MATH 357 - Honours Statistics

Pr. Masoud Asgharian

Course notes by
Léo Raymond-Belzile
Simon Szatmari
Leo.Raymond-Belzile@mail.mcgill.ca
Simon.Szatmari@mail.mcgill.ca

The current version is that of February 6, 2015
Winter 2012, McGill University

Please signal the author by email if you find any typo.

These notes have not been revised and should be read carefully.

Contents

Chapter 1
Sample moments and their distribution

Section 1.1. Sample characteristics and their distribution

Suppose that we want to estimate $F(s) = \mathsf{P}(X \leq s)$ using gathered data. Looking at the function F, a reasonable guess, if we have no further information about the random variables $X_1, \ldots, X_n$, is to consider giving equal weight to the observed realization $x_1, \ldots, x_n$ and look at the proportion falling in the category $X \leq s$ for a given value of s. We thus have

$$\widehat{F}_n(s) = \widehat{\mathsf{P}}_n(X \leq s) = \frac{1}{n} \#x_i\text{'s which are smaller or equal to } s$$

which can be written in a more formal way as

$$\widehat{F}_n(s) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{E}(s - x_i) \tag{1.1}$$

where

$$\mathcal{E}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{1.2}$$

This will be a random function if evaluated everywhere. First, we go with some definitions:

Definition 1.1 (Random sampling)

Let X be a random variable (rv) with distribution function F and let $X_1, \ldots, X_n$ be iid rv with common distribution function (DF) F. Then, the collection $X_1, \ldots, X_n$ is called a *random sample* of size n from DF F or simply as n independent observations on X.

Definition 1.2 (Statistic)

Let $X_1, \ldots, X_n$ be n independent observations on X and let $f : \mathbb{R}^n \to \mathbb{R}^k$ be a Borel measurable function. Then, the random variable $f(X_1, \ldots, X_n)$ is called a (sample) *statistic* provided that it is not a function of any unknown

parameter(s).

Example 1.1

Let $X_1, \ldots, X_n$ be a random sample from F. Then

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{1.3}$$

and

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 \tag{1.4}$$

are statistics, while $\overline{X}_n - \mu$ and S_n^2/σ^2 are not. Here, $\mu = E(X)$ and $\sigma^2 = \mathsf{Var}X$, F is the cumulative distribution function of $\mathcal{N}(\mu, \sigma^2)$; they are random variables and depend on an unknown parameter. (1.1) is there for $s = 2$, but in general it returns a cadlag function[1] and will be a random function. If we consider the mean of (1.1), the empirical distribution function (EDF) of a random sample from distribution function F, the expectation will be given by

$$\begin{aligned} \mathsf{E}\left[\widehat{F}_n(s)\right] &= \mathsf{E}\left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{E}(s - x_i)\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}\left[\mathcal{E}(s - x_i)\right] \end{aligned}$$

Recall that $\mathcal{E}(s)$ is a binary variable. If we look at $\mathsf{E}\,\mathcal{E}(s - x_i)$ as defined in (1.2), then $\mathsf{P}(\mathcal{E}(s - x_i) = 1) = \mathsf{P}(X_i \leq s) = F(s)$ so

$$\mathsf{E}\left[\mathcal{E}(s - x_i)\right] = F(s)$$

and going back to our calculation, we get

$$\frac{1}{n} \sum_{i=1}^{n} F(s) = F(s)$$

[1] The term cadlag stands for "continue à gauche, limite à droite".

