sigma})\right]\right\} - \frac{1}{2\sigma^2}\left(\sum_{i=1}^n(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right). \end{aligned}$$

This is a 2-parameter exponential family, but no explicit solutions can be derived. Numerical solutions such as Newton's method can be used with the likelihood equations.

3. Least squares.

Linear models. Response y and factor/predictor x are measured on n subjects: $(x_1, y_1), \ldots, (x_n, y_n)$. Modeling the linear dependence of y on x for estimation and prediction is of interest. With this in view, the following model is explored.
$y_i = \alpha + \beta x_i + \epsilon_i$, $i = 1, \ldots, n$ where ϵ_i are uncorrelated random errors with mean 0 and variance σ^2. Then the *least squares* method is to estimate α and β by:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$



Since we need to minimize a quadratic function in α and β, we may simply

differentiate it and set the partial derivatives to 0. We then obtain

$$\begin{aligned}
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}, \text{ and} \\
\hat{\beta} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \times \frac{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \\
&= r_{xy}\frac{s_y}{s_x},
\end{aligned}$$

where r_{xy} is the correlation coefficient between x and y, and s_x, s_y are the s.d. of x and y, respectively. $\hat{y} = \hat{\alpha} + \hat{\beta}x = \hat{\alpha} + r_{xy}\frac{s_y}{s_x}x$ is the least squares equation to predict y based on x.

Now suppose ϵ_i are i.i.d $N(0, \sigma^2)$. Then y_i are independent and $y_i \sim N(\alpha + \beta x_i, \sigma^2)$. What is the MLE of $(\alpha, \beta, \sigma^2)$? Note that the model is for $y|x$, treating x fixed. Then, we have,

$$\begin{aligned}
f(\mathbf{y}|\alpha, \beta, \sigma^2) &= (2\pi)^{-n/2}\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right), \\
\mathcal{L}(\alpha, \beta, \sigma^2) &= -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2.
\end{aligned}$$

For each fixed σ^2, maximization of $\mathcal{L}(\alpha, \beta, \sigma^2)$ over α, β is the same as minimization of $\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$. Therefore MLE of (α, β) is the same as the least squares estimate. This shows an optimality property of least squares under normality.

Now we show that the above normal linear model is a 3-parameter exponential family. Note that

$$\begin{aligned}
\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 &= \sum_{i=1}^n \left(y_i - \hat{\alpha} - \hat{\beta}x_i - (\alpha - \hat{\alpha}) - (\beta - \hat{\beta})x_i\right)^2 \\
&= \sum_{i=1}^n \left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2 + \sum_{i=1}^n \left\{-(\alpha - \hat{\alpha}) - (\beta - \hat{\beta})x_i\right\}^2 \\
&\quad -2\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)\left\{(\alpha - \hat{\alpha}) + (\beta - \hat{\beta})x_i\right\} \\
&= \sum_{i=1}^n \left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2 + \sum_{i=1}^n \left\{(\alpha - \hat{\alpha}) + (\beta - \hat{\beta})x_i\right\}^2,
\end{aligned}$$

since

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i-\hat\alpha-\hat\beta x_i)\left\{(\alpha-\hat\alpha)+(\beta-\hat\beta)x_i\right\}\\
&= \sum_{i=1}^{n}(y_i-\bar y-\hat\beta(x_i-\bar x))\left\{(\alpha-\hat\alpha)+(\beta-\hat\beta)x_i\right\}\\
&= (\alpha-\hat\alpha)\sum_{i=1}^{n}(y_i-\bar y)-(\alpha-\hat\alpha)\hat\beta\sum_{i=1}^{n}(x_i-\bar x)+(\beta-\hat\beta)\sum_{i=1}^{n}(y_i-\bar y)x_i\\
&\quad -\hat\beta(\beta-\hat\beta)\sum_{i=1}^{n}x_i(x_i-\bar x)\\
&= (\beta-\hat\beta)\sum_{i=1}^{n}x_i\left\{(y_i-\bar y)-\hat\beta(x_i-\bar x)\right\}\\
&= (\beta-\hat\beta)\left\{\sum_{i=1}^{n}(x_i-\bar x)(y_i-\bar y)-\hat\beta\sum_{i=1}^{n}(x_i-\bar x)^2\right\}\\
&= 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
f(\mathbf{y}|\alpha,\beta,\sigma^2) &= (2\pi)^{-n/2}\sigma^{-n}\\
&\times \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\hat\alpha-\hat\beta x_i)^2+\sum_{i=1}^{n}\{(\alpha-\hat\alpha)+(\beta-\hat\beta)x_i\}^2\right]\right),
\end{aligned}$$

and hence $\{\hat\alpha,\hat\beta,\sum_{i=1}^{n}\left(y_i-\hat\alpha-\hat\beta x_i\right)^2\}$ is sufficient for (α,β,σ^2). To show that this is an exponential family, note

$$\begin{aligned}
\sum_{i=1}^{n}\left\{(\alpha-\hat\alpha)+(\beta-\hat\beta)x_i\right\}^2 &= \sum_{i=1}^{n}\left\{(\alpha+\beta x_i)-(\hat\alpha+\hat\beta x_i)\right\}^2\\
&= \sum_{i=1}^{n}(\alpha+\beta x_i)^2+\sum_{i=1}^{n}(\hat\alpha+\hat\beta x_i)^2\\
&\quad -2\sum_{i=1}^{n}(\hat\alpha+\hat\beta x_i)(\alpha+\beta x_i).
\end{aligned}$$

Therefore,

$$\begin{aligned} & f(\mathbf{y}|\alpha,\beta,\sigma^2) \\ &= \exp\left(\frac{\alpha}{\sigma^2}\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i)+\frac{\beta}{\sigma^2}\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i)x_i\right. \\ &\quad -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2+\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i)^2\right] \\ &\quad \left.-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\alpha+\beta x_i)^2-n\log(\sigma)-n\log(\sqrt{2\pi})\right), \end{aligned}$$

so that we can take $c_1(\alpha,\beta,\sigma^2)=\alpha/\sigma^2$, $T_1(\mathbf{y})=\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i)$, $c_2(\alpha,\beta,\sigma^2)=\beta/\sigma^2$, $T_2(\mathbf{y})=\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i)x_i$, $c_3(\alpha,\beta,\sigma^2)=-1/(2\sigma^2)$,
$T_3(\mathbf{y})=\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2+\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i)^2$. Then, we can also use exponential family methods to find MLE of (α,β,σ^2). Since we already know the MLE for α and β, and since, for each σ^2,

$$\max_{\alpha,\beta}L(\alpha,\beta,\sigma^2)=L(\hat{\alpha},\hat{\beta},\sigma^2)=\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2\right),$$

we can find the MLE of σ^2 by finding

$$\max_{\sigma^2}L(\hat{\alpha},\hat{\beta},\sigma^2)=\max_{\sigma^2}(\sigma^2)^{-n/2}\exp\left(-\frac{t^2}{2\sigma^2}\right),$$

where $t^2=\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2$. Then,

$$\begin{aligned} \mathcal{L}(\hat{\alpha},\hat{\beta},\sigma^2) &= -\frac{n}{2}\log(\sigma^2)-\frac{1}{2\sigma^2}t^2, \\ \frac{\partial\mathcal{L}(\hat{\alpha},\hat{\beta},\sigma^2)}{\partial\sigma^2} &= -\frac{n}{2}\sigma^2+\frac{t^2}{2}\frac{1}{(\sigma^2)^2}=\frac{1}{2\sigma^4}(t^2-n\sigma^2). \end{aligned}$$

Check that MLE of σ^2 is

$$\hat{\sigma}^2=\frac{t^2}{n}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2.$$

Criteria of Estimation – Optimality

How should one choose an estimate when many are available? In other words, what criteria are to be used to determine the procedure of estimation? This prompts the question: how good is an estimate?

Suppose $X \sim P_\theta$ and $T(X)$ estimates $q(\theta)$. Then $|T(X) - q(\theta)|$ is the discrepancy in estimation. Does there exist $T(X)$ which can minimize this discrepancy (uniformly) for all θ? No. Take $q(\theta) = \theta$. For $\theta = \theta_1$, $T(X) = \theta_1$ is the best, but this is not optimal for any other θ. Define $L(q(\theta), T(X))$ to be the loss due to estimating $q(\theta)$ by $T(X)$. Standard losses in the theory of estimation are

$$L(q(\theta), T(X)) = \begin{cases} |q(\theta) - T(X)| & \text{absolute error loss;} \\ (q(\theta) - T(X))^2 & \text{squared error loss.} \end{cases}$$

We have already noted that the loss cannot be minimized uniformly. Also, it depends on X which is random. Therefore, we average it over all samples. Then we get $R(q(\theta), T(X)) = E_\theta L(q(\theta), T(X))$ which is called the risk. Thus, we have $E_\theta(|T(X) - q(\theta)|) = \int |q(\theta) - T(x)| f(x|\theta)\, dx$ = mean absolute error and $E_\theta((T(X) - q(\theta))^2) = \int (q(\theta) - T(x))^2 f(x|\theta)\, dx$ = mean square error (MSE). Since these cannot be minimized uniformaly for all θ, one may restrict $T(X)$ to some class of estimators and then choose the best in that class.

Unbiased estimators. $T(X)$ is said to be unbiased for $q(\theta)$ if $E_\theta(T(X)) = q(\theta)$ for all $\theta \in \Theta$.

Example. $X_1, \ldots, X_n$ i.i.d Exp(λ), $q(\lambda) = 1/\lambda$. Then $\bar{X}$ is an unbiased estimator of $q(\lambda)$ since $E(\bar{X}) = E(X) = 1/\lambda$ for all λ.

Note that MSE for unbiased estimators is just the variance of the estimate:

$$E_\theta(T(X) - q(\theta))^2 = E_\theta(T(X) - E_\theta(T(X)))^2 = Var(T(X)).$$

Unbiasedness means only that $E_\theta(T(X)) - q(\theta) = 0$. i.e., if used over and over again, on the average, underestimation will balance overestimation; no consideration is given to how often or by how much the estimate will depart from the parameter.

It is possible in many situations to find an estimate which is best among all unbiased estimates in terms of variance. Such an estimate is called *Uniformly Minimum Variance Unbiased Estimate* or UMVUE or *Best Unbiased Estimate*. Note that

- UMVUE may not exist;
- unbiasedness may be ridiculous;
- there may be better and simpler estimates which are not unbiased.

Example (Unbiased estmates do no exist). Suppose $X \sim \text{Binomial}(n, p)$. We want to estimate $q(p) = \frac{1}{p}$. Since $\hat{p} = X/n$ for both method of moments and MLE, n/X is the corresponding estimate, except when $X = 0$. (No estimate when $X = 0$.) Unbiasedness means,

$$E_p(T(X)) = \frac{1}{p} \text{ for all } p \in (0, 1).$$

(Note that $T(x)$ needs to be defined for all realizable x for computing $E(T)$.) Then, we must have,

$$\sum_{x=0}^{n} T(x) \binom{n}{x} p^x (1-p)^{n-x} = \frac{1}{p} \text{ for all } p \in (0, 1). \text{ i.e.,}$$

$$T(0)(1-p)^n + \sum_{x=1}^{n} T(x) \binom{n}{x} p^x (1-p)^{n-x} = \frac{1}{p} \text{ for all } p \in (0, 1).$$

As $p \to 0$, LHS $\to T(0)$ which is a real number, whereas RHS $\to \infty$. No such T exists.

Example (Unbiased estimates are silly). Suppose $X \sim$ truncated Poisson:

$$P_\lambda(X = x) = \frac{\exp(-\lambda)\lambda^x / x!}{1 - \exp(-\lambda)}, x = 1, 2, \ldots$$

Estimate $q(\lambda) = \exp(-\lambda)$, a positive quantity. Consider any unbiased estimate $T(X)$. Then,

$$\exp(-\lambda) = E(T(X)) = \sum_{x=1}^{\infty} T(x) \frac{\exp(-\lambda)\lambda^x / x!}{1 - \exp(-\lambda)},$$

so that

$$1 - \exp(-\lambda) = E(T(X)) = \sum_{x=1}^{\infty} T(x) \frac{\lambda^x}{x!}, \qquad \forall \lambda > 0.$$

Now the power series expansion gives,

$$\begin{aligned} \text{LHS} &= 1 - \left[1 - \lambda + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + \cdots\right] \\ &= -\sum_{x=1}^{\infty} \frac{(-1)^x \lambda^x}{x!} = \sum_{x=1}^{\infty} \frac{(-1)^{x+1} \lambda^x}{x!}. \end{aligned}$$

Therefore, we must have that

$$\sum_{x=1}^{\infty} \frac{(-1)^{x+1}\lambda^x}{x!} = \sum_{x=1}^{\infty} T(x)\frac{\lambda^x}{x!}, \qquad \forall \lambda > 0.$$

Since two power series agree on an interval, their coefficients must be equal. Therefore the only unbiased estimate for $\exp(-\lambda)$ is

$$T(x) = \begin{cases} 1 & \text{if } x \text{ is odd;} \\ -1 & \text{if } x \text{ is even.} \end{cases}$$

i.e., our estimate $T(x) < 0$ if x is even!

Let $\hat{\theta}(x)$ be an estimator of θ. Consider $L(\theta, d) = (\theta - d)^2$, the squared error loss. Then

$$\begin{aligned} \text{MSE} &= R(\theta, d) = E_\theta(d(X) - \theta)^2 \\ &= E_\theta\left[d(X) - E_\theta(d(X)) + E_\theta(d(X)) - \theta\right]^2 \\ &= E_\theta\left[d(X) - E_\theta(d(X))\right]^2 + \left[E_\theta(d(X)) - \theta\right]^2 \\ &\quad +2\left[E_\theta(d(X)) - \theta\right] E_\theta\left[d(X) - E_\theta(d(X))\right] \\ &= Var_\theta(d(X)) + \text{Bias}^2(\theta), \end{aligned}$$

where $\text{Bias}(\theta) = E_\theta(d(X)) - \theta$. If $E_\theta(\hat{\theta}(X)) = \theta$ for all θ, i.e., $\hat{\theta}(X)$ is unbiased for θ, then MSE = Variance.

Example. Let $X_1, \ldots, X_n$ be i.i.d $N(\mu, \sigma^2)$. Consider the following estimates for σ^2.

$$\begin{aligned} T_1(\mathbf{X}) &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{S^2}{n}, \\ T_2(\mathbf{X}) &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{S^2}{n-1}. \end{aligned}$$

Since $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \sigma^2\chi^2_{n-1}$, $E(T_1) = \frac{n-1}{n}\sigma^2$ and $E(T_2) = \sigma^2$. Thus,

T_2 is unbiased and T_1 is not. However, compare their MSE.

$$\begin{aligned}
\text{MSE}(T_1) &= E(T_1 - \sigma^2)^2 = Var(T_1) + \text{Bias}^2 \\
&= Var\left(\frac{S^2}{n}\right) + \left(E(\frac{S^2}{n}) - \sigma^2\right)^2 \\
&= \frac{1}{n^2} Var(S^2) + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 \\
&= \frac{1}{n^2} 2(n-1)\sigma^4 + \frac{1}{n^2}\sigma^4 = \sigma^4 \left[\frac{2(n-1)}{n^2} + \frac{1}{n^2}\right] \\
&= \sigma^4 \left(\frac{2n-1)}{n^2}\right), \text{ and} \\
\text{MSE}(T_2) &= Var(T_2) = Var\left(\frac{S^2}{n-1}\right) \\
&= \frac{1}{(n-1)^2} 2(n-1)\sigma^4 = \sigma^4 \left(\frac{2}{n-1}\right).
\end{aligned}$$

Note that,

$$\begin{aligned}
\frac{2}{n-1} - \frac{2n-1)}{n^2} &= \frac{2n^2 - (2n-1)(n-1)}{n^2(n-1)} \\
&= \frac{2n^2 - \{2n^2 - 2n - n + 1)\}}{n^2(n-1)} = \frac{3n-1}{n^2(n-1)} > 0.
\end{aligned}$$

Thus, T_1 has a smaller MSE than T_2 for all σ^2 even though it is not unbiased. T_2 is often preferred because T_1 can underestimate σ^2 which is undesirable.

How does one derive the UMVUE when it exists?

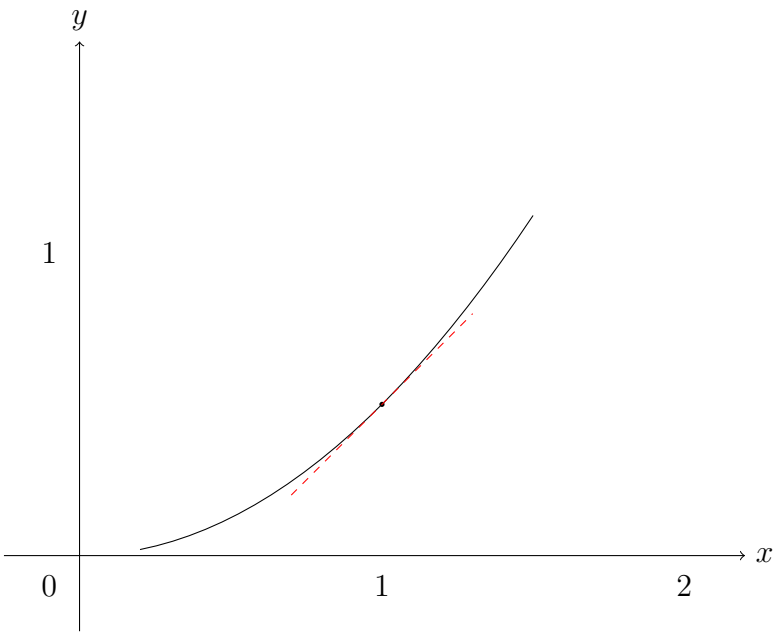
Let $\mathcal{T}$ = set of all unbiased estimators of $q(\theta)$. i.e.,

$$\mathcal{T} = \{T(X) : E_\theta(T(X)) = q(\theta) \qquad \forall \theta \in \Theta\}.$$

Then $\text{MSE}_\theta(T) = R(\theta, T) = Var_\theta(T)$ for $T \in \mathcal{T}$. Sometimes it is possible to find $T^* \in \mathcal{T}$ such that $Var_\theta(T^*) \leq Var_\theta(T)$ for all θ and all $T \in \mathcal{T}$. Then T^* is called the UMVUE of $q(\theta)$. One method of deriving this is by using the *Rao-Blackwell* Theorem. Some basic results in mathematics must precede it.

Result. Let ϕ be a convex function defined on (a, b) and let $a < t < b$. Then there exists a line
$y = L(x) = \phi(t) + c(x - t)$ passing through $(t, \phi(t))$ such that

$$(*) \qquad L(x) \leq \phi(x) \qquad \forall x \in (a, b).$$



Jensen's Inequality. If ϕ is a convex function defined on $I = (a, b)$ and X is a random variable such that $P(X \in I) = 1$ and $E(|X|) < \infty$, then

$$(**) \qquad \phi(E(X)) \leq E(\phi(X)).$$

If ϕ is strictly convex, the inequality is strict unless X is degenerate.

Proof. Let $y = L(x)$ be as in (*) for which $L(t) = \phi(t)$ when $t = E(X)$. Then, from (*),

$$(***) \qquad E(\phi(X)) \geq E(L(X)) = L(E(X)) = L(t) = \phi(t) = \phi(E(X)).$$

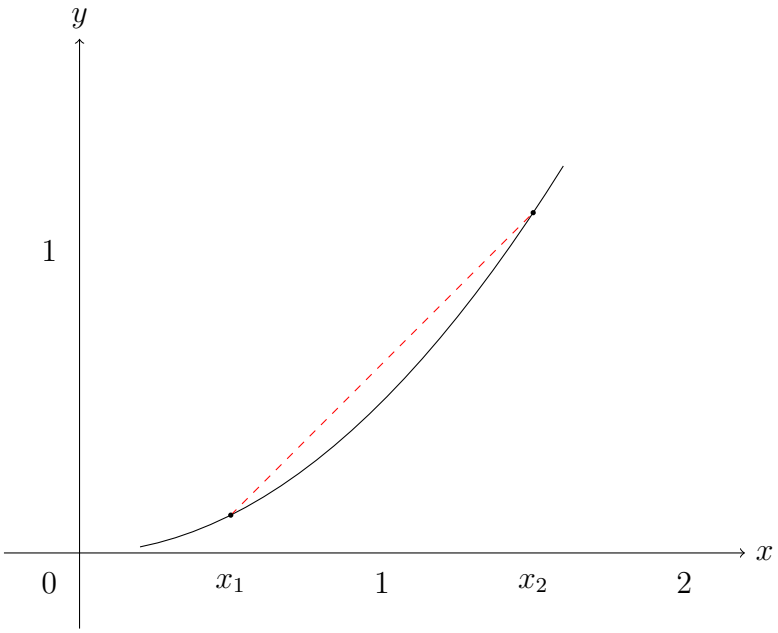
If ϕ is strictly convex, then the inequality in (*) is strict for all $x \neq t$, so inequality in (***) is strict unless $\phi(X) = E(\phi(X))$ w.p. 1. The proof can be extended to random vectors giving the following version of Jensen's Inequality, which will be used here.

Jensen's Inequality. If ϕ is a convex real-valued function defined on a non-empty convex set $S \subset \mathcal{R}^k$ and $\mathbf{Z}$ is a random vector with $E(||\mathbf{Z}||^2) < \infty$ and $P(Z \in S) = 1$, then $E(Z) \in S$ and

$$\phi\left(E(\mathbf{Z})\right) \leq E\left(\phi(\mathbf{Z})\right),$$

the inequality being strict if ϕ is strictly convex and Z is not degenerate.

Note: S is a convex set means, $x, y \in S$ implies $\alpha x + (1-\alpha)y \in S$, for $0 < \alpha < 1$.
ϕ is a convex function means, $\phi(\alpha x + (1-\alpha)y) \leq \alpha\phi(x) + (1-\alpha)\phi(y)$. Refer to *Rudin: Real and Complex Analysis* for the above discussion.



Theorem (Rao-Blackwell). Let $\mathbf{X}$ be a random vector with distribution P_θ, $\theta \in \Theta$, and let T be sufficient for θ. Let $\delta(X)$ be an estimator of θ and $\delta^*(t) = E\left[\delta(X)|T = t\right]$. Let $L(\theta, d)$ be a strictly convex loss function (in d) and $R(\theta, d) = E\left[d(X)|T = t\right]$. Then, if $R(\theta, \delta) = E\left[L(\theta, \delta(X))\right] < \infty$, we obtain

$$R(\theta, \delta^*) < R(\theta, \delta), \text{ for all } \theta\ ,$$

unless $\delta(x) = \delta^*(T(x))$ w.p.1.

Proof. Fix θ and define $\phi(d) = L(\theta, d)$. Then it is given that ϕ is strictly convex. Therfore,

$$L(\theta, \delta^*(t)) = \phi\left(E\left[\delta(X)|T=t\right]\right) < E\left[\phi\left(\delta(X)\right)|T=t\right] = E\left[L(\theta, \delta(X))|T=t\right].$$

Taking expectations on both sides w.r.t the distribution of T, we get,

$$E\left[L(\theta, \delta^*(T))\right] < E\left[E\left\{L(\theta, \delta(X))|T\right\}\right], \text{ i.e.,}$$
$$R(\theta, \delta^*) < E\left[L(\theta, \delta(X))\right] = R(\theta, \delta).$$

Note 1. δ^* is an estimator for θ since $E_\theta\left[\delta(X)|T=t\right]$ is free of θ, and depends on t only.

2. R-B says that any estimator can be improved by conditionally averaging with respect to T if the loss is convex (since T has all the information about θ). Therefore, an estimate depending on T is as good as $\delta(X)$. Given $\delta(x)$, get $\delta^*(t)$ by averaging on partitioning sets.

3. Corollary. Let $\hat{\theta}(X)$ be an unbiased estimator of θ. Then $\hat{\theta}^*(T) = E\left[\hat{\theta}(X)|T\right]$ is also unbiased and has smaller variance than $\hat{\theta}(X)$ for all θ if $Var(\hat{\theta}^*(T)) < \infty$.

Proof. $L(\theta, d) = (\theta - d)^2$. Therefore, $R(\theta, d(X)) = E(\theta - d(X))^2 = Var(d(X))$ if $E_\theta(d(X)) = \theta$. Since L is strictly convex, $R(\theta, \hat{\theta}^*) < R(\theta, \hat{\theta})$ if $\hat{\theta}^*(t) = E\left[\hat{\theta}(X)|T=t\right]$. Further, $E_\theta\left[\hat{\theta}^*(T)\right] = E\left\{E\left[\hat{\theta}^(X)|T\right]\right\} = E\left[\hat{\theta}(X)\right]$ $= \theta$. Therefore $\hat{\theta}^*$ is unbiased and hence $R(\theta, \hat{\theta}^*) = E\left[\hat{\theta}^* - \theta\right]^2 = Var(\hat{\theta}^*)$.

An alternative proof of R-B without using the Jensen's inequality exists and is as follows.

Theorem (Rao-Blackwell), another version. If T is an unbiased estimate of $\tau(\theta)$ and S is a sufficient statistic, the $T' = E(T|S)$ is also unbiased for $\tau(\theta)$ and

$$\text{Var}(T'|\theta) \leq \text{Var}(T|\theta)\ \forall\theta.$$

Proof. By the property of conditional expectations,

$$E\left(T'|\theta\right) = E\left\{E(T|S) \mid \theta\right\} = E\left(T|\theta\right).$$

(You may want to verify this at least for the discrete case.) Also,

$$\begin{aligned}\text{Var}(T|\theta) &= E\left[\left\{(T-T') + (T' - \tau(\theta))\right\}^2 \mid \theta\right] \\ &= E\left\{(T-T')^2 \mid \theta\right\} + E\left\{(T' - \tau(\theta))^2 \mid \theta\right\},\end{aligned}$$

because

$$\begin{aligned}
\mathrm{Cov}\left\{T - T', T' - \tau(\theta) \mid \theta\right\} &= E\left\{(T - T')(T' - \tau(\theta) \mid \theta\right\} \\
&= E\left[E\left\{(T' - \tau(\theta))(T - T') \mid S\right\} |\theta\right] \\
&= E\left[(T' - \tau(\theta))E(T - T'|S) \mid \theta\right] \\
&= 0.
\end{aligned}$$

The decomposition of $\mathrm{Var}(T|\theta)$ above shows that it is greater than or equal to $\mathrm{Var}(T'|\theta)$.

ANOVA Formula

If Z and Y are jointly distributed (with finite second moments), then

$$E(Y) = E\left[E(Y|Z)\right],$$

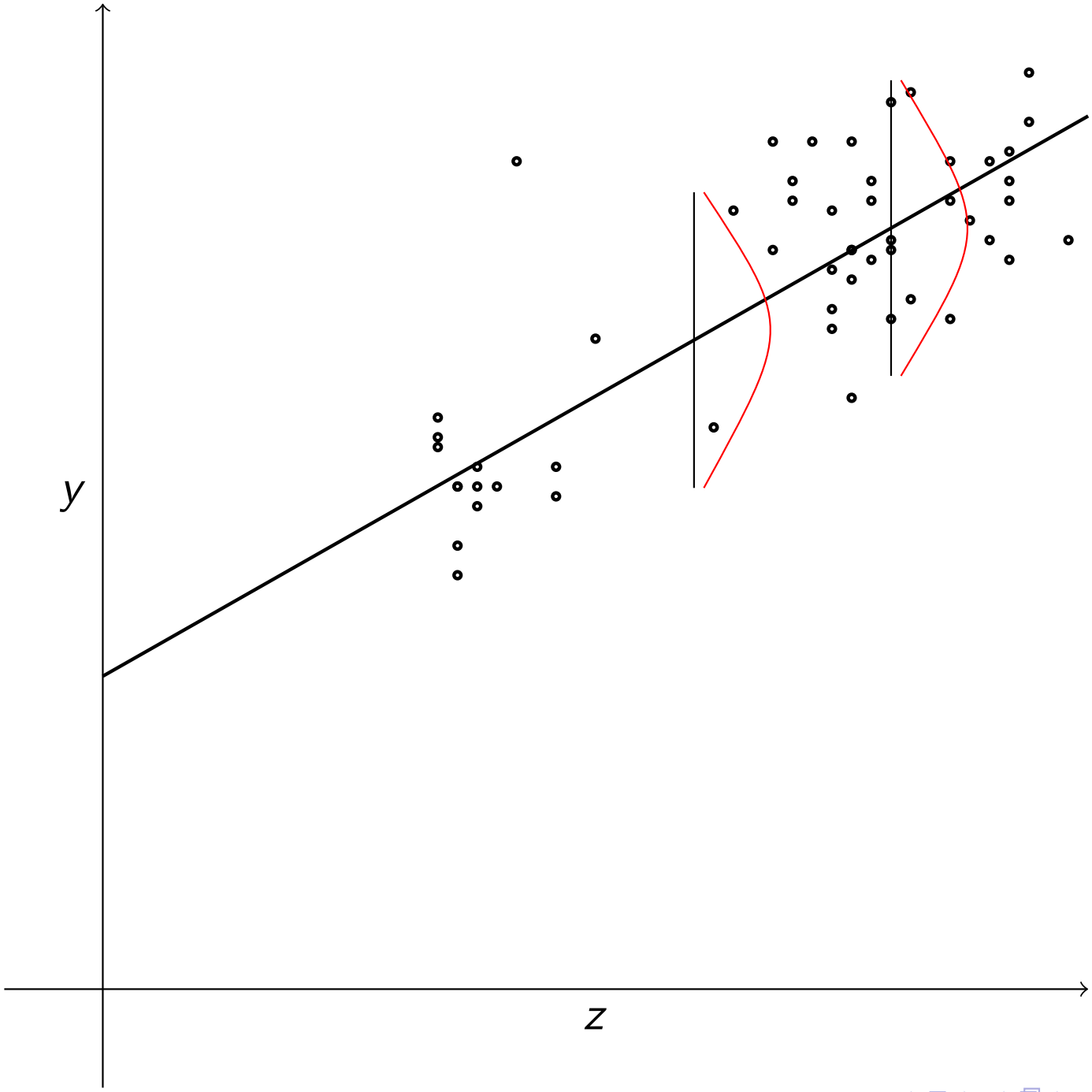
$$Var(Y) = E\left[Var(Y|Z)\right] + Var\left[E(Y|Z)\right] \geq Var\left[E(Y|Z)\right].$$

The first term on RHS is the 'within variation': if Y is partitioned according to values of Z, how much is left to be explained in Y for given Z. The second term is the variation between $\hat{Y}(Z)$ values, and is the 'between variation'. In a study, $Var(Y)$ may be large, but if $Var(Y|Z)$ is small, it makes sense to use Z to predict Y using Z. This result is known as the Analysis of Variance formula, and the ANOVA for regression is based on it.

z = duration and y = interval (both in minutes) for eruptions of Old Faithful Geyser

z	y	z	y	z	y	z	y	z	y	z	y
4.4	78	3.9	74	4.0	68	4.0	76	3.5	80	4.1	84
2.3	50	4.7	93	1.7	55	4.9	76	1.7	58	4.6	74
3.4	75	4.3	80	1.7	56	3.9	80	3.7	69	3.1	57
4.0	90	1.8	42	4.1	91	1.8	51	3.2	79	1.9	53
4.6	82	2.0	51	4.5	76	3.9	82	4.3	84	2.3	53
3.8	86	1.9	51	4.6	85	1.8	45	4.7	88	1.8	51
4.6	80	1.9	49	3.5	82	4.0	75	3.7	73	3.7	67
4.3	68	3.6	86	3.8	72	3.8	75	3.8	75	2.5	66
4.5	84	4.1	70	3.7	79	3.8	60	3.4	86		

Table: Eruptions of Old Faithful Geyser, August 1 – 4, 1978


y
z

Definition A statistic T or its distribution $\{P_\theta, \theta \in \Theta\}$ is said to be (boundedly) complete if for any real valued (bounded) function $h(T)$ with $E(|h(T)|) < \infty$,

$$E_\theta h(T) = 0 \; \forall \; \theta \text{ implies } h(T) = 0$$

(with probability one under all θ).