This property (namely having the mean as being the value of our estimate at any point) is called *unbiasedness*; we will go back to it very soon. Does it necessarily make sense to look at this property? Not always; one could be considering the distribution of income and wondering whether the fact the mean is high is an indicator of how much people have as a living; we could very well be better of estimating with the median if we have some extremely high value. We could also think in a very silly way that a slap on the left side of the face doesn't cancel a slap on the right. If we look at the variance of the $\widehat{F}_n$

$$\mathsf{Var}\left(\widehat{F}_n(s)\right) = \mathsf{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}(s-x_i)\right] = \frac{1}{n^2}\mathsf{Var}\left[\sum_{i=1}^{n}\mathcal{E}(s-x_i)\right]$$

since $\mathcal{E}$ is a measurable function and X_i are mutually independent, then $Y_i = \mathcal{E}(s-x_i)$ are independent and

$$Y_i = \begin{cases} 1 & F(s) = p \\ 0 & 1 - F(s) = 1 - p = q \end{cases}$$

and $\mathsf{P}(Y_1 = 1) = \mathsf{P}(X_i \leq s) = F(s)$ so $\mathsf{Var}(Y_i) = p(1-p) = F(s)[1-F(s)]$ and

$$\mathsf{Var}\left(\widehat{F}_n(s)\right) = \frac{1}{n^2}\sum_{i=1}^{n}p(1-p) = \frac{p(1-p)}{n} = \frac{F(s)[1-F(s)]}{n}$$

We can easily show that

Theorem 1.3

$$\mathsf{P}\left(\widehat{F}_n(s) = \frac{j}{n}\right) = \binom{n}{j}\left(F(s)\right)^j\left(1-F(s)\right)^{n-j}$$

for $j = 0, 1, \ldots, n$ since

$$\widehat{F}_n(s) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{E}(s-x_i) = \frac{1}{n}\sum_{i=1}^{n}Y_i$$

where Y_i are iid Bernoulli $\mathcal{B}(1, p = F(s))$ random variables, thus $\sum_{i=1}^{n} Y_i \sim \mathcal{B}(n, p = F(s))$; the sum is Binomial. Thus, getting $\widehat{F}_n(s)$ from above and

$$\mathsf{P}\left(\widehat{F}_n(s) = \frac{j}{n}\right) = \mathsf{P}(Z_n = j) = \binom{n}{j}[F(s)]^j[1 - F(s)]^{n-j},$$

we could immediately get the expected value and the variance from the theorem.

Corollary 1.4

The mean and the variance are from above

$$\mathsf{E}\left(\widehat{F}_n(s)\right) = F(s) \qquad \mathsf{Var}\left(\widehat{F}_n(s)\right) = \frac{F(s)(1 - F(s))}{n}$$

Corollary 1.5

We have convergence in probability of the empirical distribution function to the true distribution function:

$$\widehat{F}_n(s) \xrightarrow{\mathsf{p}} F(s)$$

using the Weak Law of Large Numbers (WLLN). We can also directly compute it using Chebychev inequality

$$\mathsf{P}\left(\left|\widehat{F}_n(s) - F(s)\right| > \varepsilon\right) \leq \frac{\mathsf{Var}\left(\widehat{F}_n(s)\right)}{\varepsilon^2}$$

and noting that $\mathsf{Var}\left(\widehat{F}_n(s)\right) \xrightarrow{n\to\infty} 0$, we have convergence for any fixed $\varepsilon > 0$ and even if it is random as long as the denominator of

$$\frac{F(s)[1 - F(s)]}{n\varepsilon^2}$$

goes to ∞. This property is called *consistency*; we have convergence to the target. This is crucial as when we take n large enough, we can get the population and we should clearly be having the true mean. Note this is necessary, but not sufficient.