Suppose T is discrete. The condition then simply means the family of p.m.f.'s $f_T(t|\theta)$ of T is rich enough that there is no non-zero $h(t)$ that is orthogonal to $f^T(t|\theta)$ for all θ in the sense $\sum_t h(t) f_T(t|\theta) = 0$ for all θ. In general, T is complete iff $h(T) \equiv 0$ is the only unbiased estimator of 0.

Complete implies boundedly complete.

Example. Let $X_1, \ldots, X_n$ be i.i.d Bernoulli(p), $0 < p < 1$. $E(X_1 - X_2) = 0$ for all p, so $\mathbf{X}$ is not complete (but sufficient). $T = \sum_{i=1}^n X_i$ is minimal sufficient for p, and $T \sim$ Binomial(n, p).
Claim: T is complete.
Suppose $E_p h(T) = 0$ for all $p \in (0, 1)$. i.e.,

$$\begin{aligned} \sum_{t=0}^{n} h(t) \binom{n}{t} p^t (1-p)^{n-t} &= 0 \; \forall \; p \in (0,1), \text{ or} \\ h(0)(1-p)^n + \sum_{t=1}^{n} h(t) \binom{n}{t} p^t (1-p)^{n-t} &= 0 \; \forall \; p \in (0,1). \end{aligned}$$

As $p \to 0$, LHS $\to h(0)$ and RHS $= 0$. Therefore, $h(0) = 0$. Hence,

$$\sum_{t=1}^{n} h(t) \binom{n}{t} p^t (1-p)^{n-t} = 0 \; \forall \; p \in (0,1).$$

Thus we get $h(1) = 0 = h(2)$ and finally $h(n) = 0$.

Example. Let $X_1, \ldots, X_n$ be i.i.d $N(\theta, \theta^2)$, $\theta > 0$. Then

$$\begin{aligned} f(x_1, \ldots, x_n|\theta) &= (2\pi)^{-n/2} \theta^{-n} \exp\left(-\frac{1}{2\theta^2} \sum_{i=1}^{n} (x_i - \theta)^2 \right) \\ &= (2\pi)^{-n/2} \theta^{-n} \exp\left(-\frac{1}{2\theta^2} \left[\sum_{i=1}^{n} x_i^2 - 2\theta \sum_{i=1}^{n} x_i + n\theta^2 \right] \right) \\ &= (2\pi)^{-n/2} \theta^{-n} \exp\left(\frac{1}{\theta} \sum_{i=1}^{n} x_i - \frac{1}{2\theta^2} \sum_{i=1}^{n} x_i^2 - \frac{n}{2} \right). \end{aligned}$$

Thus, we see that, $T(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is (minimal) sufficient for θ (even though θ is now one-dimensional).
Claim: $T = (T_1, T_2)$ is not complete.
Note that $h(t_1, t_2) = t_2 - \frac{2}{n+1}t_1^2$ is not the 0 function, but $E_\theta[T_2 - T_1^2] = 0$ for all θ. To see this, observe, $T_1 = \sum_{i=1}^n X_i \sim N(n\theta, n\theta^2)$, so

$$E_\theta(T_1^2) = E_\theta\left[\sum_{i=1}^n X_i\right]^2 = n\theta^2 + (n\theta)^2 = n(n+1)\theta^2.$$

Also,

$$E_\theta(T_2) = E_\theta\left[\sum_{i=1}^n X_i^2\right] = \sum_{i=1}^n E_\theta(X_i^2) = n(\theta^2 + \theta^2) = 2n\theta^2,$$

and hence

$$E_\theta\left[(\frac{2}{n+1})T_1^2\right] = \frac{2n(n+1)}{n+1}\theta^2 = 2n\theta^2 = E_\theta(T_2).$$

Theorem. Let $\mathbf{X}$ have distribution $P_\theta, \theta \in \Theta$ and let $T = T(\mathbf{X})$ be complete sufficient for θ (or $P_\theta, \theta \in \Theta$). Then every function $h(T)$ is the unique unbiased estimate of its own expected value. i.e., for any h, if $q(\theta) = E_\theta h(T)$, then $h(T)$ is the only unbiased estimate available for $q(\theta)$.

Proof. Suppose $h_1(T)$ and $h_2(T)$ are both unbiased estimates of a parametric function $\psi(\theta)$. Then

$$E_\theta[h_1(T) - h_2(T)] = 0 \; \forall \; \theta \in \Theta.$$

i.e., if we let $h^*(T) = h_1(T) - h_2(T)$, then

$$E_\theta[h^*(T)] = 0 \; \forall \; \theta \in \Theta.$$

Since T is complete, $h^*(T) \equiv 0$. i.e. $h_1(T) = h_2(T)$.

Theorem (Lehmann-Scheffe). Suppose $T(\mathbf{X})$ is complete sufficient (for $P_\theta, \theta \in \Theta$) and $S(\mathbf{X})$ is any unbiased estimate of $q(\theta)$. Then $T^*(\mathbf{X}) = E(S(\mathbf{X})|T)$ is the unique UMVUE of $q(\theta)$ if $Var_\theta(T^*(\mathbf{X})) < \infty$ for all θ.

Proof. Both S and T^* are unbiased, so that their MSE is the respective variance. By the Rao-Blackwell theorem, $Var_\theta(T^*) < Var_\theta(S)$ unless $S = T^*$. To show uniqueness, we show that T^* is the same, whichever S we start with, so that T^* cannot be improved upon further.

Let S_1 and S_2 be two unbiased estimators of $q(\theta)$, and let $g_1(T) = E(S_1(X)|T)$ and $g_2(T) = E(S_2(X)|T)$. But then,

$$E(g_1(T)) = E(S_1) = q(\theta) = E(S_2) = E(g_2(T)),\ \forall\ \theta \in \Theta.$$

Since T is complete, there can be only one unbiased estimate of $q(\theta)$ based on T. Therefore $g_1 \equiv g_2$.

Note. 1. Given any $S(X)$ unbiased for $q(\theta)$, UMVUE is found by obtaining $T^*(X) = E(S(X)|T)$ where T is complete sufficient.

2. If we already have $h(T)$ unbiased for $q(\theta)$ and T is complete sufficient, then $h(T)$ is UMVUE, since $T^* = E(h(T)|T) = h(T)$.

Remark. The idea behind the L-S method is that, conditioning on a sufficient statistic (possibly) improves the estimator (R-B), and conditioning on a complete sufficient statistic gives the most possible improvement.

Example. Let $X_1, \ldots, X_n$ be i.i.d Poisson(λ), $\lambda > 0$. Find the UMVUE of $q(\lambda) = 1 - \exp(-\lambda)$. Note that $q(\lambda) = 1 - \exp(-\lambda) = P_\lambda(X_1 > 0)$. Therefore,

$$S(X_1, \ldots, X_n) = I(X_1 > 0) = \begin{cases} 1 & \text{if } X_1 \geq 1; \\ 0 & \text{if } X_1 = 0 \end{cases}$$

is an unbiased estimator of $q(\lambda)$. Also, $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is complete sufficient for λ. (Check this.) Therefore,

$$h(T) = E(S(\mathbf{X})|T(\mathbf{X})) = E(I(X_1 > 0)|\sum_{i=1}^n X_i) = P(X_1 > 0|\sum_{i=1}^n X_i)$$

is the unique UMVUE. We need the conditional distribution of $X_1|\sum_{i=1}^n X_i$. Note that $X_1|(\sum_{i=1}^n X_i = t) \sim$ Binomial$(t, \frac{1}{n})$. (This is from the fact that the conditional joint distribution of the X_i's is a multinomial, as shown previously.) Therefore

$$h(t) = P(X_1 > 0|\sum_{i=1}^n X_i = t) = 1 - \binom{t}{0}\left(\frac{1}{n}\right)^0\left(\frac{n-1}{n}\right)^t = 1 - \left(\frac{n-1}{n}\right)^t.$$

Thus, $1 - \left(\frac{n-1}{n}\right)^T$ is the UMVUE. (How does one show directly that this is unbiased?)

Example. Let $X_1, \ldots, X_n$ be i.i.d Bernoulli(p), $0 < p < 1$. Find UMVUE of p.
(i) Consider $S(X_1, \ldots, X_n) = X_1$. Then $E(X_1) = p$ for all $0 < p < 1$, so that

S is an unbiased estimate of p. Therefore, the UMVUE of p is $E(X_1|\sum_{i=1}^n X_i)$ since $\sum_{i=1}^n X_i$ is complete sufficient. Now see that

$$E(X_1|\sum_{i=1}^n X_i) = E(X_2|\sum_{i=1}^n X_i) = \cdots = E(X_n|\sum_{i=1}^n X_i).$$

Therefore,

$$E(X_1|\sum_{i=1}^n X_i) = \frac{1}{n}\sum_{j=1}^n E(X_j|\sum_{i=1}^n X_i) = \frac{1}{n}E\left(\sum_{j=1}^n X_j|\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n X_i.$$

(ii) $\sum_{i=1}^n X_i$ is complete sufficient. Also, $E(\sum_{i=1}^n X_i) = np$, so that $h(\sum_{i=1}^n X_i) = \frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased estimate of p depending on the complete sufficient statistics only; thus unique UMVUE.

Theorem. Let $P_\theta, \theta \in \Theta$ be a k-parameter exponential family with density $f(\mathbf{x}|\theta) = \exp\left(\sum_{j=1}^k c_j(\theta)T_j(\mathbf{x}) + d(\theta) + S(\mathbf{x})\right) I_A(\mathbf{x})$. Suppose $\{\mathbf{c}(\theta) = (c_1(\theta), \ldots, c_k(\theta), \theta \in \Theta\}$ contains an open set (open rectangle) in $\mathcal{R}^k$. (This property is called full-rank.) Then $\mathbf{T}(\mathbf{X}) = (T_1, \ldots, T_k)$ is complete sufficient.

Proof. See Lehmann: Testing Statistical Hypotheses for a proof involving uniquness of Laplace transforms. We will use the result.

Example. Let $X_1, \ldots, X_n$ be i.i.d $N(\mu, \sigma^2)$, $-\infty < \mu < \infty$, $\sigma^2 > 0$. $T(\mathbf{X}) = (\bar{X}, \sum_{i=1}^n (X_i - \bar{X})^2)$ is sufficient. Since the set of $c(\theta) = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ for $-\infty < \mu < \infty$, $\sigma^2 > 0$ is $\mathcal{R}^1 \times \mathcal{R}^+$, an open set, $T(\mathbf{X})$ is complete also.

Therefore, since $E(\bar{X}) = \mu$ and $\bar{X} = h(T(\mathbf{X}))$, we obtain that $\bar{X}$ is the unique UMVUE of μ.

Further, $\sum_{i=1}^n (X_i - \bar{X})^2) \sim \sigma^2 \chi^2_{n-1}$, and hence

$$E\left[\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2\right] = \sigma^2, \ \forall \ (\mu, \sigma^2).$$

Therefore, $\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$, which is a function of $T(\mathbf{X})$ alone, is the unique UMVUE of σ^2.

Suppose we want to estimate $q(\mu, \sigma^2) = \mu/\sigma$. Then, for $n \geq 3$, the UMVUE of $q(\mu, \sigma^2)$ is

$$c(n)\frac{\bar{X}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}},$$

where $c(n)$ may be found from

$$E\left[\frac{\bar{X}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}\right] = E(\bar{X})E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right]^{-1/2},$$

where $E(\bar{X}) = \mu$ and $\sum_{i=1}^n (X_i - \bar{X})^2/\sigma^2 \sim \Gamma(1/2, (n-1)/2)$.

Example. Let $X_1, \ldots, X_n$ be i.i.d $N(\theta, \theta^2)$, $\theta > 0$. Then $T(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is (minimal) sufficient for θ, but it is not complete. Therefore it is not possible to apply L-S to find the UMVUE of θ.

There is an interesting technical theorem, due to D. Basu, which establishes independence of a sufficient statistic and an ancillary statistic. The result is useful in many calculations. Recall that parts of a minimal sufficient statistic may be ancillary, so conditions are needed for this to happen.

Theorem (Basu). Suppose T is complete sufficient for $\{P_\theta, \theta \in \Theta\}$. Let S be any ancillary statistic. Then T and S are independent for all θ.

Proof. Because T is sufficient, the conditional probability of S being in some set B given T is free of θ and may be written as $P_\theta(S \in B|T) = \phi(T)$. Since S is ancillary, $E_\theta(\phi(T)) = P_\theta(S \in B) = c$, where c is a constant. Consider a B for which $0 < c < 1$. Let $\psi(T) = \phi(T) - c$. Then $E_\theta\psi(T) = 0$ for all θ, implying $\psi(T) = 0$ (with probability one), i.e., $P_\theta(S \in B|T) = P_\theta(S \in B)$.

Alternatively, let $f(t, s|\theta)$ be the joint density of (T, S) and $F_S(s)$ be the cdf of S. Then F_S is free of θ since S is ancillary. Let $f_T(t|\theta)$ be the density of T and $F_{S|T}(s)$ be the conditional cdf of $S|T$ (which is again free of θ since T is sufficient). Note that, for any s,

$$\begin{aligned}
\int F_S(s) f_T(t|\theta)\, dt &= F_S(s) \int f_T(t|\theta)\, dt = F_S(s). \\
\int F_{S|T=t}(s) f_T(t|\theta)\, dt &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{s} f_{S|T=t}(u)\, du \right] f_T(t|\theta)\, dt \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{s} f_{S|T=t}(u)\, f_T(t|\theta)\, du dt \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{s} f_{S,T}(u, t)\, du dt \\
&= \int_{-\infty}^{s} \left(\int_{-\infty}^{\infty} f_{S,T}(u, t)\, dt \right) du \\
&= \int_{-\infty}^{s} f_S(u)\, du = F_S(s).
\end{aligned}$$

Therefore,

$$\int_{-\infty}^{\infty} \left[F_S(s) - F_{S|T=t}(s) \right] f_T(t|\theta)\, dt = 0, \ \forall\ \theta.$$

Fix s and let $h(T) = F_S(s) - F_{S|T=t}(s)$, which involves T but is totally free of θ. Then $E_\theta h(T) = 0$ for all θ. Since T is complete, $h \equiv 0$. That means, $F_{S|T} \equiv F_S$, implying independence of S and T for all θ.

Example. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. Then $\bar{X}$ and $S^2 = \sum(X_i - \bar{X})^2$ are independent. To prove this, treat σ^2 as fixed to start with and μ as the parameter. Then $\bar{X}$ is complete sufficient and $S^2 = \sum(X_i - \bar{X})^2 = \sum \left[(X_i - \mu) - (\bar{X} - \mu) \right]^2 = \sum(Z_i - \bar{Z})^2$ is ancillary. Hence $\bar{X}$ and S^2 are independent for each σ^2 by Basu's theorem.

Example. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d $U(\theta_1, \theta_2)$. Then for any $1 < r < n$, $Y = (X_{(r)} - X_{(1)})/(X_{(n)} - X_{(1)})$ is independent of $(X_{(1)}, X_{(n)})$. This follows

because Y is ancillary.

It is shown below that a complete sufficient statistic is minimal sufficient. In general, the converse isn't true. Also, technically, neither may exist or only one of them may exist.

Theorem. A (bdd) complete sufficient statistic is minimal sufficient, assuming minimal sufficient statistic exists.

Proof. Let T be minimal sufficient and U be complete sufficient. Then $T = h(U)$ for some function h since minimal sufficiency provides coarser partition. We need to show that T and U are equivalent statistics (i.e., produce the same partition). It is enough to show that for all (integrable) ψ,

$$E\left[\psi(U)|T\right] = \psi(U).$$

(Note that $T = h(U)$ and the above requirement is simply that averaging $\psi(U)$ where $h(U)$ is fixed, reproduces $\psi(U)$.)

Suppose not. That is, let there exist ψ such that $E\left[\psi(U)|h(U)\right]$ is not identical to $\psi(U)$. Define

$$k(U) = \psi(U) - E\left[\psi(U)|h(U)\right].$$

Then $k \neq 0$ but

$$\begin{aligned} E\left[k(U)\right] &= E\left\{\psi(U) - E\left[\psi(U)|h(U)\right]\right\} \\ &= E(\psi(U)) - E\left\{E\left[\psi(U)|h(U)\right]\right\} \\ &= E(\psi(U)) - E\left\{\psi(U)\right\} = 0. \end{aligned}$$

However, U is complete!

Corollary. Let $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ be a k-parameter exponential family with density

$$f(\mathbf{x}|\theta) = \exp(\left(\sum_{i=1}^{k} c_i(\theta)T_i(\mathbf{x} + d(\theta) + S(\mathbf{x})\right) I_A(\mathbf{x}),$$

where $C = \{(c_1(\theta, \ldots, c_k(\theta)) : \theta \in \Theta\}$ contains an open set. Then $(T_1, \ldots, T_k)$ is minimal sufficient.

Proof. $(T_1, \ldots, T_k)$ is complete sufficient as remarked previously, hence minimal sufficient.

Information contained in an experiment

It is of interest to know how informative is an experiment about the unknown parameters. Binomial and negative binomial sampling provide different amount of information depending on how large or small p is. In *Information Theory*, Shannon information is mostly used, which is a measure of entropy or randomness, but in statistics different measures are used. The notion that is described and used here is based on 'the difference that we see when we change the model continuously from one to another'.

Information number (Fisher). Let $\{P_\theta, \theta \in \Theta\}$ be a family of probability distributions satisfying the following mathematical regularity conditions.

(A) $A = \{x : f(x|\theta) > 0\}$ does not depend on θ. For all $x \in A$ and $\theta \in \Theta$, the score function,

$$S(x) = \frac{\partial}{\partial\theta} \log f(x|\theta) = \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}$$

exists and is finite.

$S(x)$ measures the relative rate at which $f(x|\theta)$ changes at x. Since x varies (due to X being random) this needs averaging.

$$\begin{aligned} I(\theta) &= E_\theta \left[\frac{\partial}{\partial\theta} \log f(X|\theta) \right]^2 \\ &= \int \left(\frac{\partial}{\partial\theta} \log f(x|\theta) \right)^2 f(x|\theta)\, dx \end{aligned}$$

is called the Fisher Information number of θ contained in $f(.|\theta)$ or P_θ. Clearly, $0 \leq I(\theta) \leq \infty$. To get a feeling for $I(\theta)$, consider an extreme case where $f(x|\theta)$ is free of θ. Clearly, in this case there can be no information about θ in $\mathbf{X}$. On the other hand, if $I(\theta)$ is large, then on an average a small change in θ leads to a big change in $\log f(x|\theta)$, i.e., f depends strongly on θ and one expects there is a lot that can be learned about θ.

Example. $X \sim$ Bernoulli(p). i.e., how much information is there in a single

toss of a coin on its success probability, p?

$$\begin{aligned} f(x|p) &= p^x(1-p)^{1-x},\ x=0,1 \\ \log f(x|p) &= x\log p + (1-x)\log(1-p) \\ \frac{\partial}{\partial p}\log f(x|p) &= \frac{x}{p} - \frac{1-x}{1-p} = \frac{x-xp-p+xp}{p(1-p)} = \frac{x-p}{p(1-p)},\ \text{so} \\ I(p) &= E_p\left[\frac{\partial}{\partial p}\log f(X|p)\right]^2 = E_p\left[\frac{(X-p)^2}{p^2(1-p)^2}\right] \\ &= \frac{p(1-p)}{p^2(1-p)^2} = \frac{1}{p(1-p)} = \frac{1}{Var_p(X)}. \end{aligned}$$

In other words, Information is inversely proportional to the variance. Intuitively, if the variance is large, or if p is far away from 0 or 1, then one will need a large number of observations to get a reliable estimate of p. If p is close to 0 or 1, the observations will be mostly 0, or mostly 1, so that estimation is easy. On the other hand, if p is close to 1/2, there will be a lot of fluctuation, and much variability. Then it will be difficult to distinguish between models.

Theorem. If (A) holds, and
(B) the derivative with respect to θ of $\int f(x|\theta)\,dx$ can be obtained by differentiating under the integral sign, then
(i) $E_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right) = 0$, and
(ii) $I(\theta) = Var_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]$.
In addition, if
(C) the second derivative (w.r.t. θ) of $\log f(x|\theta)$ exists for all x and θ, and the second derivative of $\int f(x|\theta)\,dx$ can be obtained by differentiating twice under the integral sign, then
(iii) $I(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right]$.

(A)-(C) are called Cramer-Rao (C-R) regularity conditions.

Proof. (i). Since $\int_{-\infty}^{\infty} f(x|\theta)\,dx = 1$, we have,

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f(x|\theta)\,dx = \int\frac{\partial}{\partial\theta}f(x|\theta)\,dx = \int\left\{\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right\}f(x|\theta)\,dx \\ &= \int\left\{\frac{\partial}{\partial\theta}\log f(x|\theta)\right\}f(x|\theta)\,dx = E_\theta\left\{\frac{\partial}{\partial\theta}\log f(X|\theta)\right\}. \end{aligned}$$

(ii). Now, using this, we get,

$$\begin{aligned} I(\theta) &= E_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2 = E_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta) - E_\theta\left\{\frac{\partial}{\partial\theta}\log f(X|\theta)\right\}\right]^2 \\ &= Var_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]. \end{aligned}$$

(iii). To obtain this alternative formula, note that,

$$\begin{aligned} \frac{\partial^2}{\partial\theta^2}\log f(x|\theta) &= \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] = \frac{\partial}{\partial\theta}\left[\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right] \\ &= \frac{\frac{\partial^2}{\partial^2\theta}f(x|\theta)}{f(x|\theta)} - \frac{\left(\frac{\partial}{\partial\theta}f(x|\theta)\right)^2}{\left[f(x|\theta)\right]^2}. \end{aligned}$$

Therefore,

$$\begin{aligned} E_\theta\left[\frac{\partial^2}{\partial^2\theta}\log f(X|\theta)\right] &= E_\theta\left\{\frac{\frac{\partial^2}{\partial^2\theta}f(X|\theta)}{f(X|\theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(X|\theta)}{f(X|\theta)}\right)^2\right\} \\ &= \int \frac{\frac{\partial^2}{\partial^2\theta}f(x|\theta)}{f(x|\theta)} f(x|\theta)\,dx - E_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2 \\ &= \int \frac{\partial^2}{\partial^2\theta} f(x|\theta)\,dx - E_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2 \\ &= 0 - I(\theta), \end{aligned}$$

since

$$0 = \frac{\partial^2}{\partial\theta^2}\int_{-\infty}^{\infty} f(x|\theta)\,dx = \int \frac{\partial^2}{\partial\theta^2} f(x|\theta)\,dx.$$

The condition (A) that the support of $f(\cdot|\theta)$ is free of θ is essential. The exponential families satisfy these regularity conditions. Location-scale families may or may not satisfy, usually the critical assumption is that relating to the support of f. Thus the Cauchy location-scale family satisfies these conditions but not the uniform or the exponential density

$$f(x|\mu,\sigma) = \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right),\ x > \mu.$$

Theorem. Let X and Y be independently distributed random observables with density $f_1(x|\theta)$ and $f_2(y|\theta)$. If $I(\theta)$, $I_1(\theta)$ and $I_2(\theta)$ are the information numbers about θ contained in (X, Y), X and Y, respectively, then

$$I(\theta) = I_1(\theta) + I_2(\theta).$$

Proof. To see this additive property, note that,

$$\begin{aligned} I(\theta) &= Var_\theta \left[\frac{\partial}{\partial \theta} \log f(X, Y|\theta) \right] = Var_\theta \left[\frac{\partial}{\partial \theta} \log \{ f_1(X|\theta) f_2(Y|\theta) \} \right] \\ &= Var_\theta \left[\frac{\partial}{\partial \theta} \log f_1(X|\theta) + \frac{\partial}{\partial \theta} \log f_2(Y|\theta) \right] \\ &= Var_\theta \left[\frac{\partial}{\partial \theta} \log f_1(X|\theta) \right] + Var_\theta \left[\frac{\partial}{\partial \theta} \log f_2(Y|\theta) \right] \\ &= I_1(\theta) + I_2(\theta). \end{aligned}$$

Example. Let $X_1, \ldots, X_n$ be i.i.d Bernoulli(p). Then the information in the sample is $I(p) = \frac{n}{p(1-p)}$ since the information in each of the observations is $I_1(p) = \frac{1}{p(1-p)}$. Further, $\sum_{i=1}^n X_i \sim$ Binomial(n, p) is sufficient for p and has the same likelihood function as that of the sample. Thus Binomial(n, p) has the same $I(p)$ of $\frac{n}{p(1-p)}$.

For multi-parameter problems, one defines the Information Matrix,

$$I(\theta) = ((I_{ij}(\theta))), \text{ where } I_{ij}(\theta) = E \left[\frac{\partial}{\partial \theta_i} \log f(X|\theta) \frac{\partial}{\partial \theta_j} \log f(X|\theta) \right].$$

$I(\theta)$ depends on the particular parametrization chosen: Suppose $\eta = c(\theta)$, where $c(.)$ is one-one and differentaible. Then $\theta = h(\eta) = c^{-1}(\eta)$. Therefore, letting $f^*(x|\eta) = f(x|\theta)|_{\theta=h(\eta)}$,

$$\begin{aligned} I^*(\eta) &= E_\eta \left[\frac{\partial}{\partial \eta} \log f^*(X|\eta) \right]^2 \\ &= E_\eta \left[\frac{\partial}{\partial \theta} \log f(X|\theta) \, |_{\theta=h(\eta)} \, \frac{\partial}{\partial \eta} h(\eta) \right]^2 \\ &= E_\eta \left[\frac{\partial}{\partial \theta} \log f(X|\theta) \right]^2 |_{\theta=h(\eta)} \left(\frac{\partial}{\partial \eta} h(\eta) \right)^2 \\ &= I(\theta) \left(\frac{d\theta}{d\eta} \right)^2 |_{\theta=h(\eta)} \, . \end{aligned}$$

Example. Let $f(x|\alpha) = \alpha x \exp\left(-\frac{\alpha}{2}x^2\right)$, $x > 0$, $\alpha > 0$. What is $I(\alpha)$? Using the definition, since

$$\begin{aligned} \log f(x|\alpha) &= \log\alpha + \log x - \frac{\alpha}{2}x^2, \\ \frac{\partial}{\partial\alpha}\log f(x|\alpha) &= \frac{1}{\alpha} - \frac{x^2}{2}, \text{ so} \\ I(\alpha) &= E_\alpha\left[\left(\frac{X^2}{2} - \frac{1}{\alpha}\right)^2\right], \end{aligned}$$

which is difficult to compute, but using the alternative formula, we get,

$$\frac{\partial^2}{\partial\alpha^2}\log f(x|\alpha) = -\frac{1}{\alpha^2},$$

so that $I(\alpha) = 1/\alpha^2$. Sometimes, reparametrization can help too.

Example. Let $f(x|\theta) = \frac{x}{\theta^2}\exp\left(-\frac{x^2}{2\theta^2}\right)$, $x > 0$, $\theta > 0$. Find $I(\theta)$. We note

$$\begin{aligned} \log f(x|\theta) &= \log x - 2\log\theta - \frac{x^2}{2\theta^2}, \\ \frac{\partial}{\partial\theta}\log f(x|\theta) &= -\frac{2}{\theta} + \frac{x^2}{\theta^3} = \frac{1}{\theta^3}(x^2 - 2\theta^2), \text{ so} \\ I(\theta) &= \frac{1}{\theta^6}E_\theta\left(X^2 - 2\theta^2\right)^2, \end{aligned}$$

which is again difficult to compute. We may look at

$$\frac{\partial^2}{\partial\theta^2}\log f(x|\theta) = \frac{2}{\theta^2} - \frac{3x^2}{\theta^4} = -\frac{1}{\theta^4}(3x^2 - 2\theta^2),$$

and try to compute $I(\theta) = \frac{1}{\theta^4}E_\theta(3X^3 - 2\theta^2)$, which doesn't seem to simplify the calculation. Instead, let $\alpha = \alpha(\theta) = 1/\theta^2$. Then from the previous example, $I(\alpha) = 1/\alpha^2$. Therefore,

$$I^*(\theta) = I(\alpha(\theta))(\alpha'(\theta))^2 = \theta^4\frac{4}{\theta^6} = \frac{4}{\theta^2}.$$

Example. Let $X \sim N(\mu, \sigma^2)$. Then $I(\mu, \sigma^2) = ((I_{ij}(\mu, \sigma^2)))$, where

$$\begin{aligned} I_{11}(\mu,\sigma^2) &= E_{\mu,\sigma^2}\left[\frac{\partial}{\partial\mu}\log f(X|\mu,\sigma^2)\right]^2 \\ I_{22}(\mu,\sigma^2) &= E_{\mu,\sigma^2}\left[\frac{\partial}{\partial\sigma^2}\log f(X|\mu,\sigma^2)\right]^2 \\ I_{12}(\mu,\sigma^2) &= E_{\mu,\sigma^2}\left[\frac{\partial}{\partial\mu}\log f(X|\mu,\sigma^2)\frac{\partial}{\partial\sigma^2}\log f(X|\mu,\sigma^2)\right]. \end{aligned}$$

Since

$$\begin{aligned}
\log f(x|\mu,\sigma^2) &= -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(x-\mu)^2, \\
\frac{\partial}{\partial\mu}\log f(x|\mu,\sigma^2) &= -\frac{1}{2\sigma^2}2(x-\mu)(-1), \\
\frac{\partial}{\partial\sigma^2}\log f(x|\mu,\sigma^2) &= -\frac{1}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2,
\end{aligned}$$

we obtain

$$\begin{aligned}
I_{11}(\mu,\sigma^2) &= E_{\mu,\sigma^2}\left[\frac{(X-\mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^2}, \\
I_{22}(\mu,\sigma^2) &= \frac{1}{4\sigma^8}E_{\mu,\sigma^2}\left[(X-\mu)^2-\sigma^2\right]^2 = \frac{2\sigma^4}{4\sigma^8} = \frac{1}{2\sigma^4}. \\
I_{12}(\mu,\sigma^2) &= \frac{1}{2}E_{\mu,\sigma^2}\left[\left(\frac{X-\mu}{\sigma^2}\right)\left(\frac{(X-\mu)^2-\sigma^2}{\sigma^4}\right)\right] = 0.
\end{aligned}$$

Thus,

$$I(\mu,\sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

Theorem. Let X have one-parameter exponential family density

$$f(x|\theta) = \exp(c(\theta)T(x) + d(\theta) + S(x))I_A(x).$$

Consider the mean-value parametrization, $\delta(\theta) = E_\theta(T)$. Then

$$I(\delta) = \frac{1}{Var(T)}.$$

Poof. The natual parametrization, $\eta = c(\theta)$ gives

$$\begin{aligned} f^*(x|\eta) &= \exp(\eta T(x) + d_0(\eta) + S(x))I_A(x), \\ \log f^*(x|\eta) &= \eta T(x) + d_0(\eta) + S(x)), \\ \frac{\partial}{\partial \eta} \log f^*(x|\eta) &= T(x) + d_0'(\eta) = T(x) - E_\eta(T). \end{aligned}$$

Therefore,

$$I^*(\eta) = E_\eta \left(T - E_\eta(T)\right)^2 = Var_\eta(T).$$

Now, $\delta = E(T) = -d_0'(\eta) = h(\eta)$, so

$$\frac{d\eta}{d\delta} = \left(\frac{d\delta}{d\eta}\right)^{-1} = \left(-d_0''(\eta)\right)^{-1} = \frac{1}{Var(T)}.$$

Therefore,

$$\begin{aligned} I(\delta) &= I^*(\eta) \left(\frac{d\eta}{d\delta}\right)^2 \Big|_{\eta = h^{-1}(\delta)} \\ &= Var(T) \frac{1}{\left(Var(T)\right)^2}. \end{aligned}$$

Information Inequality (Cramer-Rao). Suppose the conditions (A) and (B) hold, and $0 < I(\theta) < \infty$. Let $T(X)$ be any statistic with $Var(T) < \infty$ and such that the derivative w.r.t. θ of

$$E_\theta(T) = \int T(x) f(x|\theta)\, dx$$

exists and can be obtained by differentiating under the integral sign. Then

$$Var_\theta(T(X)) \geq \frac{\left[\frac{d}{d\theta} E_\theta(T)\right]^2}{I(\theta)}.$$

Note. This is called the C-R lower bound on the variance of a statistic.

Proof. Note that

$$\begin{aligned}\frac{d}{d\theta}E_\theta(T) &= \frac{d}{d\theta}\int_A T(x)f(x|\theta)\,dx = \int_A T(x)\frac{\partial}{\partial\theta}f(x|\theta)\,dx \\ &= \int_A T(x)\left[\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right]f(x|\theta)\,dx = \int T(x)S(x)f(x|\theta)\,dx \\ &= E_\theta\left(T(X)S(X)\right),\end{aligned}$$

where $S(x) = \frac{\partial}{\partial\theta}\log f(x|\theta)$. Further,

$$\begin{aligned}E_\theta S(X) &= \int_A S(x)f(x|\theta)\,dx = \int_A \frac{\partial}{\partial\theta}\log f(x|\theta)f(x|\theta)\,dx \\ &= \int_A \left(\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right)f(x|\theta)\,dx = \int \frac{\partial}{\partial\theta}f(x|\theta)\,dx \\ &= \frac{d}{d\theta}\int f(x|\theta)\,dx = 0.\end{aligned}$$

Therefore,

$$\frac{d}{d\theta}E_\theta(T) = Cov_\theta(T(X), S(X)).$$

Since $|Cov(T(X), S(X))| \leq \sqrt{Var(T(X))Var(S(X))}$, and $Var_\theta(S(X)) = Var_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right) = I(\theta)$, we obtain

$$|\frac{d}{d\theta}E_\theta(T)| \leq \sqrt{Var_\theta(T)}\sqrt{I(\theta)}.$$

Therefore,

$$Var_\theta(T) \geq \frac{\left[\frac{d}{d\theta}E_\theta(T)\right]^2}{I(\theta)}.$$

Note. 1. For the class of all unbiased estimators of θ, we have $E_\theta(T) = \theta$ and $\frac{d}{d\theta}E_\theta(T) = 1$, so that

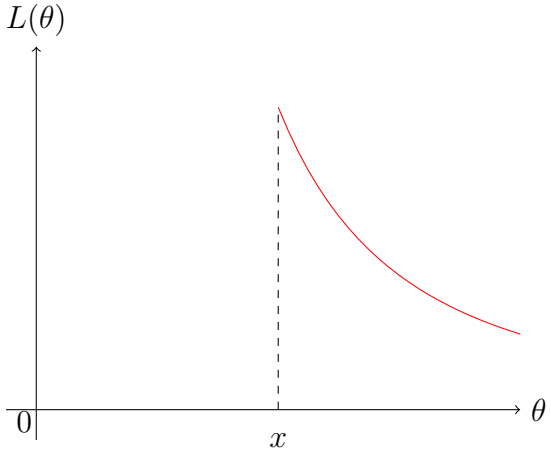
$$Var_\theta(T) \geq \frac{1}{I(\theta)}.$$

This lower bound is independent of any any particular T. Therefore, if there exists an unbiased estimator which attains this lower bound at all θ, it is the UMVUE.

Usually $T = T(X_1, \ldots, X_n) = T_n$, and $E_\theta(T_n) \to \theta$, which is called asymptotic unbiasedness. Then one would like to know what the asymptotic variance is, or whether it is the least it can be. MLE usually has this property.

Here is an example which shows that regularity conditions are indeed required.

Example. Let $X \sim U[0, \theta]$, $\theta > 0$. Then the likelihood function is as follows.



$$\begin{aligned} f(x|\theta) &= \frac{1}{\theta} \text{ for } x \le \theta, \\ \log f(x|\theta) &= -\log\theta \text{ for } x \le \theta, \\ \frac{\partial}{\partial\theta}\log f(x|\theta) &= \begin{cases} -\frac{1}{\theta} & \text{if } \theta > x; \\ \text{undefined} & \text{if } \theta = x; \end{cases} \end{aligned}$$

Since $P_\theta(X = \theta) = 0$,

$$\frac{\partial}{\partial\theta}\log f(X|\theta) = -\frac{1}{\theta}, \quad \text{w.p. 1 under } P_\theta.$$

Therefore,

$$\begin{aligned} E_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] &= -\frac{1}{\theta} \neq 0, \\ Var_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] &= 0, \\ E_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2 &= \frac{1}{\theta^2}. \end{aligned}$$

Then, what is $I(\theta)$? For unbiased estimators T of θ, do we have

$$Var_\theta(T) \geq \frac{1}{I(\theta)} = \theta^2 \text{ or } \infty?$$

Consider $T(X) = 2X$. Since $E(X) = \theta/2$, T is an unbiased estimator of θ. Note that $Var(T) = Var(2X) = 4Var(X) = 4\left(\frac{\theta^2}{12}\right) = \theta^2/3 < \theta^2 < \infty$. Note that conditions (A) and (B) are violated in this model.

Example. $X_1, \ldots, X_n$ i.i.d Poisson(λ), $\lambda > 0$. Consider $T(\mathbf{X}) = \bar{X}$ for estimating λ. Since $E(T) = E(\bar{X}) = \lambda$, T is an unbiased estimator. We also have $Var(T) = Var(\bar{X}) = \frac{Var(X)}{n} = \frac{\lambda}{n}$. Further,

$$\begin{aligned} f(x|\lambda) &= \exp(-\lambda)\frac{\lambda^x}{x!},\ x = 0, 1, 2, \ldots \\ \log f(x|\lambda) &= -\lambda + x\log(\lambda) - \log(x!), \\ \frac{\partial}{\partial\lambda}\log f(x|\lambda) &= -1 + \frac{x}{\lambda}, \\ \frac{\partial^2}{\partial\lambda^2}\log f(x|\lambda) &= -\frac{x}{\lambda^2}, \end{aligned}$$

so that

$$\begin{aligned} I_1(\lambda) &= E_\lambda\left[\frac{\partial}{\partial\lambda}\log f(X|\lambda)\right]^2 = Var_\lambda\left[\frac{\partial}{\partial\lambda}\log f(X|\lambda)\right] \\ &= -E_\lambda\left[\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda)\right] \\ &= \frac{1}{\lambda^2}E_\lambda(X) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}. \end{aligned}$$

Therefore, $I(\lambda) = I_n(\lambda) = n/\lambda$. This yields the Information bound of

$$Var_\lambda(T) \geq \frac{1}{I(\lambda)} = \frac{\lambda}{n},$$

for any unbiased estimator T. Note that $T(\mathbf{X}) = \bar{X}$ achieves this bound, hence it is UMVUE.

Example. $X \sim$ Poisson(θ), $\theta > 0$. $q(\theta) = \exp(-\theta)$. $I(\theta) = \frac{1}{\theta}$. Consider

$$T(X) = \begin{cases} 1 & \text{if } X = 0; \\ 0 & \text{otherwise.} \end{cases}$$

Then $E(T) = P_\theta(X = 0) = \exp(-\theta) = q(\theta)$ and $Var_\theta(T) = \exp(-\theta)(1 - \exp(-\theta))$ since $T \sim$ Bernoulli($q(\theta)$). The C-R bound on all unbiased estimators U of $q(\theta)$ is

$$Var_\theta(U) \geq \frac{\left(\frac{d}{d\theta}q(\theta)\right)^2}{I(\theta)} = \frac{\exp(-2\theta)}{1/\theta} = \theta\exp(-2\theta) = \text{ C-R } (\theta).$$

Therefore,

$$\begin{aligned} \frac{Var_\theta(T)}{\text{C-R } (\theta)} &= \frac{\exp(-\theta)(1 - \exp(-\theta))}{\theta\exp(-2\theta)} = \frac{1 - \exp(-\theta)}{\theta\exp(-\theta)} = \frac{\exp(\theta) - 1}{\theta} \\ &= \frac{1 + \theta + \theta^2/2 + \cdots - 1}{\theta} > 1. \end{aligned}$$

However, X is complete sufficient, hence $T(X)$ is UMVUE of $q(\theta)$.

Confidence Statements

It is not enough to give just an estimate of the parameter of interest, however good the procedure of estimation is. Usually one also wants to know what the likely error of estimation is.

Suppose θ is the parameter of interest, and we have available, a random sample, $X_1, \ldots, X_n$ from P_θ. Suppose, further, $\hat{\theta}(X_1, \ldots, X_n)$ is an estimator of θ. It is desirable to have an estimate of the magnitude of $\hat{\theta} - \theta$. Typical estimates are asymptotically unbiased. Therefore, an estimate of s.d.$(\hat{\theta})$, called the standard error, s.e.$(\hat{\theta})$, is an indicator of the likely error of estimation of θ by $\hat{\theta}$. This means, $\hat{\theta} \pm$ s.e.$(\hat{\theta})$ is our interval estimate of θ, in the sense that $\hat{\theta}$ is the point estimate but it may be off by s.e.$(\hat{\theta})$.

Example. $\bar{X} \pm \frac{\sigma}{\sqrt{n}}$ or $\bar{X} \pm \frac{s}{\sqrt{n}}$ for μ of $N(\mu, \sigma^2)$; $\hat{p} \pm \sqrt{\hat{p}(1-\hat{p})/n}$ for p of Binomial(n, p); $\hat{\lambda} \pm \sqrt{\hat{\lambda}}$ for λ of Poisson(λ), and so on.

Another more formal approach is through confidence statements.

Interval Estimation

Let $X \sim P_\theta$ and a confidence interval is of interest for $q(\theta)$.

Definition. An interval $[\underline{T}(X), \bar{T}(X)]$, where $\underline{T} \leq \bar{T}$ is a $100(1-\alpha)\%$ confidence interval for $q(\theta)$ if

$$\inf_\theta P_\theta \left\{ \underline{T}(X) \leq \theta \leq \bar{T}(X) \right\} \geq 1 - \alpha.$$

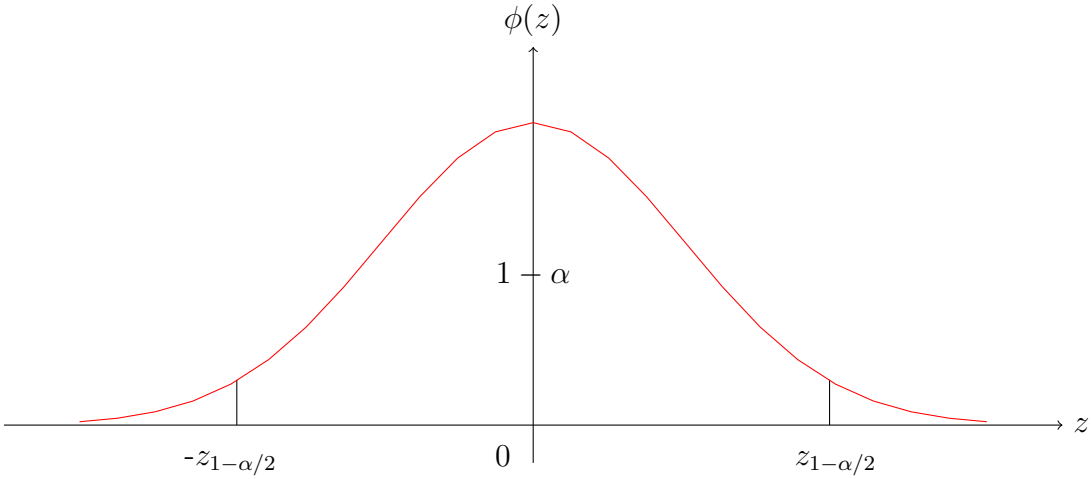
i.e., $P_\theta \left\{ \underline{T}(X) \leq \theta \leq \bar{T}(X) \right\} \geq 1 - \alpha$ for all θ.

Example. Let $X_1, \ldots, X_n$ be i.i.d $N(\theta, \sigma^2)$, σ^2 known. Want $[\underline{T}(X), \bar{T}(X)]$ such that $P_\theta \left\{ \underline{T}(X) \leq \theta \leq \bar{T}(X) \right\} \geq 1 - \alpha$ for all θ. Note,

$$Z = \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} \sim N(0,1) \text{ for all } \theta.$$

Therefore,

$$P_\theta \left(\left| \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} \right| \leq z_{1-\alpha/2} \right) = 1 - \alpha.$$



(For example, if $\alpha = 0.05$, then $z_{1-\alpha/2} = 1.96$.) Thus,

$$\begin{aligned} P_\theta\left(-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{X} - \theta \le z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1-\alpha \text{ for all } \theta, \text{ or} \\ P_\theta\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \theta \le \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1-\alpha \text{ for all } \theta. \end{aligned}$$

Therefore, $\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ is a $100(1-\alpha)\%$ confidence interval for θ.

Example. Let $X_1, \ldots, X_n$ be i.i.d $N(\mu, \sigma^2)$, σ^2 unknown. Now $\theta = (\mu, \sigma^2)$. Want $[\underline{T}(X), \bar{T}(X)]$ such that $P_\theta\left\{\underline{T}(X) \le \mu \le \bar{T}(X)\right\} \ge 1-\alpha$ for all $\theta = (\mu, \sigma^2)$. Note, since

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1} \text{ for all } \mu, \sigma^2,$$

$$P_{\mu,\sigma^2}\left(\bar{X} - t_{n-1}(1-\alpha/2)\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{n-1}(1-\alpha/2)\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

for all μ, σ^2. Thus, $\left[\bar{X} - t_{n-1}(1-\alpha/2)\frac{s}{\sqrt{n}}, \bar{X} + t_{n-1}(1-\alpha/2)\frac{s}{\sqrt{n}}\right]$ is a $100(1-\alpha)\%$ confidence interval for μ.

Interpretation of confidence statements.

Consider again the confidence interval for μ in $N(\mu, \sigma^2)$ with σ^2 known, which is $\bar{X} \pm z_{1-\alpha/2}\,\sigma/\sqrt{n}$. This means

$$\begin{aligned} & P_{\mu,\sigma^2}\{\mu \in \text{ confidence interval }\} \\ = \;& P_{\mu,\sigma^2}\left\{\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha. \end{aligned}$$

In this statement, as in all other areas of classical statistics, μ is a constant, and the probability statement is about $\bar{X}$. So $(1-\alpha)$ is the proportion of times the interval $[\underline{T}, \bar{T}]$ covers μ over repetitions of the experiment and data sets. Let $\alpha = 0.05$. Then, if the interval $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ is used for a large number of data sets in repetitions of the experiment, then in about 95% of the time μ will be inside the interval, and will lie outside rest of the time. If one has a data set with $\bar{X} = 3$, and asks for the probability that μ lies in $3 \pm 1.96\sigma/\sqrt{n}$, the answer isn't 95% but trivially zero or one depending on the value of μ. Though the idea of such intervals is quite old, it was Neyman who formalized them.

The simplest way to generate confidence intervals is to find what Fisher called a pivotal quantity, namely, a real valued function $T(\mathbf{X}, \theta)$ of both $\mathbf{X}$ and θ such that the distribution of $T(\mathbf{X}, \theta)$ does not depend on θ. Suppose then we choose two numbers t_1 and t_2 such that $P_\theta\{t_1 \le T(\mathbf{X}, \theta) \le t_2\} = 1-\alpha$. If for each $\mathbf{x}$, $T(\mathbf{x}, \theta)$ is monotone in θ, say, an increasing function of θ, then we can find $\underline{T}(\mathbf{x})$ and $\bar{T}(\mathbf{x})$ such that $T(\mathbf{x}, \bar{T}(\mathbf{x})) = t_2$ and $T(\mathbf{x}, \underline{T}(\mathbf{x})) = t_1$. Clearly $(\underline{T} \le \theta \le \bar{T})$ iff $t_1 \le T \le t_2$ and hence $\underline{T} \le \theta \le \bar{T}$ with probability $1-\alpha$.

In the normal example, $T(\mathbf{X}, \mu) = \bar{X} - \mu$, the distribution of which is $N(0, \sigma^2/n)$, free of μ.

Example. Let $X_1, \ldots, X_n$ be i.i.d $N(\mu, \sigma^2)$, both μ and σ^2 unknown. How do we construct a confidence interval for σ^2? We need a pivot involving σ^2 only. Consider the MLE of σ^2: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$, and note that $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 \sim \sigma^2 \chi^2_{n-1}$. Therefore, $S^2/\sigma^2 \sim \chi^2_{n-1}$ can be used as a pivotal statistic. i.e., we can find $c_1 < c_2$ such that

$$\begin{aligned} P_{\sigma^2}\left(c_1 \le \frac{S^2}{\sigma^2} \le c_2\right) &= 1-\alpha \text{ for all } \sigma^2, \\ P_{\sigma^2}\left(\frac{1}{c_2} \le \frac{\sigma^2}{S^2} \le \frac{1}{c_1}\right) &= 1-\alpha \text{ for all } \sigma^2, \\ P_{\sigma^2}\left(\frac{S^2}{c_2} \le \sigma^2 \le \frac{S^2}{c_1}\right) &= 1-\alpha \text{ for all } \sigma^2. \end{aligned}$$

Many choices exist for (c_1, c_2). One may take them to satisfy

$$\frac{\alpha}{2} = P\left(\chi^2_{n-1} \leq c_1\right) = 1 - P\left(\chi^2_{n-1} > c_2\right),$$

or take

$$f_{n-1}(c_1) = f_{n-1}(c_2), \text{ and } \int_{c_1}^{c_2} f_{n-1}(x)\, dx = 1 - \alpha.$$

Example. Let $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ be i.i.d. $N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$. Construct confidence interval for $\mu_1 - \mu_2$. Since $\widehat{\mu_1 - \mu_2} = \bar{X} - \bar{Y}$, look for a pivot involving this. Let $D_i = X_i - Y_i$, $i = 1, 2, \ldots, n$. Then D_i are i.i.d $N(\mu_1 - \mu_2, 2\sigma^2(1-\rho) = \sigma_D^2)$. Now, $\bar{D} \sim N(\mu_1 - \mu_2, \sigma_D^2/n)$ independent of $S_D^2 = \sum_{i=1}^n (D_i - \bar{D})^2 \sim \sigma_D^2 \chi^2_{n-1}$. Therefore,

$$T = \frac{\sqrt{n}\left(\bar{D} - (\mu_1 - \mu_2)\right)}{\sqrt{\sum_{i=1}^n (D_i - \bar{D})^2/(n-1)}} \sim t_{n-1}.$$

Hence, from the previous discussion, $\bar{D} \pm t_{n-1}(1-\alpha/2) s_D/\sqrt{n}$ is a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$. (Here $s_D^2 = S_D^2/(n-1)$.)

Example. Let $X_1, \ldots, X_n$ be i.i.d Bernoulli(θ). How do we find $\left[\underline{T}(\mathbf{X}), \bar{T}(\mathbf{X})\right]$ such that

$$P_\theta\left[\underline{T}(\mathbf{X}) \leq \theta \leq \bar{T}(\mathbf{X})\right] = 1 - \alpha \;\forall\; \theta?$$

$S_n = \sum_{i=1}^n X_i$ is sufficient for θ and $S_n \sim$ Binomial(θ). An exact pivot involving S_n is not available, so it is difficult to construct an exact confidence interval using the above approach. An approximate large sample interval is constructed as follows. If n is large then $\hat{\theta} = S_n/n \sim N(\theta, \theta(1-\theta))$ approximately. Therefore, for large n, approximately,

$$\frac{\hat{\theta} - \theta}{\sqrt{\theta(1-\theta)/n}} \sim N(0,1).$$

In fact, we have, for large n, approximately,

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{\theta}(1-\hat{\theta})/n}} \sim N(0,1).$$

This gives the usual, large sample, approximate confidence interval:

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{\theta}(1-\hat{\theta})/n}.$$

Testing Hypotheses

This is the problem of choosing between two theories, two models or two hypotheses.

(1) H_0 or the Null Hypothesis: This is the established hypothesis saying that the standard model is true or correct. If we are testing a new treatment against the current standard, H_0 denotes no difference between them.

Example. In a quality control setup, suppose that the procedure in use produces 2% defectives on average. Then new quality control measures are brought in and there is a claim of reduction in the proportion (p) of defectives. Then, we have

$$H_0 : p = 0.02 \text{ i.e., no change has occured.}$$

This effectively means that any observed changes can be explained as due to purely chance variations, and a new model is not needed.

(2) The hypothesis to be tested against H_0 is called H_1 or the Alternative Hypothesis.

In the example above,

$$H_1 : p < 0.02 \text{ i.e., a positive change has occured.}$$

This effectively means that the observed changes cannot be explained as due to chance variations under H_0 and a new model is needed.

Question. Is there enough evidence in the data to reject H_0 in favour of H_1?
We proceed as follows by considering the consequences of actions in a test procedure under the different possible states of nature.

	decision	
	accept H_0	reject H_0
H_0 is true	✓	type I error
H_1 is true	type II error	✓

Decisions (or actions, accept or reject H_0) are made using evidence from random samples or data. Therefore, we can only compute the probabilities of errors, and not know when they are committed. As shown above, *Type I Error* is the incorrect rejection of H_0, and
$P($ Type I Error $) = \alpha =$ level of significance.

Type II Error is the incorrect acceptance of H_0, and
$P($ Type II Error $) = \beta$; $1 - \beta =$ power of test.

Our approach is to fix α (at say, 0.05 or 0.01) and minimize β (or maximize the power, $1 - \beta$) to get the "most powerful" tests.

Let us consider some examples to intuitively see

- how to derive test statistics;
- how to derive test criteria.

Example. A pack of a certain brand of cigarettes displays the statement, "1.5 mg nicotine on average per cigarette". Let μ denote the actual average nicotine content per cigarette for all cigarettes of this brand. It is required to test if the actual average is higher than what is claimed. Suppose a sample of cigarettes is selected, and the nicotine content of each cigarette is determined. The observed contents are $X_1, \ldots, X_n$. Conduct the test at the level of significance of 5%. We assume that X_i are distributed as i.i.d $N(\mu, \sigma^2)$. Then, we desire to test

$$H_0 : \mu = 1.5 = \mu_0 \text{ versus } H_1 : \mu > 1.5 \text{ i.e., average is higher.}$$

How to we calculate evidence? Estimate μ from data. $\hat{\mu} = \bar{X}$. Compare $\hat{\mu}$ with μ_0. When do we say that H_0 is not true?

$\bar{X} - \mu_0 =$ observed difference between estimated mean and the hypothesized value.

However, $\bar{X}$ has variation from sample to sample. The expected variation in $\bar{X}$ values is s.e.$(\bar{X}) = \frac{s}{\sqrt{n}}$.
Now compare the observed difference with the expected, and enquire: Is

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} >> 1?$$

i.e., is the observed difference much larger than the expected difference?

If so, reject H_0. If

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim 1 \text{ or } < 1$$

there is no evidence to reject H_0. Thus, we have the statistic for testing:

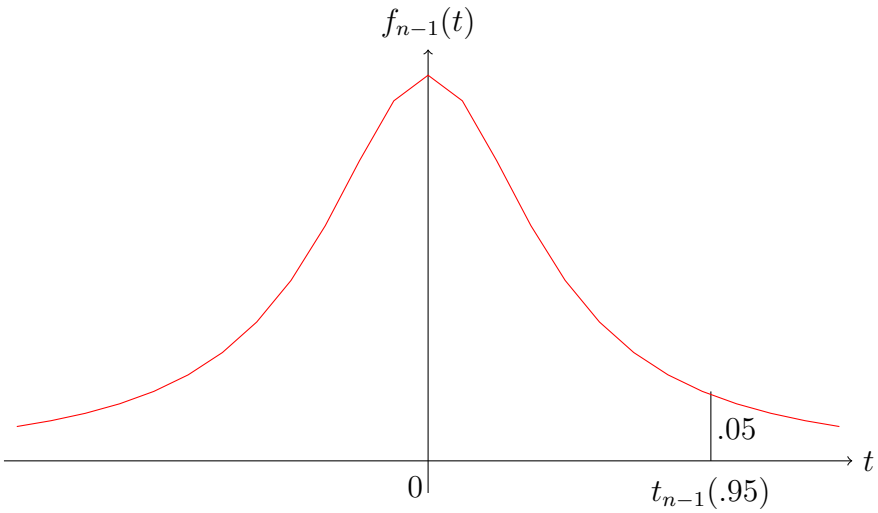
$$\text{Test statistic } = \frac{\text{observed departure from } H_0}{\text{expected departure}}.$$

In the present problem, the test statistic is

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} \text{ if } H_0 \text{ is true.}$$

Reject H_0 if the observed value of the test statistic is large. But, how large?

$$\begin{aligned} 0.05 &= \alpha = P(\text{ Type I Error }) = P(\text{ Reject } H_0 \text{ when } H_0 \text{ is true }) \\ &= P\left(\frac{\bar{X} - \mu_0}{s/\sqrt{n}} > C\right) = P\left(t_{n-1} > C\right). \end{aligned}$$



Therefore, $C = t_{n-1}(1-\alpha) = t_{n-1}(.95)$ and,
we claim to have evidence against the null hypothesis at the 5% level of significance
if the observed value of $\frac{\bar{X}-\mu_0}{s/\sqrt{n}} > t_{n-1}(.95)$.

Example. $X \sim \text{Binomial}(n, p)$ and it is of interest to test

$$H_0 : p = p_0 \text{ versus } H_1 : p \neq p_0 \text{ a two-sided alternative.}$$

$\hat{p} = \frac{X}{n}$ and it would be natural to use the statistic:

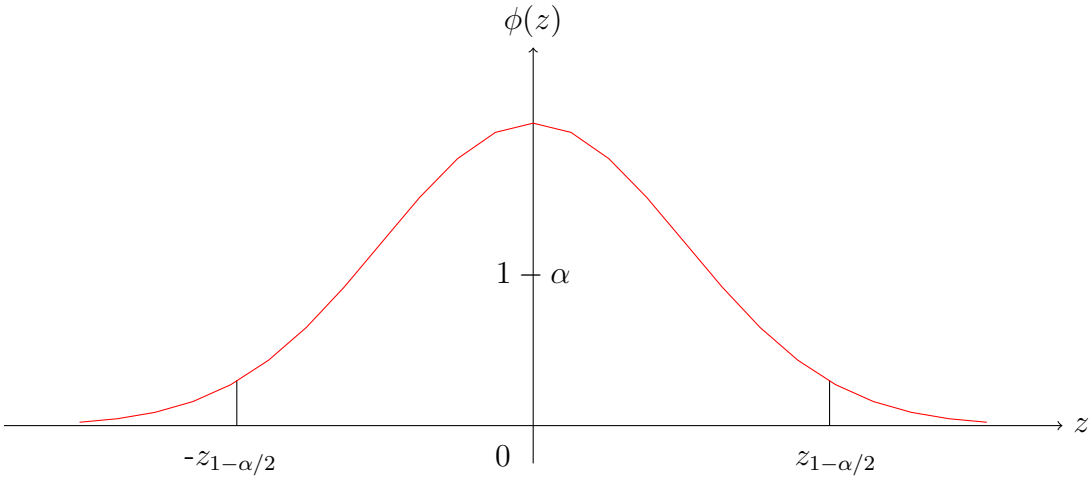
$$\frac{\text{estimated departure from } H_0}{\text{expected departure or s.e.}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1),$$

approximately for large n, if H_0 is true. Since the alternative hypothesis is two-sided, the test statistic is $\left|\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}\right|$, large values of which provide

evidence against H_0 and in favour of H_1. To obtain the crtitical value (above which H_0 is to be rejected), note

$$\alpha = P_{H_0}\left(\left|\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}\right| > C\right) = P(|Z| > C),$$

implying that $C = z_{1-\alpha/2}$.



At the significance level of α, evidence to reject H_0 exists
if $|\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}| > z_{1-\alpha/2}$.

How does one derive 'most powerful' tests?

Power of a test. As noted previously, power of a test is P_{H_0}(reject H_0). Suppose $X_1, \ldots, X_n$ are i.i.d $N(\mu, \sigma^2)$ where σ^2 is known and we desire to test

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu > \mu_0$$

at the level of significance α. We use the test statistic

$$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \text{ if } H_0 \text{ is true.}$$

Reject H_0 if the observed value of $Z > z_{1-\alpha}$ since

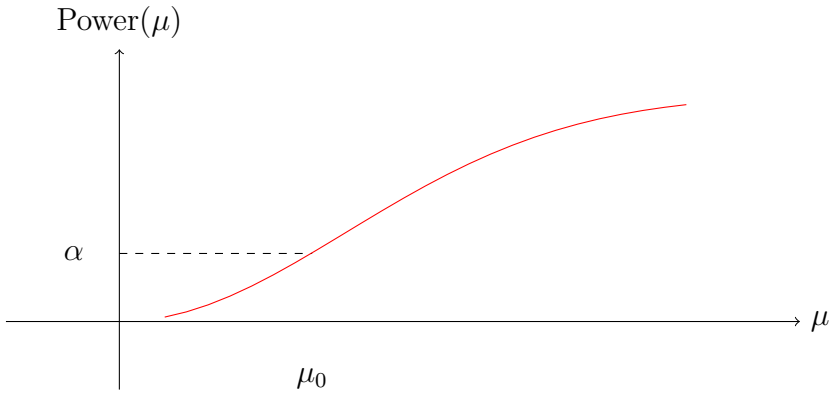
$$\alpha \;=\; P_{\mu=\mu_0}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}\right).$$

To compute the power of this test at any $\mu > \mu_0$ (under H_1), we have,

$$\begin{aligned}\text{Power}(\mu) &= P_\mu\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}\right) = P_\mu\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} + \frac{\mu-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}\right) \\ &= P_\mu\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > z_{1-\alpha} - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right) = P\left(Z > z_{1-\alpha} - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right),\end{aligned}$$

which is an increasing function of μ. i.e., if μ is much larger than μ_0, the test rejects H_0 easily.

Power(μ) as a function of μ is called the power curve.



P-values. The error probabilities of a test (significance level α and the power $1-\beta$ which are predetermined) do not provide a measure of the strength of evidence in a particular data set against H_0. The P-values defined below try to capture that.
Suppose $H_0 : \theta = \theta_0$ and your test is to reject H_0 for large values of a test statistic $W = W(\mathbf{X})$. Then, when $\mathbf{X} = \mathbf{x}$ is observed, the P-value is defined as

$$p(\mathbf{x}) = P_{\theta_0}\left(W > W(\mathbf{x})\right),$$

Any data point $\mathbf{x}$ for which $p(\mathbf{x}) \leq \alpha$ may be considered strong enough eveidence to reject H_0 at the significance level of α.

Example. Let $X_1, X_2, \ldots, X_9$ be i.i.d $N(\mu, 1)$. It is of interest to test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. Compute P-value for the Z-test if $\bar{x} = 0.9$ is observed. The Z-test rejects H_0 when

$$|Z| = \left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| = 3|\bar{X}|$$

is large. $Z \sim N(0, 1$ when H_0 is true. Therefore the P-value when $\mathbf{x}$ is observed is $p(\mathbf{x}) = P(|Z| > 3\bar{x})$. If $\bar{x} = 0.9$, then the P-value is $2[1 - \Phi(2.7)]$ $= 0.007$.

How do we know that these are the 'best' tests avaiable to us? Let us discuss the Neyman-Pearson theory of deriving optimal tests for this.

Neyman-Pearson Theory of Testing

$X \sim P_\theta$, $\theta \in \Theta$. $\mathcal{X}$ = sample space = set of all values that X can take. It is of interest to test

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_1,$$

where $\Theta_i \subset \Theta$ and $\Theta_0 \cap \Theta_1 = \phi$.
Simple hypothesis: $\Theta_0 = \{\theta_0\}$, i.e., $H_0 : \theta = \theta_0$.
Composite hypothesis: $\Theta_0 = (-\infty, \theta_0]$, i.e., $H_0 : \theta \leq \theta_0$.

Nonrandomized test. Find a subset S of $\mathcal{X}$, and reject H_0 if the observed value $x \in S$. $S \subset \mathcal{X}$ is called the critical region or the rejection region of the test. One defines a test function for nonrandomized tests ϕ as

$$\phi(x) = \begin{cases} 1 & \text{if } x \in S; \\ 0 & \text{if } x \in S^c. \end{cases}$$

Note that $\phi(x)$ is also the probability of rejecting H_0 upon observing x.

For a level α test, one must have,

$$\sup_{\theta \in \Theta_0} P_\theta(X \in S) \leq \alpha.$$

Note that, if $\Theta_0 = \{\theta_0\}$, then we need $P_{\theta_0}(X \in S) \leq \alpha$.
For nonrandomized tests, it may happen that $\sup_{\theta \in \Theta_0} P_\theta(X \in S) < \alpha$.

Power(θ) = Power of test = $P_\theta(X \in S) = E_\theta[\phi(X)]$ for $\theta \in \Theta_1$ is the power function of the test associated with S.

Randomized test. Any ϕ such that $0 \leq \phi(x) \leq 1$ for all $x \in \mathcal{X}$, and at any x, $\phi(x)$ is the probability of rejecting H_0 if x is observed. Nonrandomized tests form a subset of randomized tests. The power function for randomized tests is given by

$$P_\theta(\text{Reject}H_0) = E_\theta\left[P(\text{Reject}H_0|X)\right] = E_\theta\phi(X) = \int_{\mathcal{X}} \phi(x)\, dP_\theta(x).$$

Problem. Find ϕ such that $E_\theta\phi$ is maximized when $\theta \in \Theta_1$ and subject to $\sup_{\theta \in \Theta_0} E_\theta\phi(X) \leq \alpha$. Such a test, if it exists, is called a Uniformly Most Powerful (UMP) test.

Consider the two kinds of the errors defined earlier. It turns out that in general if one tries to reduce one error probability the other error probability goes up, so one cannot reduce both simultaneously. Because probability of

error of first kind is more important, one first makes it small (by fixing it at a small value such as 0.01 or 0.05). Among all tests satisfying this, one then tries to minimize the probability of committing error of second kind or equivalently, to maximize the power uniformly for all θ in H_1.

N-P Lemma. Suppose $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. i.e., simple null versus simple alternative. Also, let p_{θ_0} and p_{θ_1} denote the respective densities (pdf or pmf). Then
(a) there exists a test ϕ and a constant $k \geq 0$ such that

$$E_{\theta_0}\phi(X) = \alpha \tag{1}$$

and

$$\phi(x) = \begin{cases} 1 & \text{when } p_{\theta_1}(x) > kp_{\theta_0}(x); \\ 0 & \text{when } p_{\theta_1}(x) < kp_{\theta_0}(x). \end{cases} \tag{2}$$

(b) If a test satisfies (1) and (2) for some k, then it is most powerful for testing $H_0 : P_\theta = P_{\theta_0}$ versus $H_1 : P_\theta = P_{\theta_1}$ at level α.
(c) If ϕ is a most powerful test at level α for testing $H_0 : P_\theta = P_{\theta_0}$ versus $H_1 : P_\theta = P_{\theta_1}$, then it satisfies (1) and (2) for some k. (This answers whether MP test can have some other form.)

Remark. The most powerful test is of the form: reject H_0 if $p_{\theta_1}(x)/p_{\theta_0}(x) > k$. i.e., when the likelihood ratio exceeds a threshold. This is intuitively meaningful because, on the one hand we want $\int_S p_{\theta_0}(x)\,dx \leq \alpha$ and on the other $\int_S p_{\theta_1}(x)\,dx =$ maximum. To achieve this one must put all x that give very large values of $p_{\theta_1}(x)/p_{\theta_0}(x)$ in S.

Proof. Let $0 < \alpha < 1$ and define

$$\alpha(c) = P_{\theta_0}\left(p_{\theta_1}(X) > cp_{\theta_0}(X)\right) = P_{\theta_0}\left(\frac{p_{\theta_1}(X)}{p_{\theta_0}(X)} > c\right).$$

(The second equality is because $P_{\theta_0}(p_{\theta_0}(X) = 0) = 0$.) Then,

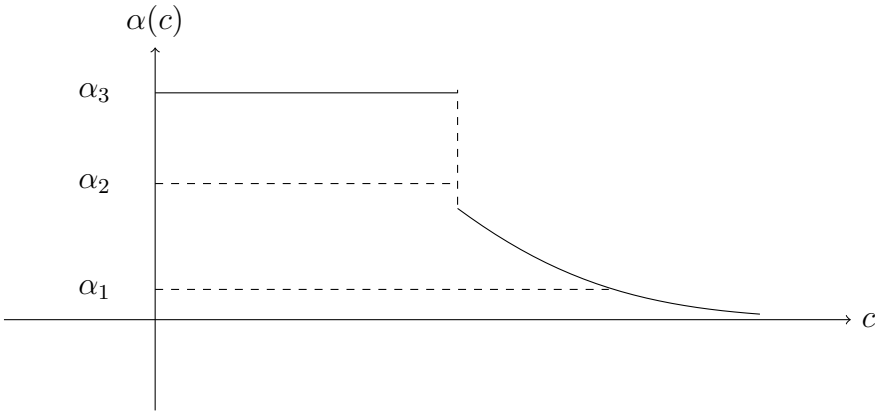
$$1 - \alpha(c) = P_{\theta_0}(r(X) \leq c), \text{ where } r(X) = \tfrac{p_{\theta_1}(X)}{p_{\theta_0}(X)}.$$

Therefore, $1 - \alpha(c)$ is the cdf of $r(X)$. Thus, $1 - \alpha(c)$ is nondecreasing and right continuous. Therefore, $\alpha(c)$ is nonincreasing and right continuous. i.e., $1 - \alpha(-\infty) = 0$, $1 - \alpha(\infty) = 1$, $1 - \alpha(c) \nearrow$ and $1 - \alpha(c+) = 1 - \alpha(c)$. Therefore, $\alpha(-\infty) = 1$, $\alpha(\infty) = 0$, $\alpha(c) \searrow$ and $\alpha(c+) = \alpha(c)$. We would like to find c_0, if possible, such that $\alpha(c_0) = \alpha$. Note that

$$\alpha(c-) - \alpha(c) = P_{\theta_0}(r(X) = c) = P_{\theta_0}\left(\frac{p_{\theta_1}(X)}{p_{\theta_0}(X)} = c\right).$$

For fixed $0 < \alpha < 1$, find c_0 such that $\alpha(c_0) \leq \alpha \leq \alpha(c_0-)$ and define

$$\phi(x) = \begin{cases} 1 & \text{if } p_{\theta_1}(x) > c_0 p_{\theta_0}(x); \\ 0 & \text{if } p_{\theta_1}(x) < c_0 p_{\theta_0}(x); \\ \frac{\alpha - \alpha(c_0)}{\alpha(c_0-) - \alpha(c_0)} & \text{if } p_{\theta_1}(x) = c_0 p_{\theta_0}(x). \end{cases}$$



If $\alpha(c_0) = \alpha(c_0-)$, (i.e., α_1 or α_3 in the figure) then

$$P_{\theta_0}\left(\frac{p_{\theta_1}(X)}{p_{\theta_0}(X)} = c_0\right) = 0,$$

i.e., $P_{\theta_0}(p_{\theta_1}(X) = c_0 p_{\theta_0}(X)) = 0$. Therefore, $\phi(.)$ is defined a.e. w.r.t. P_{θ_0}.

Note that

$$\begin{aligned} E_{\theta_0}\phi(X) &= P_{\theta_0}\left(\frac{p_{\theta_1}(X)}{p_{\theta_0}(X)} > c_0\right) + \frac{\alpha - \alpha(c_0)}{\alpha(c_0-) - \alpha(c_0)} P_{\theta_0}\left(\frac{p_{\theta_1}(X)}{p_{\theta_0}(X)} = c_0\right) \\ &= \alpha(c_0) + \frac{\alpha - \alpha(c_0)}{\alpha(c_0-) - \alpha(c_0)} \left(\alpha(c_0-) - \alpha(c_0)\right) = \alpha \end{aligned}$$

Choose $k = c_0$ to satify (1) and (2).

(b) Suppose ϕ is a test which satisfies (1) and (2). We want to show ϕ is MP at level α. Let ϕ^* be any other test such that $E_{\theta_0}\phi^*(X) \leq \alpha$. (Note that $0 \leq \phi(x) \leq 1$ and $0 \leq \phi^*(x) \leq 1$.) Let

$$S^+ = \{x : \phi(x) - \phi^*(x) > 0\} \text{ and } S^- = \{x : \phi(x) - \phi^*(x) < 0\}.$$

Then, if $x \in S^+$, $\phi(x) > \phi^*(x) \geq 0$. Therefore,

$$\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \geq k, \text{ or } p_{\theta_1}(x) \geq k p_{\theta_0}(x).$$

If $x \in S^-$, $\phi(x) < \phi^*(x) \leq 1$, so that $p_{\theta_1}(x) \leq k p_{\theta_0}(x)$. Therefore,

$$\begin{aligned}
&\int_{\mathcal{X}} (\phi(x) - \phi^*(x))(p_{\theta_1}(x) - k p_{\theta_0}(x))\, dx \\
&\quad = \int_{S^+ \cup S^-} (\phi(x) - \phi^*(x))(p_{\theta_1}(x) - k p_{\theta_0}(x))\, dx \geq 0.
\end{aligned}$$

Hence,

$$\begin{aligned}
\int_{\mathcal{X}} (\phi(x) - \phi^*(x)) p_{\theta_1}(x)\, dx &\geq k \int_{\mathcal{X}} (\phi(x) - \phi^*(x)) p_{\theta_0}(x)\, dx \\
&= k \left[\int_{\mathcal{X}} \phi(x) p_{\theta_0}(x)\, dx - \int_{\mathcal{X}} \phi^*(x) p_{\theta_0}(x)\, dx \right] \\
&= k \left[\alpha - E_{\theta_0} \phi^*(X) \right] \geq 0.
\end{aligned}$$

i.e., $E_{\theta_1} \phi(X) \geq E_{\theta_1} \phi^*(X)$.

(c) Suppose ϕ^* is the most powerful test and ϕ satisfies (1) and (2). Let

$$S = \left(S^+ \cup S^- \right) \cap \left\{ x : p_{\theta_1}(x) \neq k p_{\theta_0}(x) \right\}.$$

Suppose S has positive probability (or positive Lebesgue measure in the continuous case). Then, since $(\phi(x) - \phi^*(x))(p_{\theta_1}(x) - k p_{\theta_0}(x)) > 0$ on S,

$$\begin{aligned}
&\int_{S^+ \cup S^-} (\phi(x) - \phi^*(x))(p_{\theta_1}(x) - k p_{\theta_0}(x))\, dx \\
&\quad = \int_{S} (\phi(x) - \phi^*(x))(p_{\theta_1}(x) - k p_{\theta_0}(x))\, dx > 0.
\end{aligned}$$

Therefore,

$$\int_{\mathcal{X}} (\phi(x) - \phi^*(x)) p_{\theta_1}(x)\, dx > k \int_{\mathcal{X}} (\phi(x) - \phi^*(x)) p_{\theta_0}(x)\, dx \geq 0,$$

and hence,

$$\int_{\mathcal{X}} \phi(x) p_{\theta_1}(x)\, dx > \int_{\mathcal{X}} \phi^*(x) p_{\theta_1}(x)\, dx.$$

i.e., $E_{\theta_1} \phi(X) > E_{\theta_1} \phi^*(X)$, which contradicts the assumption that ϕ^* is MP. Therefore, S must have probability 0, which implies that ϕ and ϕ^* can differ only on the set $\{x : p_{\theta_1}(x) = k p_{\theta_0}(x)\}$. (So, the MP tests may differ by picking different subsets of this set to satisfy the level condition.)

Corollary. The power of the MP level α test for testing $H_0 : P_\theta = P_{\theta_0}$ versus $H_1 : P_\theta = P_{\theta_1}$ is strictly larger than α unless $P_{\theta_1} = P_{\theta_0}$.

Proof. Consider $\phi(x) \equiv \alpha$. This is a level α test since $E_{\theta_0}\phi(X) = \alpha$. Since $E_{\theta_1}\phi(X) = \alpha$ also, the MP test has power at least $E_{\theta_1}\phi(X) = \alpha$. If the power of the MP test is exactly α, then $\phi(x) \equiv \alpha$ is also MP. Then from the earlier theorem, part (c), ϕ should satisfy:

$$\phi(x) = \begin{cases} 1 & \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k; \\ 0 & \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < k. \end{cases}$$

Therefore $p_{\theta_1}(x) = p_{\theta_0}(x)$ w.p. 1. (Why is $k = 1$?) i.e., $P_{\theta_1} = P_{\theta_0}$.

Example. $X \sim$ Binomial$(2, p)$. Test $H_0 : p = 1/2$ versus $H_1 : p = 3/4$ at level $\alpha = 0.01$. MP test ϕ has the form:

$$\phi(x) = \begin{cases} 1 & \frac{p_{3/4}(x)}{p_{1/2}(x)} > k; \\ 0 & \frac{p_{3/4}(x)}{p_{1/2}(x)} < k. \end{cases}$$

Note that

$$\frac{p_{3/4}(x)}{p_{1/2}(x)} = \frac{\binom{2}{x}(3/4)^x(1/4)^{2-x}}{\binom{2}{x}(1/2)^x(1/2)^{2-x}} > k \text{ iff}$$

$$3^x 2^2 4^{-2} > k \text{ iff}$$

$$3^x > 4k \text{ iff}$$

$$x > \frac{\log(4k)}{\log(3)} = k_1.$$

i.e.,

$$\phi(x) = \begin{cases} 1 & \text{if } x > k_1; \\ 0 & \text{if } x < k_1; \\ \gamma & \text{if } x = k_1. \end{cases},$$

where k_1 and γ are chosen to satisfy the level condition. Now note that

$$p_{1/2}(x) = \begin{cases} 1/4 & \text{if } x = 0; \\ 1/2 & \text{if } x = 1; \\ 1/4 & \text{if } x = 2. \end{cases}$$

Since $p_{1/2}(2) = 1/4 > 0.01$, $k_1 = 2$. Since

$$\alpha = 0.01 = E_{1/2}\phi(X) = \gamma p_{1/2}(2) = \gamma/4,$$

the MP test rejects H_0 with probability 0.04 if $x = 2$; accepts otherwise.

Example. $X_1, \ldots, X_n$ i.i.d $N(\mu, \sigma^2)$, σ^2 known. Test $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ where $\mu_1 > \mu_0$. $\bar{X}$ is sufficient for μ and $\bar{X} \sim N(\mu, \sigma^2/n)$. We switch notation now from $p_\theta(x)$ to $f(x|\theta)$. MP level α test is of the form:

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } \frac{f(\bar{x}|\mu_1)}{f(\bar{x}|\mu_0)} > k; \\ \gamma & \text{if } \frac{f(\bar{x}|\mu_1)}{f(\bar{x}|\mu_0)} = k; \\ 0 & \text{if } \frac{f(\bar{x}|\mu_1)}{f(\bar{x}|\mu_0)} < k. \end{cases}$$

Since $f(\bar{x}|\mu) = (2\pi)^{-1/2} \frac{\sqrt{n}}{\sigma} \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\right)$,

$$\begin{aligned} \frac{f(\bar{x}|\mu_1)}{f(\bar{x}|\mu_0)} &= \exp\left(-\frac{n}{2\sigma^2}\left\{\bar{x}^2 + \mu_1^2 - 2\mu_1\bar{x} - \bar{x}^2 - \mu_0^2 + 2\mu_0\bar{x}\right\}\right) \\ &= \exp\left(-\frac{n}{2\sigma^2}\left\{\mu_1^2 - \mu_0^2 - 2\bar{x}(\mu_1 - \mu_0)\right\}\right) \\ &= \exp\left(\frac{n(\mu_1 - \mu_0)}{\sigma^2}\bar{x} - \frac{n}{2\sigma^2}(\mu_1^2 - \mu_0^2)\right), \end{aligned}$$

which is a monotone increasing function of $\bar{x}$. Therefore,

$$\frac{f(\bar{x}|\mu_1)}{f(\bar{x}|\mu_0)} > k \text{ iff } \bar{x} > c \text{ for some } c,$$

and hence

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } \bar{x} > c; \\ \gamma & \text{if } \bar{x} = c; \\ 0 & \text{if } \bar{x} < c. \end{cases}$$

Choice of γ is not needed since $P_\mu(\bar{X} = c) = 0$. We need

$$\alpha = P_{\mu_0}\left(\bar{X} > c\right) = P_{\mu_0}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{c - \mu_0}{\sigma/\sqrt{n}}\right),$$

so that $\frac{c-\mu_0}{\sigma/\sqrt{n}} = z_{1-\alpha}$. Equivalently,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}\ ; \\ 0 & \text{if } \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha}. \end{cases}$$

This is simply the standard Z-test.

Uniformly Most Powerful (UMP) Tests.

For testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, ϕ is level α UMP test if $\sup_{\theta \in \Theta_0} E_\theta \phi(X) \leq \alpha$ and

$E_\theta\phi(X) \geq E_\theta\phi^*(X)$ for all $\theta \in \Theta_1$, for any other test ϕ^* for which $\sup_{\theta\in\Theta_0} E_\theta\phi^*(X) \leq \alpha$. This extension of the MP theory is subject to rather strong conditions on the density $f(x|\theta)$:

I. $\Theta \subset \mathcal{R}^1$ (single parameter)

II. Monotone Likelihood Ratio (MLR). $P_\theta, \theta \in \Theta \subset \mathcal{R}^1$ with density $f(x|\theta)$ is said to have m.l.r if there exists a real valued function $T(x)$ such that for any $\theta < \theta'$, $P_\theta \neq P_{\theta'}$ and the likelihood ratio $\frac{f(x|\theta')}{f(x|\theta)}$ is a nondecreasing function of $T(x)$. i.e.,

$$\frac{f(x|\theta')}{f(x|\theta)} = h_{\theta,\theta'}(T(x)), \text{ where } h_{\theta,\theta'}(y) \nearrow y \text{ for fixed } \theta' > \theta.$$

Example. $X \sim$ Binomial(n, θ). Then

$$\begin{aligned} f(x|\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x}, \text{ and hence} \\ \frac{f(x|\theta')}{f(x|\theta)} &= \left(\frac{\theta'}{\theta}\right)^x\left(\frac{1-\theta'}{1-\theta}\right)^{n-x} \\ &= \left(\frac{1-\theta'}{1-\theta}\right)^n\left(\frac{\theta'/(1-\theta')}{\theta/(1-\theta)}\right)^x. \end{aligned}$$

For fixed $\theta' > \theta$, $\left(\frac{1-\theta'}{1-\theta}\right)^n$ is fixed and $\frac{\theta'}{1-\theta'} > \frac{\theta}{1-\theta}$, so that $\left(\frac{\theta'/(1-\theta')}{\theta/(1-\theta)}\right)^x$ is increasing in $T(x) = x$.

Theorem. If $P_\theta, \theta \in \Theta \subset \mathcal{R}^1$ belongs to a one-parameter exponential family having density

$$f(x|\theta) = \exp\left(c(\theta)T(x) + d(\theta) + S(x)\right) I_A(x)$$

with $c(.)$ strictly monotone in θ (strictly increasing or strictly decreasing) then $\{P_\theta\}$ has m.l.r. in $T(x)$ or $-T(x)$.

Proof. Note that

$$\frac{f(x|\theta')}{f(x|\theta)} = \exp\left(T(x)\left[c(\theta') - c(\theta)\right]\right)\exp\left(d(\theta') - d(\theta)\right)$$

is increasing in $T(x)$ if $c(.)$ is increasing, otherwise increasing in $-T(x)$.

Example. $X \sim U(0, \theta]$. Then $f(x|\theta) = \frac{1}{\theta}I(0 < x \leq \theta)$, so that for $\theta' > \theta > 0$,

$$\frac{f(x|\theta')}{f(x|\theta)} = \begin{cases} \frac{\theta}{\theta'} & \text{if } 0 < x \leq \theta; \\ \infty & \text{if } \theta < x \leq \theta'. \end{cases}$$

This shows that $U(0, \theta)$, $\theta > 0$ has monotone likelihood ratio (in $T(x) = x$) even though it is not an exponential family.

Theorem. Let θ be a real parameter and let X have density $f(x|\theta)$ with MLR in $T(x)$. Then
(i) for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, there exists a UMP level α test given by

$$\phi(x) = \begin{cases} 1 & \text{when } T(x) > C; \\ \gamma & \text{when } T(x) = C; \\ 0 & \text{when } T(x) < C, \end{cases} \qquad (*)$$

where C and γ are determined by

$$E_{\theta_0}\phi(X) = \alpha. \qquad (**)$$

(ii) The power function $E_\theta\phi(X)$ of this test is strictly increasing for all points θ for which $E_\theta\phi(X) < 1$.
(iii) For all θ', the test given by (*) and (**) is UMP for testing $H_0 : \theta \leq \theta'$ versus $H_1 : \theta > \theta'$ at the level $\alpha' = E_{\theta'}\phi(X)$.
(iv) For any $\theta < \theta_0$, the test (*) and (**) minimizes $E_\theta\phi(X)$ among all tests satisfying (**).

Proof. We first prove the existence of ϕ satisfying (*) and (**). Then we show that (ii) holds for this ϕ. Then we prove the UMP part of (i).

Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, $\theta_1 > \theta_0$. From N-P Lemma, MP test for this is of the form:

$$\phi(x) = \begin{cases} 1 & \text{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k; \\ 0 & \text{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < k. \end{cases}$$

But $p_\theta(x) = f(x|\theta)$ has MLR in $T(x)$. Therefore,

$$\frac{f(x|\theta_1)}{f(x|\theta_0)} > k \text{ iff } T(x) > C \text{ for some } C \ .$$

Hence,

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > C; \\ 0 & \text{if } T(x) < C. \end{cases}$$

Now, from N-P Lemma (a), there exist C and γ satisfying (*) and (**). These do not depend on θ_1. (Note, due to MLR, dependence of MP on θ_1 has been eliminated.) From N-P Lemma (b) this test is MP for testing $H_0 : \theta = \theta'$ versus $H_1 : \theta = \theta''$ at level $\alpha' = E_{\theta'}\phi(X)$ if $\theta' < \theta''$. This is because, from MLR,

$$\frac{f(x|\theta'')}{f(x|\theta')} > k_1 \text{ iff } T(x) > C_1 \text{ for some } C_1$$

and the MP test is exactly of the same form as ϕ. Now from the corollary to N-P Lemma, we have $E_{\theta''}\phi(X) > E_{\theta'}\phi(X)$, whenever $\theta'' > \theta'$. This proves (ii). The fact that $E_\theta\phi(X)$ is strictly increasing implies that

$$E_\theta\phi(X) \leq \alpha \; \forall \; \theta \leq \theta_0. \qquad (***)$$

Now let us check what test is UMP for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. For this, consider the class $\mathcal{C}_1$ of all tests that satisfy (***). (We want to know what is best in $\mathcal{C}_1$.) This class $\mathcal{C}_1$ is contained in the class $\mathcal{C}_2$ of tests that satisfy $E_{\theta_0}\phi(X) \leq \alpha$. In other words, we have,

$$\mathcal{C}_1 = \{\phi : E_\theta\phi(X) \leq \alpha \; \forall \; \theta \leq \theta_0\} \subset \mathcal{C}_2 = \{\phi : E_{\theta_0}\phi(X) \leq \alpha\},$$

since $E_\theta\phi(X) \leq \alpha \; \forall \; \theta \leq \theta_0$ implies $E_{\theta_0}\phi(X) \leq \alpha$. From N-P Lemma (b), MP level α test for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ (for any $\theta_1 > \theta_0$) is given by (*) and (**). That means that ϕ satisfying (*) and (**) is best (MP) in $\mathcal{C}_2$ and this ϕ belongs to $\mathcal{C}_1$ also. Therefore it is best in $\mathcal{C}_1$. However, this test does not depend on θ_1. Therefore it is UMP for $H_1 : \theta > \theta_0$. i.e., for any $\theta > \theta_0$, $E_\theta\phi(X) \geq E_\theta\phi^*(X)$ for all $\phi^* \in \mathcal{C}_1$.

Example. $X \sim$ Binomial(n, p) and it is of interest to test $H_0 : p \leq p_0$ versus $H_1 : p > p_0$. (Consider a clinical trial where the efficacy of a drug is being checked.) MLR exists in $T(x) = x$. Therefore, UMP exists and is given by

$$\phi(x) = \begin{cases} 1 & \text{if } x > x_0; \\ 0 & \text{if } x < x_0; \\ \gamma & \text{if } x = x_0. \end{cases}$$

Choose γ and x_0 to satisfy

$$E_{p_0}\phi(X) = P_{p_0}(X > x_0) + \gamma P_{p_0}(X = x_0) = \alpha.$$

Let $\alpha = 0.05$, $n = 10$ and $p_0 = 1/2$. We have

x	$f(x\|p = 1/2)$	$P_{p=1/2}(X \geq x)$
10	0.000977	0.00098
9	0.009766	0.01074
8	0.043945	0.05469

Thus, $x_0 = 8$ and

$$\gamma = \frac{\alpha - P(X > x_0|p = p_0)}{P(X = x_0|p = p_0)} = \frac{0.05 - 0.01074}{0.04395} = 0.8933.$$

Example. $X_1, \ldots, X_n$ i.i.d $N(\mu, \sigma^2)$, σ^2 known. Test $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. $\bar{X}$ is sufficient for μ and $\bar{X} \sim N(\mu, \sigma^2/n)$. It was shown previously,

directly, that the likelihood ratio is an increasing function of $T(\bar{x}) = \bar{x}$. This means MLR which can also be shown using the fact that $N(\mu, \sigma^2)$, σ^2 known is a one-parameter exponential family. Since

$$f(\bar{x}|\mu) = \sqrt{\frac{n}{2\pi}}\frac{1}{\sigma}\exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right) = \exp\left(\frac{n\mu}{\sigma^2}\bar{x} - \frac{n}{2\sigma^2}\mu^2 - \frac{n}{2\sigma^2}\bar{x}^2 + \cdots\right),$$

which establishes MLR in $T(\bar{x}) = \bar{x}$. Therefore UMP level α test is

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } \bar{x} > C; \\ \gamma & \text{if } \bar{x} = C; \\ 0 & \text{if } \bar{x} < C. \end{cases}$$

Choice of γ is not needed since $P_\mu(\bar{X} = C) = 0$. We need

$$\alpha = P_{\mu_0}\left(\bar{X} > C\right) = P_{\mu_0}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{C-\mu_0}{\sigma/\sqrt{n}}\right),$$

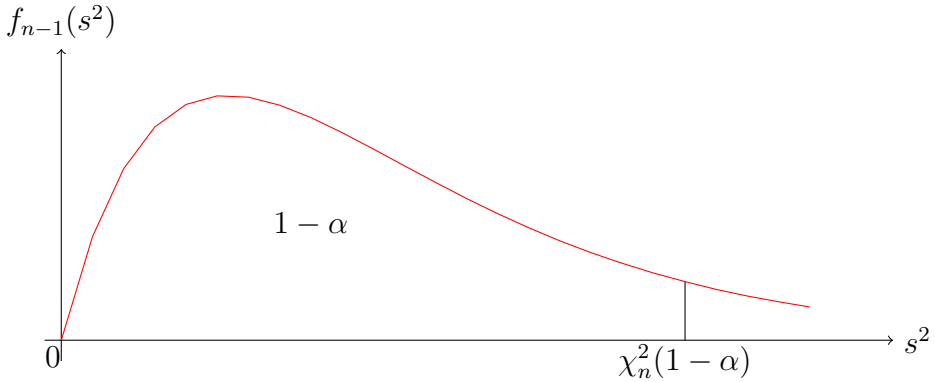
so that $\frac{C-\mu_0}{\sigma/\sqrt{n}} = z_{1-\alpha}$ or $C = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$. Thus, the UMP test simply rejects H_0 if $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}$. This is simply the standard Z-test.

Example. $X_1, \ldots, X_n$ i.i.d $N(\mu_0, \sigma^2)$, μ_0 known. Test $H_0 : \sigma^2 \leq \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$. $S^2 = \sum_{i=1}^n (X_i - \mu_0)^2$ is sufficient for σ^2 and $S^2 = \sum_{i=1}^n (X_i - \mu_0)^2/\sigma^2 \sim \chi^2_n$.

$$\begin{aligned} f_{S^2}(s^2|\sigma^2) &= \exp\left(-\frac{s^2}{2\sigma^2}\right)\left(\frac{s^2}{2\sigma^2}\right)^{n/2-1} \times \text{ constant} \\ &= \exp\left(-\frac{1}{2\sigma^2}s^2 - \frac{n}{2}\log(\sigma^2) + (\frac{n}{2}-1)\log(s^2) + \cdots\right). \end{aligned}$$

Thus we have a one-parameter exponential family with MLR in $T(s^2) = s^2$. Therefore the UMP level α test is

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } s^2 > C; \\ 0 & \text{if } s^2 < C. \end{cases}$$



C is determined by

$$\alpha = P_{\sigma_0^2}\left(S^2 > C\right) = P_{\sigma_0^2}\left(\frac{S^2}{\sigma_0^2} > \frac{C}{\sigma_0^2}\right).$$

Since $S^2/\sigma_0^2 \sim \chi_n^2$, when $\sigma^2 = \sigma_0^2$, we have that $C/\sigma_0^2 = \chi_n^2(1-\alpha)$

Are there UMP tests for all fairly simple problems? Not in most cases.

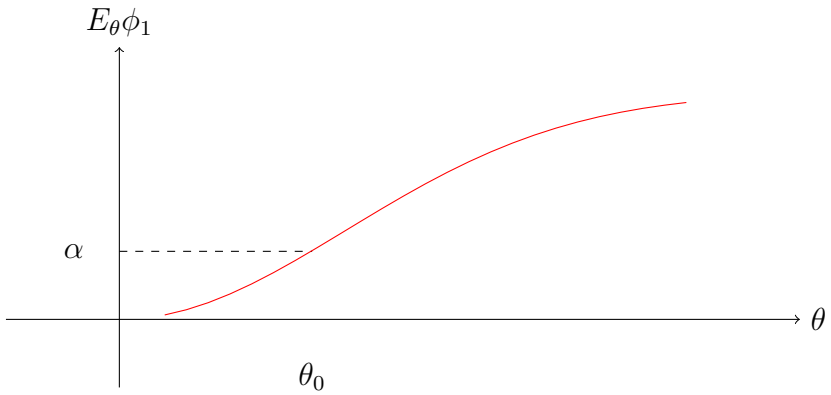
Example. $X_1, \ldots, X_n$ i.i.d $N(\mu, \sigma^2)$, σ^2 unknown. Test $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$. This seems to be a simple problem, but N-P Lemma does not apply since the hypotheses are not simple:

$$\Theta_0 = \{(\mu_0, \sigma^2), \sigma^2 > 0\}, \qquad \Theta_1 = \{(\mu_1, \sigma^2), \sigma^2 > 0\}.$$

What about problems where $\Theta \subset \mathcal{R}^1$ and MLR exists? No, not in most cases, even then. Suppose $X \sim P_\theta, \theta \in \Theta \subset \mathcal{R}^1$ and MLR exists in $T(x)$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. First consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. Then UMP test exists and is given by

$$\phi_1(x) = \begin{cases} 1 & \text{if } T(x) > C_1; \\ \gamma_1 & \text{if } T(x) = C_1; \\ 0 & \text{if } T(x) < C_1, \end{cases}$$

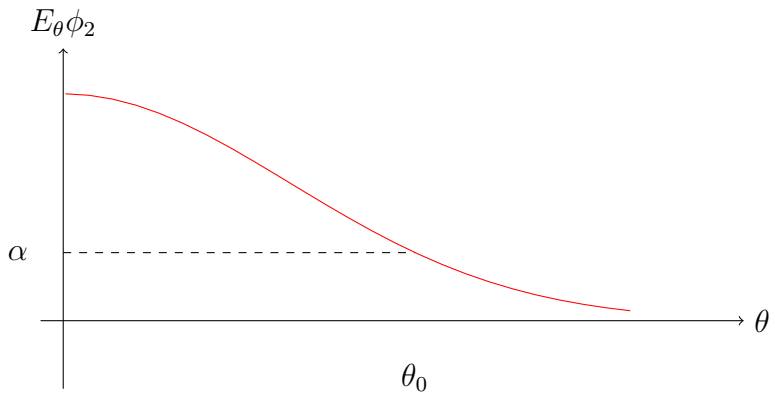
where C_1 and γ_1 are determined by $E_{\theta_0}\phi_1(X) = \alpha$. Observe the power function, $E_\theta\phi_1(X)$ of this test.



For all $\theta > \theta_0$, ϕ_1 maximizes the power among all level α tests. Does it maximize the power for any $\theta < \theta_0$? No. To see this, consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta < \theta_0$. Then the UMP test is given by

$$\phi_2(x) = \begin{cases} 1 & \text{if } T(x) < C_2; \\ \gamma_2 & \text{if } T(x) = C_2; \\ 0 & \text{if } T(x) > C_2, \end{cases}$$

where C_2 and γ_2 are determined by $E_{\theta_0}\phi_2(X) = \alpha$. This test, by definition, maximizes the power for all $\theta < \theta_0$.



Therefore, no single test ϕ uniformly maximizes the power for all $\theta \neq \theta_0$ subject to $E_{\theta_0}\phi(X) = \alpha$.

One may argue that ϕ_1 and ϕ_2 both are clearly not reasonable in this case, since the power falls below the level for certain alternatives. Eliminate these by putting the condition

$$E_\theta\phi(X) \geq \alpha \text{ for all } \theta \neq \theta_0.$$

In general, for testing

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_1$$

the test ϕ that maximizes the power $E_\theta\phi$ for $\theta \in \Theta_1$ subject to

$$\sup_{\theta\in\Theta_0} E_\theta\phi(X) \leq \alpha \text{ and } E_\theta\phi(X) \geq \alpha \text{ for all } \theta \neq \theta_0$$

is called the Uniformly Most Powerful Unbiased (UMPU) test. They exist under some very stringent conditions on the model density. These situations are rare. (See Lehmann, *TSH.*)

Generalized Likelihood Ratio Tests (GLRT)

UMP tests do not exist in all but simple situations. UMPU tests also may not exist. How does one conduct a test then? The approach that seems reasonable is to derive tests heuristically, and then check for their optimality.

Let $X \sim P_\theta, \theta \in \Theta$ having density $f(x|\theta)$. Consider testing

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_1.$$

Then the Generalized Likelihood Ratio statistic is defined to be

$$L(x) = \frac{\sup_{\theta\in\Theta_1} f(x|\theta)}{\sup_{\theta\in\Theta_0} f(x|\theta)}.$$

Reject H_0 if L is too large. This is a reasonable approach because we saw earlier that $\frac{f(x|\theta_1)}{f(x|\theta_0)}$ can be looked upon as evidence against $H_0 : \theta = \theta_0$ and in favour of $H_1 : \theta = \theta_1$. Now, $\sup_{\theta\in\Theta_1} f(x|\theta)$ is the best evidence for $H_1 : \theta \in \Theta_1$ whereas $\sup_{\theta\in\Theta_0} f(x|\theta)$ is the best evidence for $H_0 : \theta \in \Theta_0$. Suppose $\Theta = \Theta_0 \cup \Theta_1$. Consider

$$\lambda(x) = \frac{\sup_{\theta\in\Theta} f(x|\theta)}{\sup_{\theta\in\Theta_0} f(x|\theta)}.$$

Then $\lambda(x) = \max\{L(x), 1\}$ since

$$\lambda(x) = \begin{cases} 1 & \text{if } \sup_{\theta\in\Theta_0} f(x|\theta) \geq \sup_{\theta\in\Theta_1} f(x|\theta); \\ L(x) & \text{if } \sup_{\theta\in\Theta_0} f(x|\theta) < \sup_{\theta\in\Theta_1} f(x|\theta). \end{cases}$$

Note that

$$\lambda_n(x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\hat{\theta})}{f(x_1, \ldots, x_n|\hat{\theta}_0)},$$

where
$\hat{\theta}$ = MLE of θ in Θ,
$\hat{\theta}_0$ = MLE of θ in Θ_0.
If an increasing function of $\lambda(\mathbf{X})$ has a standard distribution under H_0, then it can be used to construct the test.

Generalized Likelihood Ratio Tests (GLRT)

UMP tests do not exist in all but simple situations. UMPU tests also may not exist. How does one conduct a test then? The approach that seems reasonable is to derive tests heuristically, and then check for their optimality.

Let $X \sim P_\theta, \theta \in \Theta$ having density $f(x|\theta)$. Consider testing

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_1.$$

Then the Generalized Likelihood Ratio statistic is defined to be

$$L(x) = \frac{\sup_{\theta\in\Theta_1} f(x|\theta)}{\sup_{\theta\in\Theta_0} f(x|\theta)}.$$

Reject H_0 if L is too large. This is a reasonable approach because we saw earlier that $\frac{f(x|\theta_1)}{f(x|\theta_0)}$ can be looked upon as evidence against $H_0 : \theta = \theta_0$ and in favour of $H_1 : \theta = \theta_1$. Now, $\sup_{\theta\in\Theta_1} f(x|\theta)$ is the best evidence for $H_1 : \theta \in \Theta_1$ whereas $\sup_{\theta\in\Theta_0} f(x|\theta)$ is the best evidence for $H_0 : \theta \in \Theta_0$. Suppose $\Theta = \Theta_0 \cup \Theta_1$. Consider

$$\lambda(x) = \frac{\sup_{\theta\in\Theta} f(x|\theta)}{\sup_{\theta\in\Theta_0} f(x|\theta)}.$$

Then $\lambda(x) = \max\{L(x), 1\}$ since

$$\lambda(x) = \begin{cases} 1 & \text{if } \sup_{\theta\in\Theta_0} f(x|\theta) \geq \sup_{\theta\in\Theta_1} f(x|\theta); \\ L(x) & \text{if } \sup_{\theta\in\Theta_0} f(x|\theta) < \sup_{\theta\in\Theta_1} f(x|\theta). \end{cases}$$

Note that

$$\lambda_n(x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\hat{\theta})}{f(x_1, \ldots, x_n|\hat{\theta}_0)},$$

where
$\hat{\theta}$ = MLE of θ in Θ,
$\hat{\theta}_0$ = MLE of θ in Θ_0.
If an increasing function of $\lambda(\mathbf{X})$ has a standard distribution under H_0, then it can be used to construct the test.

Example. $X_1, \ldots, X_n$ i.i.d $N(\mu, \sigma^2)$, both μ and σ^2 unknown. Test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. Then

$$\Theta_0 = \{(\mu = 0, \sigma^2), \sigma^2 > 0\}, \qquad \Theta_1 = \{(\mu, \sigma^2), -\infty < \mu < \infty, \mu \neq 0, \sigma^2 > 0\}.$$

MLE are needed to compute the GLR statistic: unrestricted and, restricted to Θ_0.

$$\hat{\theta} = (\hat{\mu} = \bar{X}, \qquad \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2),$$

$$\hat{\theta}_0 = (\hat{\mu}_0 = 0, \qquad \hat{\sigma^2}_0 = \frac{1}{n}\sum_{i=1}^{n} X_i^2).$$

Therefore,

$$\begin{aligned}
\lambda(\mathbf{x}) &= \frac{f(\mathbf{x}|\hat{\theta})}{f(\mathbf{x}|\hat{\theta}_0)} \\
&= \frac{(2\pi)^{-n/2}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{-n/2}\exp\left(-\frac{1}{2\hat{\sigma}^2}\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\hat{\mu})^2\right\}\right)}{(2\pi)^{-n/2}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right)^{-n/2}\exp\left(-\frac{1}{2\hat{\sigma}_0^2}\left\{\sum_{i=1}^{n}x_i^2\right\}\right)} \\
&= \frac{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{-n/2}}{\left(\sum_{i=1}^{n}x_i^2\right)^{-n/2}} = \left(\frac{\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^{n/2} = \left(1 + \frac{n\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^{n/2}.
\end{aligned}$$

Note that $\lambda(\mathbf{x})$ is an increasing function of

$$T^2 = \frac{n\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)},$$

and therefore of $|T|$, where

$$T = \frac{\sqrt{n}\bar{X}}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1)}} \sim t_{n-1}, \text{ if } H_0 \text{ is true.}$$

Therefore the GLRT rejects H_0 if

$$\left|\frac{\sqrt{n}\bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)}}\right| > t_{n-1}(1-\alpha/2).$$

Example. $X_1, \ldots, X_n$ i.i.d $N(\mu, \sigma^2)$, σ^2 known. Test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Derive the GL statistic and show that GLRT rejects H_0 when

$$\left|\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma}\right| > z_{1-\alpha/2}.$$

Note. Classical or Frequentist test procedure (which is what we have been discussing) is predetermined. $\mathbf{x}$ or data is used only to check whether it falls in the rejection region or not. Exact value of $\mathbf{x}$ is not relevant. What is reported is the level α and whether $\phi(\mathbf{x})$ is 1, γ or 0. If H_0 is true, then the test procedure will ensure that, if used over and over again, the long-run average rejection rate is α.

Confidence Sets and Hypothesis Tests

For a confidence set, we want $S(X) \subset \Theta$ such that

$$P_\theta\left(\theta \in S(X)\right) \geq 1-\alpha \text{ for all } \theta \in \Theta.$$

Then $S(X)$ is said to be 100(1-α)% confidence set for θ. Suppose we have available to us a test procedure for testing $H_0 : \theta = \theta'$ versus $H_1 : \theta \neq \theta'$ for any $\theta' \in \Theta$. Let $A(\theta') \subset \mathcal{X}$ be the acceptance region of the level α test of $H_0 : \theta = \theta'$ versus $H_1 : \theta \neq \theta'$. Define

$$\begin{aligned} S(x) &= \{\theta' \in \Theta \text{ such that } x \in A(\theta')\} \\ &= \{ \text{ all } \theta' \text{ for which } H_0 : \theta = \theta' \text{ will be accepted if } x \text{ is observed. } \} \end{aligned}$$

Then $\theta \in S(x)$ iff $x \in A(\theta)$. Therefore,

$$P_\theta\left(\theta \in S(X)\right) = P_\theta\left(X \in A(\theta)\right) \geq 1-\alpha.$$

Therefore, $S(X)$ is 100(1-α)% confidence set for θ.

Example. $X_1, \ldots, X_n$ i.i.d $N(\mu, \sigma^2)$, σ^2 known. Then $\bar{X}$ is sufficient and $\bar{X} \sim N(\mu, \sigma^2/n)$. Recall that GLRT for testing $H_0 : \mu = \mu'$ versus $H_1 : \mu \neq \mu'$ rejects H_0 when

$$\left|\frac{\sqrt{n}(\bar{x}-\mu')}{\sigma}\right| > z_{1-\alpha/2}.$$

Therefore, its acceptance region is

$$A(\mu') = \left\{\bar{x} : \left|\frac{\bar{x}-\mu'}{\sigma/\sqrt{n}}\right| \leq z_{1-\alpha/2}\right\}.$$

Hence,

$$S(\bar{x}) = \left\{\mu' : \left|\frac{\bar{x}-\mu'}{\sigma/\sqrt{n}}\right| \leq z_{1-\alpha/2}\right\}.$$

Therefore the resulting confidence set (interval) is

$$S(\bar{X}) = \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Example. $X_1, \ldots, X_n$ i.i.d $N(\mu, \sigma^2)$, both μ and σ^2 unknown. It is of interest to construct a confidence set for μ. $(\bar{X}, S^2 = \sum_{i=1}^n (X_i - \bar{X})^2)$ is sufficient. Let $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2)$. Consider the GLRT for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Its acceptance region is

$$\begin{aligned} A(\mu_0) &= \left\{ (\bar{x}, s^2) : \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \leq t_{n-1}(1 - \alpha/2) \right\}, \text{ so that} \\ S(\bar{x}, s^2) &= \left\{ \mu : \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \leq t_{n-1}(1 - \alpha/2) \right\}. \end{aligned}$$

This yields the confidence interval:

$$S(\bar{X}, s^2) = \bar{X} - t_{n-1}(1 - \alpha/2) \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1}(1 - \alpha/2) \frac{s}{\sqrt{n}}.$$

Bayesian Statistical Inference

An example of statistical inference is as follows.

Example. Consider a production process where the overall proportion of defectives, θ, is of interest. A random sample of size n of products from this process is checked for defectives. Let X denote the number of defectives found in the sample. Then $X \sim \text{Binomial}(n, \theta)$. i.e.,

$$P(X = x|\theta) = f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, x = 0, 1, \ldots, n.$$

The unknown quantity θ indexes the model P_θ for X. What is the 'best fit' for θ if $X = x$ is observed? We may find the mle of θ: $l(\theta|x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, as a function of θ for given x is the likelihood function of θ. This is a measure of how likely that the model with proportion θ produced the data x. With this interpretation, it makes sense to maximize this likelihood function to estimate θ.
$\hat{\theta}_{\text{mle}} : \max_\theta l(\theta|x)$

In the example, $l(\theta|x) = c(x)\theta^x(1-\theta)^{n-x}$ has unique maximum at $\hat{\theta} = x/n =$ sample proportion of defectives. Good! Now we have an estimate (for θ). How good is this estimate? What is the estimation error? What is a confidence interval for θ?

These questions cannot be answered by the likelihood approach. In the *Frequentist* approach, they require the sampling distribution of the estimator – on repeated sampling how does $\hat{\theta}$ behave? Let us consider confidence statements.

Confidence Set. Any random set (related to X) which captures θ with a prescribed level of confidence.

If n is large, a 95% confidence interval for θ is $\hat{\theta} \pm 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n}$. This is because, since $X \sim \text{Binomial}(n, \theta)$, for large n, X/n is approximately $N(\theta, \theta(1-\theta)/n)$, or

$$\frac{X/n - \theta}{\sqrt{\theta(1-\theta)/n}} \sim N(0, 1).$$

Then, approximately,

$$P\left(|\frac{X}{n} - \theta| \leq 1.96\sqrt{\theta(1-\theta)/n}\right) = 0.95.$$

For large n, $\hat{\theta}(X)$ is close to θ, so

$$P\left(|\frac{X}{n} - \theta| \leq 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n}\right) = 0.95.$$

Therefore,

$$\theta \in \left(\hat{\theta}(X) \pm 1.96\sqrt{\hat{\theta}(X)(1-\hat{\theta}(X))/n}\right)$$

with probability 0.95 for all θ.

There are two issues with this approach. First, instead of binomial sampling, suppose we did inverse binomial sampling. i.e., check products until x (same count as what we got with binomial sampling) defectives are spotted. Then, we have:
binomial likelihood: $\theta^x(1-\theta)^{n-x}$
inverse binomial likelihood: $\theta^x(1-\theta)^y$

Suppose $n - x = y$; then the observed likelihood is the same for both the models, so $\hat{\theta} = x/n$ for both. However, the confidence intervals will be different since the variances of $\hat{\theta}$ will be different.

The second issue involves the interpretation of the confidence interval. If we sample again and again from the production process and construct 95% confidence intervals with each of the samples, in about 19 cases out of 20 the intervals will contain θ. However, for the given sample we get a fixed (not random) interval: $\hat{\theta}(x) \pm 1.96\sqrt{\hat{\theta}(x)(1-\hat{\theta}(x))/n}$. What is the interpretation for this interval? Surely, θ can only lie in that interval with probability either 0 or 1.

Classical statistics is frequentist – talks only in terms of optimality w.r.t. long-run average behaviour of statistical procedures. It cannot condition on data, and cannot interpret procedures with respect to fixed data. Why is conditioning needed if we have procedures which have good long-run behaviour?

The need for conditioning on data.

First of all, repetition of experiments (as in frequentist sense) may not be meaningful – what are the chances of another catastrophe like covid-19? Another point is illustrated below.

Example. Let X_1 and X_2 be i.i.d. with

$$X_i = \begin{cases} \theta - 1 & \text{with probability } 1/2; \\ \theta + 1 & \text{with probability } 1/2, \end{cases}$$

where $-\infty < \theta < \infty$. Define a confidence set for θ as follows.

$$C(X_1, X_2) = \begin{cases} \{\frac{X_1+X_2}{2}\} & \text{if } X_1 \neq X_2; \\ \{X_1 - 1\} & \text{if } X_1 = X_2. \end{cases}$$

Then, we get,

$$\begin{aligned} P_\theta(\theta \in C) &= P_\theta\left(\theta = \frac{X_1+X_2}{2} \middle| X_1 \neq X_2\right) P_\theta(X_1 \neq X_2) \\ &\quad + P(\theta = X_1 - 1 \mid X_1 = X_2) P_\theta(X_1 = X_2) \\ &= 1 \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{3}{4}, \end{aligned}$$

for all θ, so $C(X_1, X_2)$ is a 75% confidence set for θ. Thus, if we use this procedure repeatedly, we will be correct about θ three times out of four. But, if we observe $x_1 \neq x_2$, are we not 100% sure that $\theta = (x_1 + x_2)/2$? Why say that we are only 75% sure? This shows that ther are situations where pre-experimental optimality is not the appropriate approach for inference. However, the frequentist approach does not permit any (observed) data dependent confidence statements. There are many examples like this.

Example. To estimate μ in $N(\mu, \sigma^2)$, toss a fair coin. Have a sample of size $n = 2$ if it is a head and take $n = 1000$ if it is a tail. An unbiased estimate of μ is $\bar{X}_n = \sum_{i=1}^n X_i/n$ with variance $= \frac{1}{2}\{\frac{\sigma^2}{2} + \frac{\sigma^2}{1000}\} \sim \frac{\sigma^2}{4}$. Suppose it was a tail. Would you believe $\sigma^2/4$ is a measure of accuracy?

Example. Let X_1, X_2 be i.i.d. $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. Let $\bar{X} \pm C$ be a 95% confidence interval, $C > 0$ being suitably chosen. Suppose $X_1 = 2$ and $X_2 = 1$. Then we know for sure $\theta = (X_1 + X_2)/2$ and hence $\theta \in (\bar{X} - C, \bar{X} + C)$. Should we still claim we have only 95% confidence that the confidence interval covers θ?

How is then conditioning on data to be done? Consider the example below.

Example. A laboratory test (such as RAT for COVID-19) is needed to check whether a person has a particular disease. The result of the test is either positive ($x = 1$) or negative ($x = 0$). Let θ_1 denote 'disease is present', θ_2 be 'not present'. $P(X = x|\theta)$ is as follows.

	$x = 0$	$x = 1$
θ_1	0.2	0.8
θ_2	0.7	0.3

The test is not fool-proof. 30% false positives and 20% false negatives appear.

Now suppose a patient is sent to the laboratory for this test and the test result comes out positive. What is the doctor to conclude regarding the presence or absence of the disease? Note that the question of interest is not whether the test result is positive or negative. Instead, what are the chances of the disease being present? i.e., $P(\theta = \theta_1|X = 1) = ?$

What we have are $P(X = 1|\theta = \theta_1)$ and $P(X = 1|\theta = \theta_2)$. We have the 'wrong' conditional probabilities! They need to be reversed or inverted. But how?

Suppose, in the concerned community, the disease is present in 5% of the cases. i.e., $P(\theta = \theta_1) = 0.05$. This is, however, not part of the sample data. This is pre-experimental. The doctor has this information from experience in the field. Now,

$$P(\theta = \theta_1|X = x) = \frac{P(\theta = \theta_1 \text{ and } X = x)}{P(X = x)},$$

and

$$P(X = x) = P(X = x|\theta_1)P(\theta = \theta_1) + P(X = x|\theta_2)P(\theta = \theta_2),$$

so applying the *Bayes Theorem*,

$$P(\theta = \theta_1|X = x) = \frac{P(X = x|\theta_1)P(\theta = \theta_1)}{P(X = x|\theta_1)P(\theta = \theta_1) + P(X = x|\theta_2)P(\theta = \theta_2)}. \quad (1)$$

Therefore,

$$P(\theta = \theta_1|X = 1) = \frac{0.8 \times 0.05}{0.8 \times 0.05 + 0.3 \times 0.95} = \frac{0.04}{0.04 + 0.237} = 0.123,$$

and $P(\theta = \theta_2|X = 1) = 0.877$. Positive blood test indicates only a 12.3% chance of disease being present in a random member of the community, so further diagnostic measures may be needed. On the other hand, this is important since the risk has more than doubled, from 5% to 12.3%.

(1) shows how to 'invert' the given conditional probabilities, $P(X = x|\theta)$ to derive the desired conditional probabilities, $P(\theta = \theta_i|X = x)$, which is an application of the Bayes Theorem. Since this involves an inversion, the name, *Theory of inverse probability* is used for statistical inference based on this approach. This was the usage at the time of Bayes and Laplace – late 18th century, before Fisher and Pearson. However, these days it is known simply as Bayesian inference. Note that, to obtain $P(\theta = \theta_1|X = 1)$, it is

essential to have $P(\theta = \theta_1)$ (and hence $P(\theta = \theta_0) = 1 - P(\theta = \theta_1)$). Where does this come from, and what kind of a probability is this?

Ingredients of Bayesian inference

likelihood function, $l(\theta|x) \propto f(x|\theta)$

prior distribution, $\pi(\theta) = \begin{cases} \text{probability mass function, if } \theta \text{ is discrete;} \\ \text{probability density function, if } \theta \text{ is continuous} \end{cases}$

What are the implications of using a prior distribution (π) on the unknown quantity θ?

There is usually some information (prior to sample data collection) available about θ; sometimes this may be precise but not often. Thus, there is usually a lot of uncertainty about θ. What is a good way to quantify uncertainty? Probability is the only well-accepted mathematical approach. This does not necessarily mean that θ is random. Probability is a tool to incorporate uncertainty, that is all. There is no requirement that probabilities must have a relative frequency interpretation based on a repeatable experiment. However, past data is a common source for prior probabilities. Recall the example on laboratory test for diagnosis. In this case, the prior probability, $P(\theta = \theta_1) = 0.05$ in the concerned population is simply the prevalence of the disease, about which medical experts are expected to have information. In the quality control example, the manufacturer wants to monitor the quality of his products. Random samples are taken periodically to estimate the proportion p of defectives. But the manufacturer has a lot of other information about his production process including past data on the proportion of defectives.

An important aspect of Bayesian inference is this: Whether useful prior information is available or not, a prior distribution is needed for the implementation of conditioning on data using the Bayes theorem.

Technically, a Bayesian takes the view that all unknown quantities, namely the unknown parameter and the data before observation, have a probability distribution. For the data, the distribution, given θ, comes from a model that arises from past experience in handling similar data as well as subjective judgment. The distribution of θ arises as a quantification of the Bayesian's knowledge and belief. If her knowledge and belief are weak, she may fall back on a common objective distribution in such situations.

Bayesian Inference

Informally, to make inference about θ is to learn about the unknown θ from data X, i.e., based on the data, explore which values of θ are probable, what might be plausible numbers as estimates of different components of θ and the extent of uncertainty associated with such estimates. In addition to having a model $f(x|\theta)$ yielding a likelihood function, the Bayesian needs a distribution for θ. The distribution is called a prior distribution or simply a prior because it quantifies her uncertainty about θ prior to seeing data. The prior may represent a blending of her subjective belief and knowledge, in which case it would be a subjective prior. Alternatively, it could be a conventional prior supposed to represent small or no information. Such a prior is called an objective prior.

Given all the above ingredients, the Bayesian calculates the conditional probability density of θ given $X = x$ by Bayes formula. First, the joint density of X and θ is $h(x, \theta) = f(x|\theta)\pi(\theta)$ on $\mathcal{X} \times \Theta$. The marginal (or predictive) density of X is

$$m(x) = \int_\Theta f(x|\theta)\pi(\theta)\, d\theta, \text{ or } \sum_\theta f(x|\theta)\pi(\theta).$$

Then,

$$\begin{aligned} \pi(\theta|x) &= \frac{h(x,\theta)}{m(x)} = \frac{f(x|\theta)\pi(\theta)}{m(x)} \\ &\propto f(x|\theta)\pi(\theta) \text{ for observed data, } x, \end{aligned}$$

is called the post-experimental or posterior distribution (density) of θ given x.

This summarizes all the post-data information about θ, and is a quantification of our uncertainty about θ in the light of data. The transition from $\pi(\theta)$ to $\pi(\theta|x)$ is what we have learnt from the data. All inferences about θ must be based on this posterior distribution. Nothing beyond $l(\theta|x)$ from the experiment is needed. Two different experiments with the same likelihood lead to identical inference. $\pi(\theta|x) \propto l(\theta|x) \propto f(x|\theta)$ if $\pi(\theta) \equiv 1$. In this case $\pi(\theta|x)$ does not use any information other than what is in $l(\theta|x)$.

Suppose $T = T(X)$ is sufficient for θ (or $P_\theta, \theta \in \Theta$) or for $f(x|\theta)$, $\theta \in \Theta$.

Theorem. Posterior distribution of θ given $X = x$ depends on x only through $T(x)$.

Proof. We will assume the factorization theorem: $f(x|\theta) = g(T(x), \theta)h(x)$. If $T(x) = t$, then

$$\begin{aligned}\pi(\theta|x) &= \frac{f(x|\theta)\pi(\theta)}{\int f(x|u)\pi(u)\,du} \\ &= \frac{g(T(x),\theta)h(x)\pi(\theta)}{\int g(T(x),u)h(x)\pi(u)\,du} \\ &= \frac{g(t,\theta)\pi(\theta)}{\int g(t,u)\pi(u)\,du}.\end{aligned}$$

Example. Consider an urn with Np red and $N(1-p)$ black balls, p is unknown but N is a known large number. Balls are drawn at random one by one and with replacement, selection is stopped after n draws. For $i = 1, 2, \ldots, n$, let

$$Y_i = \begin{cases} 1 & \text{if the } i\text{th ball drawn is red;} \\ 0 & \text{otherwise.} \end{cases}$$

Then Y_i's are i.i.d Binomial$(1, p)$, i.e., *Bernoulli* with probability of success p. Therefore the likelihood function for p given the data is proportional to

$$f(y_1, \ldots, y_n) = p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i} = p^x(1-p)^{n-x},$$

where $x = \sum_{i=1}^n y_i$. Since $X = \sum_{i=1}^n Y_i$ (the number of red balls drawn) is sufficient for p, we get the same likelihood function (the part involving p) if we assume that we have observed $X = x$ from a Binomial(n, p). Let p have a prior distribution $\pi(p)$. We will consider a family of priors for p that simplifies the calculation of posterior distribution and then consider some commonly used priors from this family. Let

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}, \quad 0 \le p \le 1; \alpha > 0, \beta > 0.$$

This is the density of the Beta distribution. Equivalently, under the prior distribution, the unknown parameter, $p \sim$ Beta(α, β). (Note that for convenience we take p to assume all values between 0 and 1, rather than only $0, 1/N, 2/N$, etc.) The prior mean and variance are $\alpha/(\alpha+\beta)$ and $\alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$, respectively, which may be obtainable from past data.

To derive the posterior density, note that

$$\begin{aligned}h(x,p) &= \binom{n}{x}p^x(1-p)^{n-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}, x = 0, \ldots, n; 0 < p < 1 \\ &= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{x+\alpha-1}(1-p)^{n-x+\beta-1}, x = 0, \ldots, n; 0 < p < 1,\end{aligned}$$

so that

$$\begin{aligned} m(x) &= \int_0^1 h(x,p)\,dp = \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 p^{x+\alpha-1}(1-p)^{n-x+\beta-1}\,dp \\ &= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}, x = 0,\ldots,n. \end{aligned}$$

Therefore, we obtain,

$$\begin{aligned} \pi(p|x) &= \frac{h(x,p)}{m(x)} \\ &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}p^{x+\alpha-1}(1-p)^{n-x+\beta-1}, 0<p<1. \end{aligned}$$

i.e., $p|X = x \sim \text{Beta}(x+\alpha, n-x+\beta)$. Note, however, that the computation of $m(x)$ is not needed here to derive the posterior density; it can be deduced from simply noting the functional form of the density in $h(x,p)$, namely, $p^{x+\alpha-1}(1-p)^{n-x+\beta-1}$, which is just the (unnormalized) density of the Beta distribution. This is due to the choice of the prior, and will be explored further later. Before that note the following. As a straightforward and immediate estimate of p, one could look at the most 'probable' value of p (under $\pi(p|x)$). The *highest posterior density* or HPD estimate of p is the value, denoted $\hat{p}_{\text{ hpd}}$, which maximizes $\pi(p|x)$. In the example above, $\hat{p}_{\text{ hpd}} = (x+\alpha-1)/(n+\alpha+\beta-2)$.
(i) If we take $\alpha = 1 = \beta$, i.e., $\pi(p) \equiv 1$, we get

$$\pi(p|x) = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}p^x(1-p)^{n-x}, 0<p<1.$$

As a function of p, $\pi(p|x)$ and $l(p|x)$ are the same. Therefore,
MLE of $p = x/n =$ HPD estimate of p. i.e., $\hat{p}_{\text{ mle}} = \hat{p}_{\text{ hpd}}$, but their interpretations are different.
Given x, $\hat{\theta}_{\text{ hpd}}$ is the most probable value of θ, or $P_{\hat{\theta}_{\text{ hpd}}}$ is most 'probably' the correct model (for X), whereas $\hat{p}_{\text{ mle}}$ is that value of θ, or the parameter of that model which most 'likely' produced x.

Now, coming back to the Beta prior for Binomial, note that what simplied the computation of the posterior density is that the prior and likelihood have the same functional form.

Now, coming back to the Beta prior for Binomial, note that what simplied the computation of the posterior density is that the prior and likelihood have the same functional form.

Conjugate families of prior distributions. Let $\mathcal{F}$ denote a class of density functions $f(x|\theta)$. A class $\mathcal{P}$ of prior densities is said to be a conjugate family for $\mathcal{F}$ if $\pi(.|x) \in \mathcal{P}$ for all $f \in \mathcal{F}$ and $\pi \in \mathcal{P}$.

Example. $X|\theta \sim$ Binomial(n, θ). Then

$$\mathcal{F} = \{ \text{ all Binomial}(n, \theta),\ n = 1, 2 \dots \},$$

$$\mathcal{P} = \{ \text{ all Beta}(a, b),\ a > 0,\ b > 0. \}.$$

If $\pi \in \mathcal{P}$ and $f \in \mathcal{F}$, then $\theta \sim$ Beta(a, b) for some $a > 0$, $b > 0$, and $X|f \sim$ Binomial(n, θ) for some $n > 0$, so $\theta|X = x \sim$ Beta$(x+a, n-x+b) \in \mathcal{P}$.

Example. $X|\theta \sim N(\theta, \sigma^2)$, σ^2 known. Consider $\theta \sim N(\mu, \tau^2)$, μ, τ^2 known. Then, $\theta|X = x \sim$? $h(x, \theta) = \frac{1}{2\pi}\frac{1}{\sigma\tau}\exp(-\frac{1}{2}[\frac{(x-\theta)^2}{\sigma^2} + \frac{(\theta-\mu)^2}{\tau^2}])$. One can complete the square for θ, proceed using calculus to find $m(x)$ and then determine $\pi(\theta|x)$. We will use a property of the multivariate normal instead. Note that $X|\theta \sim N(\theta, \sigma^2)$ is equivalent to $X = \theta + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ independent of θ. Since $\theta \sim N(\mu, \tau^2)$, we can obtain the joint bivariate normal distribution for X and θ as:

$$\begin{pmatrix} X \\ \theta \end{pmatrix} = \begin{pmatrix} \theta + \epsilon \\ \theta \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2 + \tau^2 & \tau^2 \\ \tau^2 & \tau^2 \end{pmatrix} \right),$$

because $E(X) = E(\theta + \epsilon) = \mu$, $Var(X) = Var(\theta + \epsilon) = \sigma^2 + \tau^2$, $Cov(X, \theta) = Cov(\theta + \epsilon, \theta) = Cov(\theta, \theta) = Var(\theta) = \tau^2$. Therefore,

$$\begin{aligned} \theta|X = x &\sim N\left(\mu + \frac{\tau^2}{\sigma^2 + \tau^2}(x - \mu), \tau^2 - \frac{\tau^4}{\sigma^2 + \tau^2}\right) \\ &= N\left(\frac{\tau^2}{\sigma^2 + \tau^2}x + \frac{\sigma^2}{\sigma^2 + \tau^2}\mu, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} = \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right). \end{aligned}$$

Remark. If instead $X_1, X_2, \dots, X_n$ are i.i.d $N(\theta, \sigma^2)$, σ^2 known, in the above example, then $\bar{X}$ is sufficient for θ and $\bar{X}|\theta \sim N(\theta, \sigma^2/n)$. Therefore replace X by $\bar{X}$ and σ^2 by σ^2/n above.

Question. Are conjugate priors reasonable for expressing prior information? If they represent actual prior information, there is no problem. Otherwise they are easy to work with but not robust – prior and likelihood have the

same functional form, so similar weight is given to prior and sample data. Mixtures of conjugate priors are much better, and computations are not too difficult because MCMC sampling methods are available.

Noninformative or vague priors

Example. $X_1, X_2, \ldots, X_n$ are i.i.d $N(\theta, \sigma^2)$, σ^2 known. Inference on θ is of interest. Consider $\pi(\theta) \equiv 1$ as an expression of lack of prior information. This is not a probability density, but that of a limit of $N(0, \tau^2)$ as $\tau^2 \to \infty$. Since $\bar{X}|\theta \sim N(\theta, \sigma^2/n)$, $h(\bar{x}, \theta) = f(\bar{x}|\theta)\pi(\theta) = f(\bar{x}|\theta)$, we get

$$m(\bar{x}) = \int_{-\infty}^{\infty} f(\bar{x}|\theta)\, d\theta = \int_{-\infty}^{\infty} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp(-\frac{n}{2\sigma^2}(\theta - \bar{x})^2)\, d\theta = 1.$$

Therefore,

$$\pi(\theta|\bar{x}) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp(-\frac{n}{2\sigma^2}(\theta - \bar{x})^2),$$

so that $\theta|\bar{X} = \bar{x} \sim N(\bar{x}, \sigma^2/n)$. Why not then use $\pi(\theta) \equiv c$ whenever no prior information is available, or when a noninformative prior is needed? Observe the problem with this approach. Suppose $\theta > 0$ and consider $\pi(\theta) \equiv c$. Then $\int_0^{\infty} \pi(\theta)\, d\theta = \infty$. Consider the reparametrization: $\eta = \eta(\theta) = \exp(\theta)$. Then $\theta = \log(\eta)$, so $d\theta = d\eta/\eta$. Therefore, the prior density on η is given by

$$\pi^*(\eta) = \pi(\log(\eta))\frac{1}{\eta} = \frac{c}{\eta},$$

which is not uniform as is the case with $\pi(\theta)$. If we did not have information about θ how did we get information about a transform of it? There are no strictly noninformative priors, there are default and reference priors for objective choice, for example, the Jeffreys' prior.

Jeffreys' Prior

The idea is to employ a prior which contains the minimal amount of prior information needed to be able to conduct Bayesian analysis for the given experiment. Let $f(x|\theta)$ be the model density of $X|\theta$ for which $I(\theta)$ be the Fisher Information. Then the Jeffreys' prior in this case is defined to be

$$\pi(\theta) = (I(\theta))^{1/2} \text{ if } \theta \text{ is univariate; more generally } \pi(\theta) = |I(\theta)|^{1/2}.$$

Example. Suppose $X|\theta \sim N(\theta, 1)$. Then $I(\theta) = 1$ since $\frac{\partial}{\partial\theta} \log f(x|\theta) = x - \theta$ and $\frac{\partial^2}{\partial\theta^2} \log f(x|\theta) = -1$. Therefore, $\pi(\theta) \equiv C$ is the Jeffreys' prior in this case.

It may be verified that this prior is invariant with respect to any one-one differentiable transformations on θ.

In practice, however, Jeffreys' employed a different group invariance argument. For any location family, his suggestion for the prior on the location parameter is the translation invariant (thus indicating lack of information) measure. Note that this agrees with the Jeffreys' prior above for the $N(\theta, 1)$. For a location-scale family $f(x|\theta, \sigma)$, his argument is as follows. If σ is fixed, then the prior for the location θ is $\pi_1(\theta|\sigma) = 1$ as above. Now note that, if σ is a scale parameter for a positive r.v. Y then $\log \sigma$ is a location for $\log Y$. So, the prior for $\log \sigma$ is a constant which in turn gives $\pi_2(\sigma) = 1/\sigma$. Thus $\pi(\theta, \sigma) = \pi_1(\theta|\sigma)\pi_2(\sigma) = 1/\sigma$. This is the right invariant Haar measure for the affine group of transaformations whereas the Jeffreys' prior according to the formal definition above is the left invariant Haar measure which is $1/\sigma^2$. Most Bayesians prefer the former one, for various reasons including posterior consistency.

Estimation

$\pi(\theta|$ data) is the probability density of θ having seen the data. It contains all the post-experimental information about θ. Any Bayesian inference about θ must be based on it. Note the following in this context.
(i) We have an actual probability distribution on the unknown *parameters* to describe their uncertainty, namely $\pi(\theta|$ data).
(ii) We can readily make probability statements on where θ lies using this distribution.

A Bayesian can simply report the posterior distribution, or report some summary descriptive measures associated with the posterior distribution. For example, as mentioned previously, $\hat{\theta}_{\text{ hpd}}$, is one such measure which is analogous to the MLE. If $\pi(\theta|x)$ is unimodal, this may be a reasonable estimate for θ. However, the usual Bayes estimate of θ is $E(\theta|x)$, which is a measure of location or centre of $\pi(\theta|x)$. For this estimate, the precision may be measured by the posterior standard deviation, $s.d.(\theta|x)$, which is a standard measure of spread or dispersion. Note that (for a real valued θ)

$$E(\theta|x) = \int_{-\infty}^{\infty} \theta\pi(\theta|x)\, d\theta$$

and the posterior variance

$$\begin{aligned} Var(\theta|x) &= E\{(\theta - E(\theta|x))^2|x\} \\ &= \int_{-\infty}^{\infty} (\theta - E(\theta|x))^2\pi(\theta|x)\, d\theta. \end{aligned}$$

Consider the binomial example again: $X|\theta \sim \text{Binomial}(n, p)$, for which the prior is $p \sim \text{Beta}(\alpha, \beta)$. Then $p|X = x \sim \text{Beta}(x + \alpha, n - x + \beta)$. Therefore, the posterior mean and variance are

$$\begin{aligned} E(p|x) &= (\alpha + x)/(\alpha + \beta + n), \\ Var(p|x) &= \frac{(\alpha + x)(\beta + n - x)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}, \\ s.d(p|x) &= \sqrt{\frac{(\alpha + x)(\beta + n - x)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}}. \end{aligned}$$

The posterior mean can be rewritten as a weighted average of the prior mean and MLE.

$$\frac{(\alpha + \beta)}{(\alpha + \beta + n)} \frac{\alpha}{(\alpha + \beta)} + \frac{n}{(\alpha + \beta + n)} \frac{x}{n}.$$

Once again, the importance of both the prior and the data comes out, the relative importance of the prior and the data being measured by $(\alpha + \beta)$ and n. It will not escape one's attention that if n is large then the posterior mean is approximately equal to the MLE, $\hat{p}_{\text{ mle}} = x/n$ and the posterior variance is quite small, so the posterior distribution is concentrated around $\hat{p}_{\text{ mle}}$ for large n. We can interpret this as an illustration of a fact that when we have lots of data, the data tend to wash away the influence of the prior.

Posterior inference

Model: $X|\theta$ has density $f(x|\theta)$
Prior: θ has density $\pi(\theta)$

Posterior density:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}$$

This is the probability density for θ after observing the data, $X = x$

It contains all the information on θ after observing the data

All inferences on θ must be based on this

Optimal estimators

$L(\theta, a) = (\theta - a)^2$ is the (squared error) loss when θ is estimated with a number a.

What is the best estimator for θ under this loss?

$\min_a E\left[(\theta - a)^2 | x\right] = ?$

$$\begin{aligned} E\left[(\theta - a)^2 | x\right] &= E\left[\{(\theta - E(\theta|x)) + (E(\theta|x) - a)\}^2 | x\right] \\ &= E\left[(\theta - E(\theta|x))^2\right] + (E(\theta|x) - a)^2 \\ &\geq E\left[(\theta - E(\theta|x))^2\right] = Var(\theta|x), \end{aligned}$$

with equality iff $a = \delta(x) = E(\theta|x)$

What is the optimal estimator if

$$L(\theta, a) = |\theta - a|?$$

Credible sets

$\pi(\theta|x)$ is the probability density for θ (after observing data, $X = x$)

Any subset $C = C(x) \subset \Theta$ which has probability

$$P(\theta \in C|x) = \int_C \pi(\theta|x)\, d\theta = 1 - \alpha$$

is a $100(1-\alpha)\%$ credible set for θ

Frequentist Confidence sets:

$$P_\theta(\underline{T}(X) \leq \theta \leq \bar{T}(X)) = 1 - \alpha$$

for all θ

Example.

$X_1, X_2, \cdots, X_n$ i.i.d. $N(\theta, \sigma^2)$, σ^2 is known. $\theta \sim N(\mu, \tau^2)$.

$$\theta|\mathbf{X} = \mathbf{x} \sim N\left(\frac{\tau^2}{\tau^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\mu, \frac{\tau^2\sigma^2/n}{\tau^2 + \sigma^2/n}\right).$$

Bayes estimate of θ is the posterior mean:

$$E(\theta|\mathbf{x}) = \frac{\tau^2}{\tau^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\mu,$$

Posterior variance:

$$Var(\theta|\mathbf{x}) = \frac{\tau^2\sigma^2/n}{\tau^2 + \sigma^2/n}.$$

$100(1-\alpha)\%$ HPD credible interval for θ is:

$$E(\theta|\mathbf{x}) \pm z_{1-\alpha/2}\,\text{s.d.}(\theta|\mathbf{x})$$

$$\frac{\tau^2}{\tau^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\mu \pm z_{1-\alpha/2}\sqrt{\frac{\tau^2\sigma^2/n}{\tau^2 + \sigma^2/n}}$$

Example. $X_1, \ldots, X_n$ i.i.d. Poisson(λ). $\lambda \sim \text{Exp}(a)$. Then

$$f(x_1, \ldots, x_n|\lambda) = \exp(-n\lambda)\lambda^{\sum_{i=1}^n x_i}/(\prod_{i=1}^n x_i!), x_i = 0, 1, 2, \ldots$$

$\pi(\lambda) = a\exp(-a\lambda)$, $\lambda > 0$, $a > 0$, so

$$\pi(\lambda|\mathbf{x}) = \frac{a\exp(-a\lambda)\exp(-n\lambda)\lambda^{\sum_{i=1}^n x_i}}{(\prod_{i=1}^n x_i!)m(\mathbf{x})} \propto \exp\left(-\lambda(n+a)\right)\lambda^{\sum_{i=1}^n x_i}, \lambda > 0.$$

Therefore $\lambda|\mathbf{x} \sim \Gamma(\sum_{i=1}^n x_i + 1, n + a)$ and hence

$$\begin{aligned} E(\lambda|\mathbf{x}) &= \frac{\sum_{i=1}^n x_i + 1}{n+a} = \frac{n}{n+a}\frac{\sum_{i=1}^n x_i}{n} + \frac{a}{n+a}\frac{1}{a}, \\ Var(\lambda|\mathbf{x}) &= \frac{\sum_{i=1}^n x_i + 1}{(n+a)^2}, \\ s.d.(\lambda|\mathbf{x}) &= \frac{\sqrt{\sum_{i=1}^n x_i + 1}}{n+a}. \end{aligned}$$

Example. $X_1, X_2, \cdots, X_n$ i.i.d. $(N(\theta, \sigma^2))$, σ^2 is known. $\theta \sim N(\mu, \tau^2)$. Then as shown previously,

$$\theta|\mathbf{X} = \mathbf{x} \sim N\left(\frac{\tau^2}{\tau^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\mu, \frac{\tau^2\sigma^2/n}{\tau^2 + \sigma^2/n}\right).$$

Therefore, the Bayes estimate of θ is the posterior mean

$$E(\theta|\boldsymbol{x}) = \frac{\tau^2}{\tau^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\mu,$$

and posterior variance

$$Var(\theta|\mathbf{x}) = \frac{\tau^2\sigma^2/n}{\tau^2 + \sigma^2/n}.$$

i.e., in the light of the data, θ shifts from prior guess μ towards a weighted average of the prior guess about θ and $\bar{x}$, while the variability reduces from σ^2 to $\frac{\sigma^2}{n}(\frac{\tau^2}{\tau^2+\sigma^2/n})$. Consider the role of τ^2 and n: If the prior information is small, implying large τ^2 or there is lot of data, i.e., n is large, the posterior mean is close to the MLE $\bar{x}$. Similarly, the posterior variance will be close to $\frac{\sigma^2}{n}$ in such a case. This can also be seen from the fact that then the posterior distribution is close to $N(\bar{x}, \frac{\sigma^2}{n})$, which is what one gets from the likelihood.

This phenomenon of the likelihood dominating any reasonable prior as the sample size grows simply says that as data accumulates, prior information becomes unimportant. As expected, prior information is especially useful when the sample size is small.

What happens to the posterior computations when there are more parameters?

Example. Suppose the data consist of i.i.d. observations $X_1, X_2, \ldots, X_n$ from a normal $N(\theta, \sigma^2)$ distribution where both θ and σ^2 are unknown. Suppose we are only interested in inferences for θ. Even then we need a joint prior on both the parameters. Consider the prior density $\pi(\theta, \sigma) = 1/\sigma$. This improper prior is recommended by Jeffreys. Then we have,

$$f(\mathbf{x}|\theta,\sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\theta-\bar{x})^2\right\}\right],$$

so that (letting $S^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2$),

$$\begin{aligned}\pi(\theta,\sigma|\mathbf{x}) &\propto f(\mathbf{x}|\theta,\sigma^2)\frac{1}{\sigma}\\ &= \text{constant}\ \sigma^{-(n+1)}\exp\left[-\frac{1}{2\sigma^2}\left\{n(\theta-\bar{x})^2+S^2\right\}\right].\end{aligned}$$

Therefore, with a transformation $v = \sigma^{-2}$, so that $dv = -2\sigma^{-3}\,d\sigma$, or $dv/v = -2d\sigma/\sigma$, and $s^2 = S^2/(n-1)$ we get

$$\begin{aligned}\pi(\theta\mathbf{x}) &= \int_0^\infty \pi(\theta,\sigma|\mathbf{x})\,d\sigma\\ &= \text{constant}\int_0^\infty \sigma^{-(n+1)}\exp\left[-\frac{1}{2\sigma^2}\left\{n(\theta-\bar{x})^2+S^2\right\}\right]d\sigma\\ &= \text{constant}\int_0^\infty \sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\left\{n(\theta-\bar{x})^2+S^2\right\}\right]\frac{d\sigma}{\sigma}\\ &= \text{constant}\int_0^\infty v^{n/2}\exp\left[-\frac{v}{2}\left\{n(\theta-\bar{x})^2+S^2\right\}\right]\frac{dv}{v}\\ &= \text{constant}\int_0^\infty \exp\left[-\frac{v}{2}\left\{n(\theta-\bar{x})^2+S^2\right\}\right]v^{n/2-1}\,dv\\ &= \text{constant}\ \left\{S^2+n(\theta-\bar{x})^2\right]^{-n/2}\\ &= \text{constant}\ (S^2)^{-n/2}\left\{1+\frac{n}{S^2}(\theta-\bar{x})^2\right]^{-n/2}\\ &\propto \left\{1+\frac{1}{n-1}\left(\frac{\sqrt{n}(\theta-\bar{x})}{s}\right)^2\right]^{-(n-1+1)/2},\end{aligned}$$

which is the density of Student's t with $n-1$ d.f. i.e.,

$$\frac{\sqrt{n}(\theta-\bar{x})}{s}|\mathbf{x} \sim t_{n-1}.$$

Therefore, the Bayes estimate for θ is $E(\theta|\mathbf{x}) = \bar{x}$ under the Jeffreys' prior.

Credible Intervals

Bayesian interval estimates for θ are similar to confidence intervals of classical inference. They are called credible intervals or sets.

Definition For $0 < \alpha < 1$, a $100(1-\alpha)\%$ credible set for θ is a subset $C \subset \Theta$ such that

$$P\{C|X = x\} = 1 - \alpha.$$

Usually C is taken to be an interval. Let θ be a continuous random variable, $\theta^{(1)}, \theta^{(2)}$ be $100\alpha_1\%$ and $100(1-\alpha_2)\%$ quantiles with $\alpha_1 + \alpha_2 = \alpha$. Let $C = [\theta^{(1)}, \theta^{(2)}]$. Then $P(C|X = x) = 1 - \alpha$. Usually equal tailed intervals are chosen so $\alpha_1 = \alpha_2 = \alpha/2$.

If θ is discrete, usually it would be difficult to find an interval with exact posterior probability $1-\alpha$. There the condition is relaxed to

$$P(C|X = x) \geq 1 - \alpha$$

with the inequality being as close to an equality as possible. In general, one may use a conservative inequality like this in the continuous case also if exact posterior probability $1-\alpha$ is difficult to attain.

Whereas the (frequentist) confidence statements do not apply to whether a given interval for a given x covers the "true" θ, this is not the case with credible intervals. The credibility $1-\alpha$ of a credible set does answer a layman's question on whether the given set covers the "true" θ with probability $1-\alpha$. This is because in the Bayesian approach, "true" θ is a random variable with a data dependent probability distribution, namely, the posterior distribution.

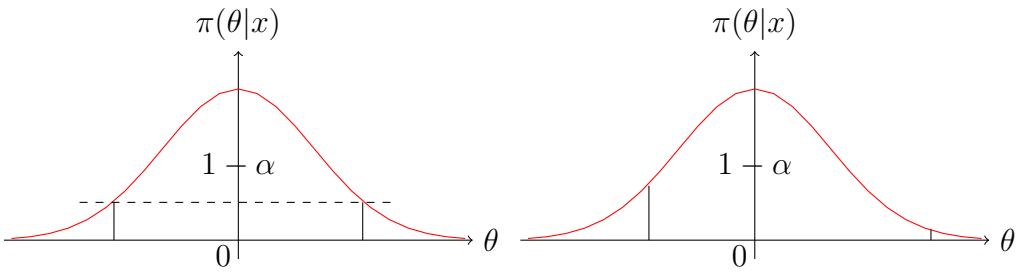
The equal tailed credible interval need not have the smallest size, namely, length or area or volume whichever is appropriate. For that one needs an HPD (Highest Posterior Density) interval.

Definition Suppose the posterior density for θ is unimodal. Then the HPD interval for θ is the interval

$$C = \{\theta : \pi(\theta|X = x) \geq k\},$$

where k is chosen such that

$$P(C|X = x) = 1 - \alpha.$$



HPD Credible Interval versus Other Credible Interval

Example. Consider a normal prior for mean of a normal population with known variance σ^2. The posterior is normal for wich the mean and variance have been derived earlier. The HPD interval is the same as the equal tailed interval centered at the posterior mean,

$$C = \text{ posterior mean } \pm z_{1-\alpha/2} \text{ posterior s.d.}$$

$X \sim N(\mu, \sigma^2)$. $I(\mu, \sigma^2) = ((I_{ij}(\mu, \sigma^2)))$, where

$$
\begin{aligned}
I_{11}(\mu, \sigma^2) &= E_{\mu,\sigma^2}\left[\frac{\partial}{\partial \mu} \log f(X|\mu, \sigma^2)\right]^2 \\
I_{22}(\mu, \sigma^2) &= E_{\mu,\sigma^2}\left[\frac{\partial}{\partial \sigma^2} \log f(X|\mu, \sigma^2)\right]^2 \\
I_{12}(\mu, \sigma^2) &= E_{\mu,\sigma^2}\left[\frac{\partial}{\partial \mu} \log f(X|\mu, \sigma^2) \frac{\partial}{\partial \sigma^2} \log f(X|\mu, \sigma^2)\right].
\end{aligned}
$$

$$
\begin{aligned}
\log f(x|\mu, \sigma^2) &= -\frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2}(x-\mu)^2, \\
\frac{\partial}{\partial \mu} \log f(x|\mu, \sigma^2) &= -\frac{1}{2\sigma^2} 2(x-\mu)(-1), \\
\frac{\partial}{\partial \sigma^2} \log f(x|\mu, \sigma^2) &= -\frac{1}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2.
\end{aligned}
$$

$$
\begin{aligned}
I_{11}(\mu,\sigma^2) &= E_{\mu,\sigma^2}\left[\frac{(X-\mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^2},\\
I_{22}(\mu,\sigma^2) &= \frac{1}{4\sigma^8}E_{\mu,\sigma^2}\left[(X-\mu)^2-\sigma^2\right]^2 = \frac{2\sigma^4}{4\sigma^8} = \frac{1}{2\sigma^4}.\\
I_{12}(\mu,\sigma^2) &= \frac{1}{2}E_{\mu,\sigma^2}\left[\left(\frac{X-\mu}{\sigma^2}\right)\left(\frac{(X-\mu)^2-\sigma^2}{\sigma^4}\right)\right] = 0.
\end{aligned}
$$

$$
I(\mu,\sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.
$$

$$
|I(\mu,\sigma^2)| \propto (\sigma^2)^{-3}.
$$

Jeffreys (formal) prior is

$$\pi(\mu, \sigma^2)\, d\mu\, d\sigma^2 = |I(\mu, \sigma^2)|^{1/2}\, d\mu\, d\sigma^2 \propto (\sigma^2)^{-3/2}\, d\mu\, d\sigma^2.$$

Why is this not the same as the left invariant Haar measure on the affine group, which is

$$\sigma^{-2}?$$

Since $d\sigma^2 = 2\sigma\, d\sigma$,

$$\pi(\mu, \sigma)\, d\mu\, d\sigma \propto (\sigma^2)^{-3/2}\, d\mu\, \sigma\, d\sigma = \sigma^{-2}\, d\mu\, d\sigma.$$

$$\frac{P^{\pi}(\Theta_0|x)}{P^{\pi}(\Theta_1|x)} = \frac{P^{\pi}(\Theta_0|x)}{1 - P^{\pi}(\Theta_0|x)} = \frac{\pi_0}{1-\pi_0} \times \text{BF}_{01}(x).$$

If $\Theta_0 = \{\theta_0\}$, then

$$\begin{aligned}
\frac{\pi(\theta_0|x)}{1-\pi(\theta_0|x)} &= \frac{\pi_0}{1-\pi_0} \times \text{BF}_{01}(x) \\
\frac{1-\pi(\theta_0|x)}{\pi(\theta_0|x)} &= \frac{1-\pi_0}{\pi_0} \times \text{BF}_{01}^{-1}(x) \\
\frac{1}{\pi(\theta_0|x)} - 1 &= \frac{1-\pi_0}{\pi_0} \times \text{BF}_{01}^{-1}(x) \\
\frac{1}{\pi(\theta_0|x)} &= 1 + \frac{1-\pi_0}{\pi_0} \times \text{BF}_{01}^{-1}(x)
\end{aligned}$$

Example. Now consider the unknown variance case. Then as discussed previously, with the Jeffreys' prior, we have

$$\frac{\sqrt{n}(\theta - \bar{x})}{s}|\mathbf{x} \sim t_{n-1}.$$

Then, since

$$P(|\frac{\sqrt{n}(\theta - \bar{x})}{s}| \leq t_{n-1}(1-\alpha/2)|\mathbf{x}) = 1-\alpha,$$

for $n \geq 2$, the HPD 100(1-α)% credible interval for θ is

$$\bar{x} \pm t_{n-1}(1-\alpha/2)\frac{s}{\sqrt{n}}.$$

Credible intervals are very easy to calculate unlike confidence intervals, the construction of which requires pivotal quantities or inversion of a family of tests. Consider the following example.

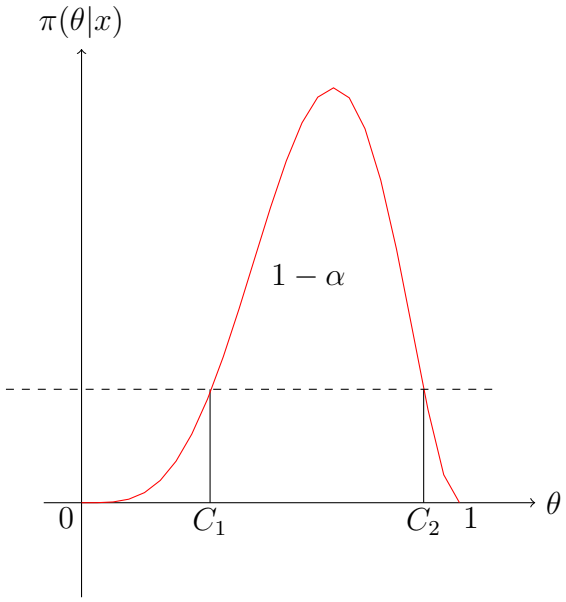
Example. $X|\theta \sim$ Binomial(n,θ) and $\theta \sim$ Beta(a,b). Then $\theta|X = x \sim$ Beta$(a+x, b+n-x)$. Therefore, the 100(1-α)% HPD credible set is (C_1, C_2) where C_1 and C_2 satisfy

$$\begin{aligned} 1-\alpha &= \int_{C_1}^{C_2} \pi(\theta|x)\, d\theta \\ &= \int_{C_1}^{C_2} \frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)}\theta^{a+x-1}(1-\theta)^{b+n-x-1}\, d\theta, \text{ and} \\ \pi(C_1|x) &= \pi(C_2|x), \text{ or} \end{aligned}$$

$$C_1^{a+x-1}(1-C_1)^{b+n-x-1} = C_2^{a+x-1}(1-C_2)^{b+n-x-1}$$

.

Solve for C_1 and C_2 numerically.



Example. $X_1, X_2, \ldots, X_n$ i.i.d $N(\theta, \sigma^2)$, σ^2 known. $\pi(\theta) \equiv C$. Then, from previous discussion,

$$\theta|x_1, x_2, \ldots, x_n \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right).$$

Therefore, the 100(1-α)% HPD credible set for θ is

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},$$

which is the same as the corresponding confidence interval. Then what is the difference between the Bayesian and Frequentist intervals? It is in the interpretation.

When viewed as a credible set, $\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ has (posterior) probability $1-\alpha$ of containing θ. But when it is viewed as a confidence interval, this fixed set has no such meaning. The random interval $\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ has probability $1-\alpha$ of containing θ. In otherwords, if the procedure is employed over and over again, then the resulting intervals have long-run relative frequency of $1-\alpha$ of capturing θ inside. If $\alpha = 0.05$, the random interval has 19 out of 20 chance of containing θ, so we can have the confidence that the confidence interval from any data set has a good chance of capturing θ.

Prediction of a Future Observation

We have already done this informally earlier. Suppose the data are $x_1, \cdots, x_n$, where $X_1, \ldots, X_n$ are i.i.d. with density $f(x|\theta)$, e.g., $N(\mu, \sigma^2)$ with σ^2 known.

We want to predict the unobserved X_{n+1} or set up a predictive credible interval for X_{n+1}.

Prediction by a single number $t(x_1, \cdots, x_n)$ based on $x_1, \cdots, x_n$ with squared error loss amounts to considering prediction loss

$$\begin{aligned} E\left\{(X_{n+1}-t)^2|\boldsymbol{x}\right\} &= E\left[\left\{(X_{n+1}-E(X_{n+1}|\boldsymbol{x}))-(t-E(X_{n+1}|\boldsymbol{x}))\right\}^2|\boldsymbol{x}\right] \\ &= E\left\{(X_{n+1}-E(X_{n+1}|\boldsymbol{x}))^2|\boldsymbol{x}\right\}+(t-E(X_{n+1}|\boldsymbol{x}))^2 \end{aligned}$$

which is minimum at

$$t = E(X_{n+1}|\boldsymbol{x}).$$

To calculate the predictor we need to calculate the predictive distribution

$$\begin{aligned} \pi(x_{n+1}|\boldsymbol{x}) &= \int_\Theta \pi(x_{n+1}|\boldsymbol{x},\theta)\pi(\theta|\boldsymbol{x})\,d\theta \\ &= \int_\Theta f(x_{n+1}|\theta)\pi(\theta|\boldsymbol{x})\,d\theta. \end{aligned}$$

Let $\mu(\theta) = \int_{-\infty}^{\infty} x f(x|\theta)\,dx$. It can be shown that

$$E(X_{n+1}|\boldsymbol{x}) = E(\mu(\theta)|\boldsymbol{x}) = \int_\Theta \mu(\theta)\pi(\theta|\boldsymbol{x})\,d\theta$$

and hence for the normal problem the predictor is $\int_{-\infty}^{\infty} \mu\pi(\mu|\boldsymbol{x})\,d\mu$ = posterior mean of μ.

Similarly in the Binomial Example, the predictive probability that the next ball is red is

$$E(X_{n+1}|\boldsymbol{x}) = E(p|\boldsymbol{x}) = \frac{\alpha+r}{\alpha+\beta+n}$$

where $r = \sum_1^n x_i$.

A predictive credible interval for X_{n+1} is (c,d) where c and d are $100\alpha_1\%$ and $100(1-\alpha_2)\%$ quantiles of the predictive distribution of X_{n+1} given $\boldsymbol{x}$. Usually, one takes $\alpha_1 = \alpha_2 = \alpha/2$ as for credible intervals.

Testing of hypotheses: Model choice/criticism

$X \sim P_\theta$, $\theta \in \Theta$, with density or mass function $f(x|\theta)$. We want to test $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, $\Theta_0 \cup \Theta_1 = \Theta$. In principle this is just another Bayesian inference problem. Simply obtain $\pi(\theta|x)$ and compute

$$P(\Theta_0|x) = \int_{\Theta_0} \pi(\theta|x)\,d\theta \text{ and } P(\Theta_1|x) = \int_{\Theta_1} \pi(\theta|x)\,d\theta = 1 - P(\Theta_0|x).$$

If $P(\Theta_0|x) > 1/2$ (or a suitable threshold), or the posterior odds ratio (of H_0 relative to H_1), $P(\Theta_0|x)/P(\Theta_1|x) > 1$, accept H_0.

Example. Consider a blood test conducted for determining the sugar level of a person with diabetes two hours after he had his breakfast. It is of interest to see if his medication has controlled his blood sugar levels. Assume that the test result X is $N(\theta, 100)$, where θ is the true level. In the appropriate population (diabetic but under this treatment), θ is distributed according to a $N(100, 900)$. Then, marginally X is $N(100, 1000)$, and the posterior distribution of θ given $X = x$ is normal with
mean $= \frac{900}{1000}x + \frac{100}{1000}100 = 0.9x + 10$ and variance $= \frac{100\times 900}{1000} = 90$.
Suppose we want to test $H_0 : \theta \leq 130$ versus $H_1 : \theta > 130$. If the blood test shows a sugar level of 130, what can be concluded? Note that, given this test result, the true mean blood sugar level (θ) may be assumed to be $N(127, 90)$, which is the posterior distribution. Consequently, we obtain,

$$\begin{aligned} P(\theta \leq 130|X = 130) &= \Phi\left(\frac{130 - 127}{\sqrt{90}}\right) = \Phi(.316) = 0.624, \text{ and hence} \\ P(\theta > 130|X = 130) &= 0.376. \end{aligned}$$

Therefore the Posterior odds ratio of H_0 relative to H_1 is $0.624/0.376 = 1.66$.

There is a simple interpretation for the posterior odds ratio, $P(\Theta_0|x)/P(\Theta_1|x)$ mentioned above. Consider the simple versus simple testing: $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. We can assume $\Theta = \{\theta_0, \theta_1\}$. Let $\pi_0 = P^\pi(\theta = \theta_0) = 1 - P^\pi(\theta = \theta_1)$. Then

$$\begin{aligned} \pi(\theta|x) &= \frac{\pi(\theta)f(x|\theta)}{m(x)} = \frac{\pi(\theta)f(x|\theta)}{\pi_0 f(x|\theta_0) + (1-\pi_0)f(x|\theta_1)} \\ &= \left\{ \begin{array}{ll} \frac{\pi_0 f(x|\theta_0)}{m(x)} & \text{if } \theta = \theta_0; \\ \frac{(1-\pi_0)f(x|\theta_1)}{m(x)} & \text{if } \theta = \theta_1. \end{array} \right\} \end{aligned}$$

Therefore, the posterior odds ratio of H_1 relative to H_0 is

$$\frac{(1-\pi_0)f(x|\theta_1)}{\pi_0 f(x|\theta_0)} = \frac{f(x|\theta_1)}{f(x|\theta_0)},$$

if $\pi_0 = 1/2$. This is nothing but the likelihood ratio used in the classical MP test. However the Bayesian use of it is simply to use it for expressing evidence against H_0 directly, without having to look for a reference distribution.

Testing a Point Null Hypothesis The problem is to test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

Conceptually testing a point null is not a different problem, but there are complications. First of all, it is not possible to use a continuous prior density because any such prior will necessarily assign prior probability zero to the null hypothesis. Consequently, the posterior probability of the null hypothesis will also be zero. Intuitively, this is clear: if the null hypothesis is *a priori* impossible, it will remain so *a posteriori* also. Therefore, a prior probability of $\pi_0 > 0$ needs to be assigned to the point θ_0 and the remaining probability of $\pi_1 = 1 - \pi_0$ will be spread over $\{\theta \neq \theta_0\}$ using a density g_1. The complication now is that the prior π is of the form

$$\pi(\theta) = \pi_0 I\{\theta = \theta_0\} + (1-\pi_0)g_1(\theta)I\{\theta \neq \theta_0\}$$

and hence has both discrete and continuous parts. This complication appears whenever Θ_0 and Θ_1 have different dimensions when we test $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$. Therefore we shall discuss this more general problem below.

Let π_0 and $1 - \pi_0$ be the prior probabilities of Θ_0 and Θ_1. Let $g_i(\theta)$ be the prior p.d.f. of θ on Θ_i (conditional on H_i being true), so that

$$\int_{\Theta_i} g_i(\theta)\, d\theta = 1.$$

Thus the prior $\pi(\theta)$ is specified by

$$\pi(\theta) = \pi_0 g_0(\theta) I\{\theta \in \Theta_0\} + (1-\pi_0) g_1(\theta) I\{\theta \in \Theta_1\}.$$

We do not require any longer that Θ_0 and Θ_1 be of the same dimension. We can now proceed as before and compute posterior probabilities or posterior odds. To obtain these posterior quantities, note that the marginal density of X under the prior π can be expressed as

$$\begin{aligned} m_\pi(x) &= \int_\Theta f(x|\theta)\pi(\theta)\, d\theta \\ &= \pi_0 \int_{\Theta_0} f(x|\theta) g_0(\theta)\, d\theta + (1-\pi_0) \int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta, \end{aligned}$$

and hence the posterior density of θ given the data $X = x$ as

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m_\pi(x)} = \begin{cases} \pi_0 f(x|\theta) g_0(\theta)/m_\pi(x) & \text{if } \theta \in \Theta_0; \\ (1-\pi_0) f(x|\theta) g_1(\theta)/m_\pi(x) & \text{if } \theta \in \Theta_1. \end{cases}$$

It follows then that

$$\begin{aligned} P^\pi(H_0|x) &= P^\pi(\Theta_0|x) = \frac{\pi_0}{m_\pi(x)} \int_{\Theta_0} f(x|\theta) g_0(\theta)\, d\theta \\ &= \frac{\pi_0 \int_{\Theta_0} f(x|\theta) g_0(\theta)\, d\theta}{\pi_0 \int_{\Theta_0} f(x|\theta) g_0(\theta)\, d\theta + (1-\pi_0) \int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta} \quad \text{and} \\ P^\pi(H_1|x) &= P^\pi(\Theta_1|x) = \frac{(1-\pi_0)}{m_\pi(x)} \int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta \\ &= \frac{(1-\pi_0) \int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta}{\pi_0 \int_{\Theta_0} f(x|\theta) g_0(\theta)\, d\theta + (1-\pi_0) \int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta}. \end{aligned}$$

Therefore, the posterior odds ratio of H_0 to H_1 is

$$\begin{aligned} \frac{P^\pi(\Theta_0|x)}{P^\pi(\Theta_1|x)} &= \frac{\pi_0 \int_{\Theta_0} f(x|\theta) g_0(\theta)\, d\theta}{(1-\pi_0) \int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta} \\ &= \frac{\pi_0}{1-\pi_0} \frac{\int_{\Theta_0} f(x|\theta) g_0(\theta)\, d\theta}{\int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta} \\ &= \frac{\pi_0}{1-\pi_0} \times \mathrm{BF}_{01}(x), \end{aligned}$$

where

$$\begin{aligned} \mathrm{BF}_{01}(x) &= \frac{P^\pi(\Theta_0|x)}{P^\pi(\Theta_1|x)} \Big/ \frac{P^\pi(\Theta_0)}{P^\pi(\Theta_1)} \\ &= \frac{\text{Posterior odds ratio}}{\text{Prior odds ratio}}. \end{aligned}$$

Thus, one may also report the *Bayes factor*, which does not depend on π_0. Note that the Bayes factor may be defined without reference to the prior odds ratio also:

$$\mathrm{BF}_{01} = \frac{\int_{\Theta_0} f(x|\theta) g_0(\theta)\, d\theta}{\int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta} = \frac{m_0(x)}{m_1(x)},$$

where $m_i(x)$ is the marginal or predictive distribution of X under H_i. Clearly, $\mathrm{BF}_{10} = 1/\mathrm{BF}_{01}$. Also, the posterior odds ratio of H_0 relative to H_1 is

$$\left(\frac{\pi_0}{1-\pi_0}\right) \mathrm{BF}_{01},$$

which reduces to BF_{01} if $\pi_0 = \frac{1}{2}$. Thus, BF_{01} is an important evidential measure that is free of π_0. The smaller the value of BF_{01}, the stronger the evidence against H_0. As noted previously, the Bayes factor is the likelihood ratio in the simple verus simple case, a weighted likelihood ratio in the general case.

Example. In the blood sugar example, $\pi_0 = P^{\pi}(\theta \leq 130) = \Phi(\frac{130-100}{30}) = \Phi(1)$, so the prior odds ratio is $\pi_0/(1-\pi_0) = \Phi(1)/(1-\Phi(1)) = .8413/.1587 = 5.3$, and thus the Bayes factor turns out to be BF_{01} = posterior odds ratio/prior odds ratio = $1.66/5.3 = .313$.

Consider an example on testing a point null hypothesis.

Example. Suppose $X \sim \mathrm{Binomial}(n, \theta)$ and we want to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, a problem similar to checking whether a given coin is biased based on n independent tosses (where θ_0 will be taken to be 0.5). Under the alternative hypothesis, suppose θ is distributed as $\mathrm{Beta}(\alpha, \beta)$. Then $m_1(x)$ is given by

$$m_1(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)},$$

so that

$$\begin{aligned} BF_{01}(x) &= \binom{n}{x}\theta_0^x(1-\theta_0)^{n-x} \Big/ \left(\binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}\right) \\ &= \theta_0^x(1-\theta_0)^{n-x} \Big/ \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}\right) \\ &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} \theta_0^x(1-\theta_0)^{n-x}. \end{aligned}$$

Hence, we obtain,

$$\begin{aligned}\pi(\theta_0|x) &= \left\{1+\frac{1-\pi_0}{\pi_0}\,BF_{01}^{-1}(x)\right\}^{-1} \\ &= \left\{1+\frac{1-\pi_0}{\pi_0}\,\frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}}{\theta_0^x(1-\theta_0)^{n-x}}\right\}^{-1}.\end{aligned}$$

Applications in statistical inference

References:
1. Bickel, D. and Doksum, K. *Mathematical Statistics*
2. Lehman, E. and Casella, G. *Theory of Point Estimation*

Let $X_1, X_2, \ldots$ be an i.i.d sequence such that $E(X) = \mu$. Then, since $\bar{X}_n \xrightarrow[n\to\infty]{P} \mu$, we have that $g(\bar{X}_n) \xrightarrow[n\to\infty]{P} g(\mu)$ for all continuous functions g.

Example. Let $X_1, X_2, \ldots$ be an i.i.d sequence such that $E(X) = \mu$ and $Var(X) = \sigma^2$. Then $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \xrightarrow[n\to\infty]{P} \sigma^2$.

Proof. Since $\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow[n\to\infty]{P} E(X^2) = \mu^2 + \sigma^2$ and $\bar{X}^2 \xrightarrow[n\to\infty]{P} (E(X))^2 = \mu^2$, we obtain, from Slusky,

$$\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 \xrightarrow[n\to\infty]{P} \{\mu^2 + \sigma^2\} - \mu^2 = \sigma^2.$$

Example. Let $X_1, X_2, \ldots$ be an i.i.d sequence such that $E(X) = \mu$ and $Var(X) = \sigma^2$. Then from CLT, we have,

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \xrightarrow[n\to\infty]{d} N(0,1).$$

From this we obtain that

$$\left(\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma}\right)^2 = \frac{n\left(\bar{X}_n - \mu\right)^2}{\sigma^2} \xrightarrow[n\to\infty]{d} \chi_1^2,$$

since $g(x) = x^2$ is continuous.

Large Sample Optimality

It is desirable to see that statistical procedures have optimality properties as more and more data become available – estimators should be close to the true quantities, their errors become small and so on.

Definition. Let $X_1, X_2, \ldots$ be i.i.d P_θ. An estimator $T_n(X_1, X_2, \ldots, X_n)$ of $q(\theta)$ is said to be consistent (strong) if

$$T_n(X_1, X_2, \ldots, X_n) \xrightarrow[n\to\infty]{P} q(\theta)$$

(strong if convergence is a.s.).

Result. Suppose $q(\theta) = g(\mu_1(\theta), \ldots, \mu_k(\theta))$ for $k \geq 1$, where $\mu_r(\theta) = E_\theta(X^r)$, $r = 1, 2, \ldots$ and g is continuous. Let the sample moments be $\hat{\mu}_r(\theta) = \frac{1}{n}\sum_{j=1}^{n} X_j^r$. Then the method of moments estimate $T_n = g(\hat{\mu}_1(\theta), \ldots, \hat{\mu}_k(\theta))$ is consistent.

This follows from the fact that

$$(\hat{\mu}_1(\theta), \ldots, \hat{\mu}_k(\theta)) \xrightarrow[n\to\infty]{P} (\mu_1(\theta), \ldots, \mu_k(\theta))$$

so that

$$g(\hat{\mu}_1(\theta), \ldots, \hat{\mu}_k(\theta)) \xrightarrow[n\to\infty]{P} g(\mu_1(\theta), \ldots, \mu_k(\theta)).$$

Note that unbiasedness does not imply consistency. For instance, consider i.i.d $X_1, \ldots, X_n$ with $E(X) = \mu$ and $Var(X) = \sigma^2 < \infty$. We know then that $T_n = \bar{X}$ is both consistent and unbiased, whereas $U_n = X_1$ is unbiased but not consistent.

To establish the consistency of a given estimator, LLN may not be useful in some situations.

Example. Let $X_1, \ldots, X_n$ be i.i.d $U(0, \theta)$. Then $E(X) = \theta/2$. A method of moments estimator is
$\hat{\theta}_1 = \frac{2}{n}\sum_{i=1}^{n} X_i$, which is consistent from WLLN. Note that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{P} E(X) = \frac{\theta}{2}.$$

The MLE, however, is different, and is
$\hat{\theta}_2 = X_{(n)}$, a function of the minimal sufficient statistic. Is this consistent? MLE is generally consistent, but regularity conditions apply, so it is easier to prove it directly rather than checking those conditions. One can establish the consistency of $\hat{\theta}_2$ by finding its mean and variance (using the distribution of o.s.; $X_{(n)}/\theta$ is Beta(1, n)) and applying the Chebychev's inequality. Alternatively, since $0 < X_{(n)} < \theta$,

$$\begin{aligned} P(|X_{(n)} - \theta| > \epsilon) &= P(X_{(n)} - \theta < -\epsilon) + P(X_{(n)} - \theta > \epsilon) \\ &= P(X_{(n)} < \theta - \epsilon) = (P(X < \theta - \epsilon))^n \\ &= \left(\frac{\theta - \epsilon}{\theta}\right)^n = \left(1 - \frac{\epsilon}{\theta}\right)^n \\ &\xrightarrow[n\to\infty]{} 0 \end{aligned}$$

for any fixed $\epsilon > 0$. MLE is not always consistent as the following example shows.

Example (Neyman-Scott problem). This problem involves estimating the precision of a measuring device by measuring a large number of different quantities. Suppose two independent measurements each of $\mu_1, \mu_2, \ldots...$ are made. In other words, let $\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_i \\ \mu_i \end{pmatrix}, \sigma^2 I_2\right)$, $i = 1, 2, \ldots$ be independent. Now there is no question of consistent estimators for μ_i since only two observations are available. σ^2 is important for calibration purposes. What is the MLE of σ^2? Note

$$f\big(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \cdots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}\big) = (2\pi)^{-n}(\sigma^2)^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(x_i-\mu_i)^2 + (y_i-\mu_i)^2\right]\right).$$

Fix σ^2. Since X_i and Y_i are i.i.d $N(\mu_i, \sigma^2)$,

$$\hat{\mu}_i = \frac{1}{2}(X_i + Y_i), i = 1, 2, \ldots \text{ independent of } \sigma^2 \ .$$

Therefore,

$$\begin{aligned} \max_{\{\mu_i\},\sigma^2} L\left(\{\mu_i\}, \sigma^2, (\mathbf{x}, \mathbf{y})\right) &= \max_{\sigma^2} L\left(\{\hat{\mu}_i\}, \sigma^2, (\mathbf{x}, \mathbf{y})\right) \\ &= \max_{\sigma^2} (\sigma^2)^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(x_i-\hat{\mu}_i)^2 + (y_i-\hat{\mu}_i)^2\right]\right). \end{aligned}$$

Thus

$$\begin{aligned} \hat{\sigma^2} &= \frac{1}{2n}\sum_{i=1}^{n}\left[(x_i-\hat{\mu}_i)^2 + (y_i-\hat{\mu}_i)^2\right] \\ &= \frac{1}{2n}\sum_{i=1}^{n} 2\left(x_i - \frac{1}{2}(x_i+y_i)\right)^2 = \frac{1}{4n}\sum_{i=1}^{n}(x_i-y_i)^2. \end{aligned}$$

Now note that $X_i - Y_i \sim N(0, 2\sigma^2)$ i.i.d., so

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-Y_i)^2 \xrightarrow[n\to\infty]{P} E(X-Y)^2 = 2\sigma^2, \text{ and hence}$$

$$\frac{1}{4n}\sum_{i=1}^{n}(X_i - Y_i)^2 \xrightarrow[n\to\infty]{P} \frac{\sigma^2}{2}.$$

This shows that the MLE of σ^2 is inconsistent here. Consistent estimators are easily available, however. For example, $\frac{1}{2n}\sum_{i=1}^{n}(X_i - Y_i)^2$ is a consistent estimator. The problem with the MLE here is that it considers the problem of estimating the infinite sequence of μ_i in addition to σ^2 in the limit as n grows.

Example. $X_1, \ldots, X_n$ i.i.d Bernoulli(p). Estimate p. Then $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}_n$ is MLE, method of moments as well as UMVUE. By WLLN, $\hat{p} = \bar{X}_n \xrightarrow[n\to\infty]{P} p$, thus showing that it is consistent. What if we want to estimate $q(p) = p(1-p)$? Then $\hat{q}(p) = \hat{p}(1-\hat{p})$ is MLE or method of moments estimator. Since $q(x) = x(1-x)$ is a continuous function and $\hat{p} \xrightarrow[n\to\infty]{P} p$, we have that $\hat{q}(p) = \hat{p}(1-\hat{p})$ is consistent for $q(p) = p(1-p)$. How good is this estimator?

How good is consistency as a measure of optimality of an estimator? We need a measure of accuracy. For large samples, we need a rate of convergence to the true parameter or parametric function. If we assume that our estimator is asymptotically unbiased (i.e. $E(\hat{\theta}-\theta) \underset{n\to\infty}{\longrightarrow} 0$) then we can use the asymptotic s.d. for this purpose. If we have an i.i.d sequence $X_1, X_2, \ldots$ with $E(X) = \mu$ and $Var(X) = \sigma^2$ then $\sqrt{n}(\bar{X}_n - \mu)$ is asymptotically normal. What about $g(\bar{X}_n)$ in such a situation for a smooth function g? We need the following result in this context.

Result. Suppose $\{a_n\} \uparrow \infty$ as $n \to \infty$, b fixed and

$$a_n(X_n - b) \xrightarrow[n\to\infty]{d} X.$$

Let g be a continuous function which is differentiable, and let g' be continuous and $g'(b) \neq 0$. Then

$$a_n \left(g(X_n) - g(b)\right) \xrightarrow[n\to\infty]{d} g'(b)X.$$

Proof. Note that

$$X_n - b = \frac{1}{a_n}\left[a_n(X_n - b)\right] \xrightarrow[n\to\infty]{d} 0 \times X = 0.$$

Therefore $X_n \xrightarrow[n\to\infty]{P} b$. Now, $a_n \left(g(X_n) - g(b)\right) = a_n \left(g'(X_n^*)(X_n - b)\right)$ where X_n^* lies between X_n and b. Therefore

$$|X_n^* - b| \leq |X_n - b| \xrightarrow[n\to\infty]{P} 0.$$

Therefore, $X_n^* \xrightarrow[n\to\infty]{P} b$ and hence $g'(X_n^*) \xrightarrow[n\to\infty]{P} g'(b)$. It follows then that

$$g'(X_n^*)a_n \left(X_n - b\right) \xrightarrow[n\to\infty]{d} g'(b)X.$$

Note, however, that if $g'(b) = 0$, then $a_n \left(g(X_n) - g(b)\right) \xrightarrow[n\to\infty]{P} 0$.

Result. Suppose we have an i.i.d sequence $X_1, X_2, \ldots$ with $E(X) = \mu$ and $Var(X) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Let h be differentiable, h' be continuous and $h'(\mu) \neq 0$. Then

$$\sqrt{n}\left(h(\bar{X}_n) - h(\mu)\right) \xrightarrow[n\to\infty]{d} N\left(0, (h'(\mu))^2\sigma^2\right).$$

Proof. From CLT, $\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow[n\to\infty]{d} N\left(0, \sigma^2\right)$. Therefore, from the previous result,

$$\sqrt{n}\left(h(\bar{X}_n) - h(\mu)\right) \xrightarrow[n\to\infty]{d} h'(\mu)N\left(0, \sigma^2\right) = N\left(0, (h'(\mu))^2\sigma^2\right).$$

Example. $X_1, \ldots, X_n$ i.i.d Bernoulli(p). Let $S_n = \sum_{i=1}^n X_i$. Then $\hat{p}_n = S_n/n = \bar{X}_n$ satisfies $\sqrt{n}(\hat{p}_n - p) = \sqrt{n}(\bar{X}_n - p) \xrightarrow[n\to\infty]{d} N(0, p(1-p))$. Now consider estimating $q(p) = p(1-p)$ with $T_n = \hat{p}_n(1-\hat{p}_n)$. We have seen earlier that it is consistent. What can be said about its asymptotic distribution? Consider $h(x) = x(1-x)$ which is differentiable with $h'(x) = 1-2x$. $h'(p) \neq 0$ if $p \neq 1/2$. Therefore, for $p \neq 1/2$,

$$\begin{aligned}\sqrt{n}\left(T_n - p(1-p)\right) = \sqrt{n}\left(h(\hat{p}_n) - h(p)\right) &\xrightarrow[n\to\infty]{d} N\left(0, (h'(p))^2 p(1-p)\right) \\ &= N(0, p(1-p)(1-2p)^2).\end{aligned}$$

What happens when $p = 1/2$? Recall how we proved the result:

$$a_n\left(g(X_n) - g(b)\right) \xrightarrow[n\to\infty]{d} g'(b)X$$

if $a_n(X_n - b) \xrightarrow[n\to\infty]{d} X$, g is differentiable, g' continuous and $g'(b) \neq 0$? We used Taylor series expansion: $g(X_n) = g(b) + g'(X_n^*)(X_n - b)$. If $g'(b) = 0$, we need a further term:

$$\begin{aligned} g(X_n) &= g(b) + g'(b)(X_n - b) + \frac{1}{2}g''(X_n^*)(X_n - b)^2 \\ &= g(b) + \frac{1}{2}g''(X_n^*)(X_n - b)^2, \end{aligned}$$

so that $g(X_n) - g(b) = \frac{1}{2}g''(X_n^*)(X_n - b)^2$. Assume that g'' is continuous at b and $g''(b) \neq 0$. Then $g''(X_n^*) \xrightarrow[n\to\infty]{P} g''(b)$ and $\left(a_n(X_n - b)\right)^2 \xrightarrow[n\to\infty]{d} X^2$. Therefore,

$$\begin{aligned} a_n^2\left(g(X_n) - g(b)\right) &= \frac{1}{2}g''(X_n^*)\left\{a_n(X_n - b)\right\}^2 \\ &\xrightarrow[n\to\infty]{d} \frac{1}{2}g''(b)X^2. \end{aligned}$$

Now consider the asymptotic distribution of $T_n = \hat{p}_n(1-\hat{p}_n)$ in the Bernoulli(p) example. Here $h(x) = x(1-x)$, $h'(x) = 1 - 2x$ and $h''(x) = -2$. Also $\sqrt{n}(\hat{p}_n - \frac{1}{2}) \xrightarrow[n\to\infty]{d} N(0, \frac{1}{4})$. Therefore

$$n(\hat{p}_n(1-\hat{p}_n) - \frac{1}{4}) \xrightarrow[n\to\infty]{d} \frac{1}{2}(-2)\frac{1}{4}\chi_1^2 = -\frac{1}{4}\chi_1^2.$$

Example. Let $X_1, X_2, \ldots$ be an i.i.d sequence such that $E(X) = \mu$ and $Var(X) = \sigma^2$. Let $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Then

$$\left(\sqrt{n}(\bar{X} - \mu), s^2\right) \xrightarrow[n\to\infty]{d} \left(N(0, \sigma^2), \sigma^2\right).$$

Therefore, by Slutsky,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{s} \xrightarrow[n\to\infty]{d} N(0, 1).$$

Example. Let $X_1, X_2, \ldots$ be i.i.d $N(\mu, \sigma^2)$. Then

$$T_n = \frac{\sqrt{n}(\bar{X} - \mu)}{s} \sim t_{n-1}.$$

Thus note that as $n \to \infty$, $t_{n-1} \xrightarrow{d} N(0, 1)$ from the previous result. This can also be seen directly since the numerator is always $N(0, \sigma^2)$ whereas s in the denominator converges to σ in probability.

Asymptotic Normality.

Large sample normal approximation for estimators is desirable:
(i) to obtain estimation error;
(ii) to be able to compare efficiencies; and
(iii) for large sample tests.

Usually reasonable estimators $\hat{\theta}_n$ converge at the rate of $O_P(\frac{1}{\sqrt{n}})$. i.e., $\hat{\theta}_n \xrightarrow[n\to\infty]{P}$ θ and $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{P}$ to some distribution.

Definition. Suppose $T_n(X_1, \ldots, X_n)$ is an estimator of $q(\theta)$. Then T_n is said to be asymptotically normal if

$$\sqrt{n}\left(T_n(X_1, \ldots, X_n) - q(\theta)\right) \xrightarrow[n\to\infty]{d} N(0, \sigma^2(\theta)).$$

Example. Let $X_1, X_2, \ldots$ be i.i.d such that $E(X) = \mu$ and $Var(X) = \sigma^2$. Consider $T_n(X_1, \ldots, X_n) = \bar{X}$. Then by CLT $\sqrt{n}(T_n - \mu) \xrightarrow[n\to\infty]{d} N(0, \sigma^2)$. We therefore say that $\bar{X}$ is asymptotically normal in this case.

Example. Let $X_1, X_2, \ldots$ be i.i.d Exp(θ). Then $E(X) = \frac{1}{\theta}$ and $Var(X) = \frac{1}{\theta^2}$. Consider $\hat{\theta} = \frac{1}{\bar{X}}$. Since, by WLLN, $\bar{X} \xrightarrow[n\to\infty]{P} \frac{1}{\theta}$, $\hat{\theta} = \frac{1}{\bar{X}} \xrightarrow[n\to\infty]{P} \theta$, hence it is

consistent. Note that we have made use of the continuity of $h(x) = 1/x$ for $x > 0$. $h'(x) = -1/x^2$ is also continuous for $x > 0$. We have, by CLT,

$$\sqrt{n}\left(\bar{X} - \frac{1}{\theta}\right) \xrightarrow[n\to\infty]{d} N(0, \frac{1}{\theta^2}).$$

Therefore,

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow[n\to\infty]{d} h'(\frac{1}{\theta})N(0, \frac{1}{\theta^2}) = N\left(0, (\theta^2)^2\frac{1}{\theta^2} = \theta^2\right).$$

Asymptotic Relative Efficiency (ARE)

Consider two estimators which are asymptotically unbiased. If one of them has a smaller variance than the other, then the former is more precise, more efficient.

Suppose $T_n^{(1)}$ and $T_n^{(2)}$ are two estimators of $q(\theta)$ such that

$$\sqrt{n}\left(T_n^{(1)} - q(\theta)\right) \xrightarrow[n\to\infty]{d} N(0, \sigma_1^2(\theta)) \text{ and } \sqrt{n}\left(T_n^{(2)} - q(\theta)\right) \xrightarrow[n\to\infty]{d} N(0, \sigma_2^2(\theta)).$$

Then the asymptotic relative efficiency of $T_n^{(1)}$ w.r.t. $T_n^{(2)}$ is defined to be

$$\mathrm{e}\left(\theta, T_n^{(1)}, T_n^{(2)}\right) = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}.$$

Notice $T_n^{(1)}$ is better if $\sigma_2^2 > \sigma_1^2$.

Example. $X_1, \ldots, X_n$ i.i.d Bernoulli(p). Let $S_n = \sum_{i=1}^n X_i$. Then MLE of p is $T_n^{(1)} = \frac{S_n}{n}$. Consider the Bayes estimate of p under the Beta(a, b) prior for p, $a > 0$, $b > 0$. Since

$$p|X_1, \ldots, X_n \sim \text{Beta}(S_n + a, n - S_n + b),$$

we obtain the Bayes estimate to be

$$T_n^{(2)} = E(p|X_1, \ldots, X_n) = \frac{S_n + a}{S_n + a + n - S_n + b} = \frac{S_n + a}{n + a + b}.$$

We know already that

$$\sqrt{n}(T_n^{(1)} - p) \xrightarrow[n\to\infty]{d} N(0, p(1-p)).$$

Since

$$\begin{aligned} T_n^{(2)} &= \frac{S_n + a}{n + a + b} = \frac{S_n}{n}\frac{n}{n + a + b} + \frac{a}{n + a + b} \\ &= \frac{S_n}{n}\left(1 - \frac{a + b}{n + a + b}\right) + \frac{a}{n + a + b} \\ &= \frac{S_n}{n} - \frac{(a + b)S_n}{n(n + a + b)} + \frac{a}{n + a + b}, \end{aligned}$$

$$\sqrt{n}(T_n^{(2)} - p) = \sqrt{n}(T_n^{(1)} - p) - \frac{(a + b)S_n}{\sqrt{n}(n + a + b)} + \frac{\sqrt{n}a}{n + a + b}.$$

Now, note that, since a and b are fixed, and $\frac{S_n}{n} \xrightarrow[n\to\infty]{P} p$,

$$\frac{\sqrt{n}a}{n+a+b} \xrightarrow[n\to\infty]{} 0, \text{ and } \frac{(a+b)S_n}{\sqrt{n}(n+a+b)} \xrightarrow[n\to\infty]{P} 0.$$

Therefore $T_n^{(2)}$ has the same asymptotic distribution as $T_n^{(1)}$. i.e.,

$$\sqrt{n}(T_n^{(2)} - p) \xrightarrow[n\to\infty]{d} N(0, p(1-p)).$$

Thus the two estimators have the same asymptotic relative efficiency, or

$$\mathrm{e}(p, T_n^{(1)}, T_n^{(2)}) = \frac{p(1-p)}{p(1-p)} = 1.$$

Example. Let $X_1, X_2, \ldots$ be i.i.d $N(0, \sigma^2)$. Consider the following two estimators for σ:

$$\hat{\sigma}_1 = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}, \qquad \hat{\sigma}_2 = \sqrt{\frac{\pi}{2}}\frac{1}{n}\sum_{i=1}^{n} |X_i|.$$

Since $X_i^2 \sim \sigma^2\chi_1^2$, $E(X_i^2) = \sigma^2$, $Var(X_i^2) = 2\sigma^4$, by CLT,

$$\sqrt{n}\left(\hat{\sigma}_1^2 - \sigma^2\right) \xrightarrow[n\to\infty]{d} N(0, 2\sigma^4).$$

Taking $h(x) = \sqrt{x}$ for $x > 0$, we have $h'(x) = \frac{1}{2\sqrt{x}}$, and so

$$\sqrt{n}\left(\hat{\sigma}_1 - \sigma\right) \xrightarrow[n\to\infty]{d} N\left(0, \left(\frac{1}{2\sigma}\right)^2 2\sigma^4\right) = N\left(0, \frac{\sigma^2}{2}\right).$$

Since $Z_i = X_i/\sigma \sim N(0,1)$, and

$$\begin{aligned} E(|Z_i|) &= \int_{-\infty}^{\infty} |z| \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\, dz \\ &= 2\int_0^{\infty} z \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\, dz \\ &= \sqrt{\frac{2}{\pi}} \int_0^{\infty} \exp(-u)\, du = \sqrt{\frac{2}{\pi}}, \end{aligned}$$

we get $E(|X_i|) = \sqrt{\frac{2}{\pi}}\sigma$ and $Var(|X_i|) = E(X_i^2) - (E(|X_i|))^2 = \sigma^2 - \frac{2}{\pi}\sigma^2 = \left(1 - \frac{2}{\pi}\right)\sigma^2$. Therefore, $E(\sqrt{\frac{\pi}{2}}|X_i|) = \sigma$ and $Var(\sqrt{\frac{\pi}{2}}|X_i|) = \frac{\pi}{2}(1 - \frac{2}{\pi})\sigma^2 = (\frac{\pi}{2} - 1)\sigma^2$. Hence

$$\sqrt{n}\left(\hat{\sigma}_2 - \sigma\right) \xrightarrow[n\to\infty]{d} N\left(0, \left(\frac{\pi}{2} - 1\right)\sigma^2\right).$$

Thus we have the ARE:

$$\mathrm{e}\left(\sigma^2, \hat{\sigma}_1, \hat{\sigma}_2\right) = \frac{\left(\frac{\pi}{2} - 1\right)\sigma^2}{\frac{1}{2}\sigma^2} = 2\left(\frac{\pi}{2} - 1\right) = \pi - 2 > 1.$$

What can be done when CLT cannot be used to obtain the asymptotic distribution?

Example. Consider i.i.d observations $X_1, X_2, \ldots$ from a location family with θ = the median. i.e., the density is $f(x|\theta) = f_0(x - \theta)$ and the cdf $F_\theta(x) = F_0(x - \theta)$, and further, $F_\theta(\theta) = F_0(0) = 1/2$ and $f_0(0) > 0$. Then we have the following result.

Result.

$$\sqrt{n}\left(\text{median}(X_1, X_2, \ldots, X_n) - \theta\right) \xrightarrow[n\to\infty]{d} N\left(0, \frac{1}{4f_0^2(0)}\right).$$

Remark. Note that there are no conditions on the moments of X.

Proof. Let $n = 2m-1$, an odd integer, so that the median of $X_1, X_2, \ldots, X_n$ is $X_{(m)}$. Also, let $Y = X - \theta$. Then $X_{(m)} - \theta = Y_{(m)}$. Fix a and consider

$$\begin{aligned} F_n(a) &= P_\theta\left(\sqrt{n}(X_{(m)} - \theta) \le a\right) \\ &= P\left(\sqrt{n}Y_{(m)} \le a\right) = P\left(Y_{(m)} \le \frac{a}{\sqrt{n}}\right). \end{aligned}$$

Let S_n = number of Y's that exceed $\frac{a}{\sqrt{n}}$. Then

$$S_n \sim \text{Binomial}\left(n, p_n = 1 - F_0(\frac{a}{\sqrt{n}})\right).$$

Also $Y_{(m)} \le \frac{a}{\sqrt{n}}$ iff number of Y_i that exceed $\frac{a}{\sqrt{n}}$ is less than or equal to $m-1$ iff $S_n \le m - 1 = \frac{n-1}{2}$. Therefore,

$$\begin{aligned} P\left(Y_{(m)} \le \frac{a}{\sqrt{n}}\right) &= P\left(S_n \le \frac{n-1}{2}\right) \\ &= P\left(\frac{S_n - np_n}{\sqrt{np_n(1-p_n)}} \le \frac{\frac{n-1}{2} - np_n}{\sqrt{np_n(1-p_n)}}\right). \end{aligned}$$

From CLT for S_n, we obtain

$$P\left(\frac{S_n - np_n}{\sqrt{np_n(1-p_n)}} \le \frac{\frac{n-1}{2} - np_n}{\sqrt{np_n(1-p_n)}}\right) - \Phi\left(\frac{\frac{n-1}{2} - np_n}{\sqrt{np_n(1-p_n)}}\right) \xrightarrow[n\to\infty]{} 0.$$

Now let

$$x_n = \frac{\frac{n-1}{2} - np_n}{\sqrt{np_n(1-p_n)}} = \frac{\sqrt{n}(\frac{1}{2} - p_n) - \frac{1}{2\sqrt{n}}}{\sqrt{F_0(\frac{a}{\sqrt{n}})(1 - F_0(\frac{a}{\sqrt{n}}))}}.$$

Then

$$\begin{aligned}
\lim_{n\to\infty} x_n &= \frac{\lim_{n\to\infty} \sqrt{n}(\frac{1}{2} - p_n) - 0}{\sqrt{F_0(0)(1 - F_0(0))}} \\
&= 2 \lim_{n\to\infty} \sqrt{n} \left(\frac{1}{2} - (1 - F_0(\frac{a}{\sqrt{n}}) \right) \\
&= 2a \lim_{n\to\infty} \frac{F_0(a/\sqrt{n}) - F_0(0)}{a/\sqrt{n}} \\
&= 2aF_0'(0) = 2af_0(0).
\end{aligned}$$

Therefore,

$$F_n(a) = P_\theta \left(\sqrt{n}(X_{(m)} - \theta) \leq a \right) \underset{n\to\infty}{\longrightarrow} \Phi \left(2f_0(0)a \right).$$

Hence, the density of the asymptotic distribution is

$$\begin{aligned}
\frac{d}{da} \Phi \left(2f_0(0)a \right) &= 2f_0(0)\phi \left(2f_0(0)a \right) \\
&= 2f_0(0) \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{1}{2} (2f_0(0))^2 a^2 \right) \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{(2f_0(0))^{-1}} \exp \left(-\frac{1}{2} \frac{a^2}{[(2f_0(0))^{-1}]^2} \right),
\end{aligned}$$

which is that of $N\left(0, (2f_0(0))^{-2}\right)$.

Example. Let $X_1, X_2, \ldots$ be a random sample from a population with symmetric density, mean θ, variance equal to 1 and $f_X(\theta) > 0$. Consider $T_n^{(1)} = \bar{X}_n$ and $T_n^{(2)} = \text{median}(X_1, X_2, \ldots, X_n)$. From the above results,

$$\sqrt{n}\left(T_n^{(1)} - \theta\right) = \sqrt{n}\left(\bar{X}_n - \theta\right) \xrightarrow[n\to\infty]{d} N(0,1), \text{ and}$$

$$\sqrt{n}\left(T_n^{(2)} - \theta\right) = \sqrt{n}\left(\text{median}(X_1, X_2, \ldots, X_n) - \theta\right) \xrightarrow[n\to\infty]{d} N\left(0, \frac{1}{4f_0^2(0)}\right), \text{ so}$$

$$\text{e}\left(\theta, T_n^{(1)}, T_n^{(2)}\right) = \frac{1}{4f_\theta^2(\theta)}.$$

Example. Suppose $X \sim N(\theta, 1)$ in the above example. Then $f_\theta(\theta) = \frac{1}{\sqrt{2\pi}} = f_0(0)$, so that $4f_\theta^2(\theta) = \frac{4}{2\pi} = \frac{2}{\pi}$. Therefore $\text{e}\left(\theta, T_n^{(1)}, T_n^{(2)}\right) = \pi/2 \approx 1.57$. If we consider X such that $X - \theta \sim t_\nu$ for various values of ν, we obtain the following table listing $\text{e}\left(\theta, T_n^{(2)}, T_n^{(1)}\right)$.

ν	3	4	5	8	∞
ARE	1.62	1.12	0.96	0.80	$\frac{2}{\pi} \approx 0.64$

Note that the sample median is more efficient than the sample mean when $\nu \leq 4$. In fact, the sample mean does not provide a consistent estimator when $\nu = 1$ since t_1 which is the same as Cauchy which does not have a mean. For $\nu = 2$ the variance is not finite.

Information bound and asymptotically efficient estimators

Suppose $T_n(X_1, \ldots, X_n)$ is an estimator of $q(\theta)$. Then the Information Inequality says

$$Var_\theta(T_n) \geq \frac{\left[\frac{\partial}{\partial\theta} E_\theta(T_n)\right]^2}{I_n(\theta)},$$

where $I_n(\theta) = nI_1(\theta)$ if X_i are i.i.d P_θ. Suppose further that T_n is asymptotically normal, in the sense,

$$\sqrt{n}\left(T_n - q(\theta)\right) \xrightarrow[n\to\infty]{d} N(0, \sigma^2(\theta)) \text{ and}$$

$$\sqrt{n}\left(E_\theta(T_n) - q(\theta)\right) \xrightarrow[n\to\infty]{} 0.$$

Then $E_\theta(T_n) = q(\theta) + o\left(\frac{1}{\sqrt{n}}\right)$ so that

$$\frac{\partial}{\partial\theta} E_\theta(T_n) = q'(\theta) + o\left(\frac{1}{\sqrt{n}}\right) \xrightarrow[n\to\infty]{} q'(\theta).$$

Then the Information bound for $\sigma^2(\theta)$ reduces to $(q'(\theta))^2/I_1(\theta)$.

Definition. $T_n = T_n(X_1, \ldots, X_n)$ is said to be asymptotically efficient for estimating $q(\theta)$ if its asymptotic variance is

$$\sigma^2(\theta) = \frac{(q'(\theta))^2}{I_1(\theta)}.$$

Example. Let $X_1, X_2, \ldots$ be i.i.d $N(\theta, 1)$. Then $I_1(\theta) = 1$. Consider $T_n^{(1)} = \bar{X}_n$. Then $\sqrt{n}\left(T_n^{(1)} - \theta\right) \xrightarrow[n\to\infty]{d} N(0, 1)$. Note that

$$1 = \sigma^2(\theta) = \frac{1^2}{1} = \frac{1}{I_1(\theta)},$$

so that $T_n^{(1)} = \bar{X}_n$ is asymptotically efficient. Now consider $T_n^{(2)} = \text{median}(X_1, X_2, \ldots, X_n)$. Then $\sqrt{n}\left(T_n^{(2)} - \theta\right) \xrightarrow[n\to\infty]{d} N(0, \frac{\pi}{2})$. Note that, now,

$$\sigma^2(\theta) = \frac{\pi}{2} > \frac{1}{I_1(\theta)} = 1.$$

Thus $T_n^{(2)}$ is not asymptotically efficient.

Result. Under regularity conditions on the model density, MLE is consistent. i.e., $\hat{\theta}(X_1, X_2, \ldots, X_n) \xrightarrow[n\to\infty]{P} \theta$.

Result. Under regularity conditions on the model density, MLE is asymptotically normal and asymptotically efficient. i.e.,

$$\sqrt{n}\left(\hat{\theta}(X_1, X_2, \ldots, X_n) - \theta\right) \xrightarrow[n\to\infty]{d} N\left(0, \frac{1}{I_1(\theta)}\right).$$

However, the conditions are typically hard to verify. It is easier to prove the result on a case by case basis using other standard limit theorems.

Result. Under regularity conditions on the model and the prior, the Bayes estimate, $\tilde{\theta} = E^{\pi}(\theta|\mathbf{X})$ is asymptotically normal and asymptotically efficient, satisfying

$$\sqrt{n}\left(\tilde{\theta}(X_1, X_2, \ldots, X_n) - \theta\right) \xrightarrow[n\to\infty]{d} N\left(0, \frac{1}{I_1(\theta)}\right).$$

Sketch of asymptotic normality for MLE assuming regularity conditions:

Let $X_1, X_2, \ldots$ be an i.i.d sequence from P_θ having density $f(x|\theta)$, $\theta \in \Theta$ which satisfies the regularity conditions to be specified in the proof below. We have

$$L(\theta, \mathbf{X}) = \prod_{i=1}^{n} f(X_i|\theta),$$

$$\mathcal{L}(\theta, \mathbf{X}) = \log L(\theta, \mathbf{X}) = \sum_{i=1}^{n} \log f(X_i|\theta),$$

Let $\hat{\theta}_n$ be the MLE of θ such that

$$\mathcal{L}'(\hat{\theta}_n(\mathbf{X}), \mathbf{X}) = 0.$$

Condition 1: $\mathcal{L}$ is differentiable three times and $\mathcal{L}'(\hat{\theta}_n) = 0$.

Now we obtain,

$$0 \;=\; \mathcal{L}'(\hat{\theta}_n) = \mathcal{L}'(\theta) + (\hat{\theta}_n - \theta)\mathcal{L}''(\theta) + \frac{1}{2}(\hat{\theta}_n - \theta)^2\mathcal{L}'''(\theta_n^*),$$

where θ_n^* lies between $\hat{\theta}_n$ and θ. Therefore,

$$(\hat{\theta}_n - \theta)\left[\mathcal{L}''(\theta) + \frac{1}{2}(\hat{\theta}_n - \theta)\mathcal{L}'''(\theta_n^*)\right] = -\mathcal{L}'(\theta).$$

Hence,

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \frac{-\sqrt{n}\mathcal{L}'(\theta)}{\mathcal{L}''(\theta) + \frac{1}{2}(\hat{\theta}_n - \theta)\mathcal{L}'''(\theta_n^*)} \\ &= \frac{\frac{1}{\sqrt{n}}\mathcal{L}'(\theta)}{-\frac{1}{n}\mathcal{L}''(\theta) - \frac{1}{2n}(\hat{\theta}_n - \theta)\mathcal{L}'''(\theta_n^*)}.\end{aligned}$$

Assuming that $\frac{1}{\sqrt{n}}\mathcal{L}'''(\theta_n^*)$ is bounded (Condition 2), and that $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta$ (Condition 3, consistency of MLE), we obtain,

$$-\frac{1}{2n}(\hat{\theta}_n - \theta)\mathcal{L}'''(\theta_n^*) \xrightarrow[n\to\infty]{P} 0.$$

Note that, we can also get

$$\begin{aligned}\frac{1}{\sqrt{n}}\mathcal{L}'(\theta) &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta), \text{ with} \\ E\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right) &= 0, \quad \text{(Condition 4)} \\ E\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2 &= I_1(\theta). \quad \text{(Condition) 5}\end{aligned}$$

Then it follows from CLT that

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta) - 0\right] \xrightarrow[n\to\infty]{d} N(0, I_1(\theta)).$$

Therefore,

$$\frac{1}{\sqrt{n}}\mathcal{L}'(\theta) \xrightarrow[n\to\infty]{d} N(0, I_1(\theta)).$$

Note that

$$\begin{aligned} -\frac{1}{n}\mathcal{L}''(\theta) &= -\frac{1}{n}\frac{\partial}{\partial\theta}\mathcal{L}'(\theta) \\ &= -\frac{1}{n}\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta). \end{aligned}$$

Now assuming (Condition 6) that

$$-E\left(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right) = I_1(\theta),$$

we have from WLLN,

$$-\frac{1}{n}\mathcal{L}''(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta) \xrightarrow[n\to\infty]{P} I_1(\theta).$$

Finally, we obtain

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{d} \frac{1}{I_1(\theta)}N(0, I_1(\theta)) = N\left(0, \frac{1}{I_1(\theta)}\right).$$

Variance Stabilizing Transformations

Result. Suppose we have an i.i.d sequence $X_1, X_2, \ldots$ with $E(X_i) = \mu(\theta)$ and $Var(X_i) = \sigma^2(\theta) < \infty$. Then, from CLT, we have

$$\sqrt{n}\left(\bar{X}_n - \mu(\theta)\right) \xrightarrow[n\to\infty]{d} N\left(0, \sigma^2(\theta)\right).$$

Note, however, that s.e.$(\hat{\mu})$ = s.e.$(\bar{X}_n) = \sigma(\theta)/\sqrt{n}$ involves the usually unknown $\sigma^2(\theta)$. This may pose difficulties while using this for procedures such as large sample confidence intervals and tests. It is desirable to remove that dependence.

Recall that if h is differentiable and $h'(\mu) \neq 0$, then

$$\sqrt{n}\left(h(\bar{X}_n) - h(\mu)\right) \xrightarrow[n\to\infty]{d} N\left(0, (h'(\mu))^2\sigma^2\right).$$

If $(h'(\mu))^2\sigma^2 = (h'(\mu(\theta)))^2\sigma^2(\theta)$ can be made independent of θ for some h then the difficulty may be alleviated. For example, then, we could use

$$h(\bar{X}_n) \pm z_{1-\alpha/2}\frac{c}{\sqrt{n}}, \qquad c^2 = (h'(\mu))^2\sigma^2$$

as our confidence interval for $h(\mu)$. (Invert it to get a CI for μ.) In other words, $\frac{h(\bar{X}_n)-h(\mu)}{c/\sqrt{n}}$ becomes a pivot. *Variance stabilizing transformation* then is to find h such that $\sigma^2(\theta)(h'(\mu(\theta)))^2 \equiv c^2$.

Example. $S_n \sim$ Binomial(n, θ). $\bar{X}_n = S_n/n$. $\mu(\theta) = \theta$.

$$\sqrt{n}\left(\bar{X}_n - \mu(\theta)\right) \xrightarrow[n\to\infty]{d} N\left(0, \sigma^2(\theta) = \theta(1-\theta)\right).$$

We require h such that

$$\begin{aligned}
\sigma^2(\theta)(h'(\mu(\theta)))^2 &\equiv c^2. \text{ i.e.,} \\
(h'(\theta))^2 &= \frac{c^2}{\theta(1-\theta)}, \text{ or} \\
h'(\theta) &= \frac{c}{\sqrt{\theta(1-\theta)}}.
\end{aligned}$$

Solving it, we get $h(\theta) \propto \sin^{-1}(\sqrt{\theta})$. Therefore the required transformation is $h(x) = \sin^{-1}(\sqrt{x})$. (Note, $\sin(h(x)) = \sqrt{x} \implies \cos(h(x))h'(x) = \frac{1}{2\sqrt{x}} \implies h'(x) = \frac{1}{2\sqrt{x}}\frac{1}{\cos(h(x)))} = \frac{1}{2\sqrt{x}}\frac{1}{\sqrt{1-(\sqrt{x})^2}}$.)

Example. $Y_1, Y_2, \ldots$ i.i.d Poisson(λ). Then $\mu(\lambda) = \lambda = \sigma^2(\lambda)$. We have $\sqrt{n}\left(\bar{Y}_n - \lambda\right) \xrightarrow[n\to\infty]{d} N(0, \lambda)$. Find h such that

$$\begin{aligned}
\sigma^2(\lambda)(h'(\lambda))^2 &\equiv c^2. \text{ i.e.,} \\
(h'(\lambda))^2 &= \frac{c^2}{\lambda}, \text{ or} \\
h'(\lambda) &= \frac{c}{\sqrt{\lambda}}.
\end{aligned}$$

Solving it, we get $h(\lambda) \propto \sqrt{\lambda}$. Therefore the required transformation is $h(x) = \sqrt{x}$.

Large Sample Theory

References:
1. Bickel, D. and Doksum, K. *Mathematical Statistics*
2. Lehman, E. and Casella, G. *Theory of Point Estimation*

Convergence and Limit Theorems – Review

Let $X_1, X_2, \ldots$ be a sequence of random variables. Then there are different modes of convergence that apply (unlike sequences of real or complex numbers).

Definition. $X_n \xrightarrow{P} X$ as $n \to \infty$ (i.e., X_n converges to X in probability) if $P(|X_n - X| \geq \epsilon) \to 0$ as $n \to \infty$ for each $\epsilon > 0$.

Note that $P(|X_n - X| \geq \epsilon) \equiv P(\{\omega : |X_n(\omega) - X(\omega)| \geq \epsilon\})$.

Example. By the Weak Law of Large Numbers (WLLN), in the i.i.d case,
(i) $\bar{X} \xrightarrow[n\to\infty]{P} \mu$;
(ii) $\hat{p} \xrightarrow[n\to\infty]{P} p$.

Definition. $X_n \xrightarrow{a.s.} X$ as $n \to \infty$ (i.e., X_n converges to X almost surely or almost everywhere) if $P\left(\lim_{n\to\infty} X_n = X\right) = 1$.

Note again that $P\left(\lim_{n\to\infty} X_n = X\right) = P\left(\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right)$.

Example. By the Strong Law of Large Numbers (SLLN), in the i.i.d case,
(i) $\bar{X} \xrightarrow[n\to\infty]{a.s.} \mu$;
(ii) $\hat{p} \xrightarrow[n\to\infty]{a.s.} p$.

Definition. $X_n \xrightarrow{d} X$ as $n \to \infty$ (i.e., X_n converges to X in distribution) if $F_{X_n}(x) \to F_X(x)$ as $n \to \infty$ for all x where F_X is continuous.

Note that X_n and X need not be on the same space, or have a joint distribution.

Example. By the Central Limit Theorem (CLT), $\sqrt{n}(\bar{X} - \mu) \xrightarrow[n\to\infty]{d} N(0, \sigma^2)$.

Result. We have that

$$X_n \xrightarrow[n\to\infty]{a.s.} X \implies X_n \xrightarrow[n\to\infty]{P} X \implies X_n \xrightarrow[n\to\infty]{d} X.$$

Example. Consider independent $X_n \sim \text{Bernoulli}(\frac{1}{n})$. Then

$$P(|X_n - 0| > \epsilon) = P(X_n > \epsilon) = P(X_n = 1) = \frac{1}{n} \quad \underset{n\to\infty}{\longrightarrow} 0,$$

so that $X_n \xrightarrow[n\to\infty]{P} 0$, but does $X_n \xrightarrow[n\to\infty]{a.s.} 0$?

Result. We have

$$X_n \xrightarrow[n\to\infty]{d} X, \quad X \equiv c \text{ (a constant) then } X_n \xrightarrow[n\to\infty]{P} X.$$

Proof. We have $F_{X_n}(x) \to F_X(x)$ for all $x \neq c$. Thus, $F_{X_n}(c+\epsilon) \to 1$ and $F_{X_n}(c-\epsilon) \to 0$. Therefore,

$$\begin{aligned} P(|X_n - c| \geq \epsilon) &= P(X_n \leq c - \epsilon \text{ or } X_n \geq c + \epsilon) \\ &\leq F_{X_n}(c-\epsilon) + (1 - F_{X_n}(c + \epsilon/2)) \to 0. \end{aligned}$$

Result. Let g be a continuous function. Then

$$X_n \longrightarrow X \implies g(X_n) \longrightarrow g(X)$$

for all three modes of convergence.

This is easy to see for a.s. convergence.

Theorem (Slutsky). Suppose $X_n \xrightarrow[n\to\infty]{d} X$ and $U_n \xrightarrow[n\to\infty]{P} u_0$. Then
(a) $X_n + U_n \xrightarrow[n\to\infty]{d} X + u_0$;
(b) $U_n X_n \xrightarrow[n\to\infty]{d} u_0 X$.

Proof. Since $(X_n, U_n) \xrightarrow[n\to\infty]{d} (X, u_0)$, and $g(x,y) = x + y$ and $g(x,y) = xy$ are continuous functions, the result follows from the previous result.

Result. If $X_n - Y_n \xrightarrow[n\to\infty]{P} 0$ and $X_n \xrightarrow[n\to\infty]{d} X$, then $Y_n \xrightarrow[n\to\infty]{d} X$.

Proof. Note that $(X_n, X_n - Y_n) \xrightarrow[n\to\infty]{d} (X, 0)$. Therefore, $Y_n = X_n - (X_n - Y_n) \xrightarrow[n\to\infty]{d} X$, by Slutsky.

Chebychev's Inequality. For any random variable X, and $a > 0$,

$$P(|X| \geq a) \leq \frac{E(X^2)}{a^2}.$$

Proof. If $Y > 0$ and $a > 0$, we have

$$\begin{aligned} E(Y) &= \int_0^a y f_Y(y)\, dy + \int_a^\infty y f_Y(y)\, dy \\ &\geq \int_a^\infty y f_Y(y)\, dy \geq a P(Y \geq a), \end{aligned}$$

so that

$$P\left(Y \geq a\right) \leq \frac{E(Y)}{a}.$$

Therefore,

$$P\left(|X| \geq a\right) = P\left(X^2 \geq a^2\right) \leq \frac{E(X^2)}{a^2}.$$

The familiar form of this inequality is

$$P\left(|X - \mu| \geq \epsilon\right) = P\left((X-\mu)^2 \geq \epsilon^2\right) \leq \frac{E((X-\mu)^2)}{\epsilon^2} = \frac{Var(X)}{\epsilon^2}.$$

WLLN (Khinchine). If $X_1, X_2, \ldots$ are i.i.d such that $E(X)$ exists, then $\bar{X} \xrightarrow[n\to\infty]{P} E(X)$.

WLLN (Chebychev). If $X_1, X_2, \ldots$ is a sequence of random variables with $E(X_i) = \mu_i$, $Var(X_i) = \sigma_i^2$, $Cov(X_i, X_j) = 0$ for $i \neq j$, then

$$\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2 \underset{n\to\infty}{\longrightarrow} 0 \implies \bar{X}_n - \bar{\mu}_n \xrightarrow[n\to\infty]{P} 0.$$

Proof. Applying Chebychev's inequality,

$$\begin{aligned} P\left(|\bar{X}_n - \bar{\mu}_n| \geq \epsilon\right) &\leq \frac{1}{\epsilon^2} E\left(\bar{X}_n - \bar{\mu}_n\right)^2 = \frac{1}{\epsilon^2} E\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu_i)\right)^2 \\ &= \frac{1}{\epsilon^2}\frac{1}{n^2}\sum_{i=1}^n Var(X_i) \end{aligned}$$

If $\mu_i \equiv \mu$ then we obtain $\bar{X}_n \xrightarrow[n\to\infty]{P} \mu$ subject to conditions on σ_i^2 such as $\sigma_i^2 \equiv \sigma^2$. WLLN for an i.i.d sequence then is a special case.

SLLN (Kolmogorov). If $X_1, X_2, \ldots$ is an i.i.d sequence such that $E(X) = \mu$ exists and is finite, then $\bar{X}_n \xrightarrow[n\to\infty]{a.s.} \mu$.

CLT (Lindberg-Levy). Let $X_1, X_2, \ldots$ is an i.i.d sequence such that $E(X) = \mu$ and $Var(X) = \sigma^2$, $0 < \sigma^2 < \infty$. Then

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \xrightarrow[n\to\infty]{d} N(0,1).$$