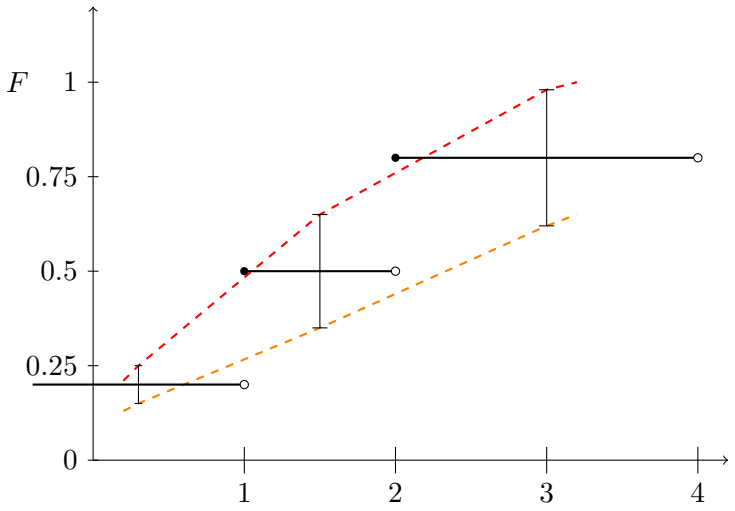


Figure 1: Confidence band

We can even get

Corollary 1.6

As a result of the Central Limit Theorem (CLT)

$$\frac{\widehat{F}_n(s) - F(s)}{\sqrt{\frac{F(s)[1-F(s)]}{n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

This is remarkable since with only the mean and the variance, we can get confidence intervals -we don't need the distribution from which they are drawn. What can I say about this as a function is beyond the scope of the course, but nevertheless here is an amazing result.

Theorem 1.7 (Glivenko-Cantelli theorem)

Suppose that $X_1, \ldots, X_n$ are independent and have a common distribution function F; put

$$D_n(\omega) = \sup_x |F_n(x,\omega) - F(x)|.$$

Then $D_n \to 0$ almost surely.

PROOF First, we note that $F_n(x,\omega) \xrightarrow{\text{a.s.}} F(x)$ using the Strong Law of Large Numbers (SLLN). Another application of the SLLN with $I_{(-\infty,x)}$ instead of $I_{(-\infty,x]}$ in the definition of the empirical cumulative distribution function (CDF) shows that

$$\lim_n F_n(x^-,\omega) = F(x^-)$$

except on a set B_x of probability 0.

Let $\phi(u) = \inf[x : u \le F(x)]$ for $0 < u < 1$ and put $x_{m,k} = \phi(k/m)$ for $m \ge 1, 1 \le k \le m$. It is not hard to see that $F(\phi(u)^-) \le u \le F(\phi(u))$; hence

$$\begin{aligned} F(x_{m,k}^-) - F(x_{m,k-1}) &\le \frac{1}{m} \\ F(x_{m,1}^-) &\le \frac{1}{m} \\ F(x_{m,m}) &\ge 1 - \frac{1}{m} \end{aligned}$$

Let

$$D_{m,n}(\omega) = \max\left\{\left|F_n(x_{m,k},\omega) - F(x_{m,k})\right|, \left|F_n(x_{m,k}^-,\omega) - F(x_{m,k}^-)\right|\right\}$$

for $k = 1,\ldots,m$. If $x_{m,k-1} \le x \le x_{m,k}$, then

$$F_n(x,\omega) \le F_n(x_{m,k}^-,\omega) \le F(x_{m,k}^-) + D_{m,n}(\omega) \le F(x) + \frac{1}{m} + D_{m,n}(\omega)$$

and

$$F_n(x,\omega) \ge F_n(x_{m,k-1},\omega) \ge F(x_{m,k-1}) - D_{m,n}(\omega) \ge F(x) - \frac{1}{m} - D_{m,n}(\omega)$$

Together with similar arguments for the cases $x < x_{m,1}$ and $x \ge x_{m,m}$, this shows that

$$D_n(\omega) \le D_{m,n}(\omega) + \frac{1}{m} \tag{1.5}$$

If ω lies outside the union of all the $A_{x_{m,k}}$ and $B_{x_{m,k}}$, then $\lim_{n\to\infty} D_{m,n}(\omega) = 0$ and hence using (1.5), $\lim_{n\to\infty} D_n(\omega) = 0$ except at $A = \bigcup_{m,k}(A_{x_{m,k}} \cup$

$B_{m,k}$) where $A_x = \{\omega \in \Omega : \lim_{n\to\infty} F_n(x\omega) \neq F(x)\}$. □

One can also show using the SLLN that

$$\|\widehat{F_n} - F\|_\infty \overset{\text{a.s.}}{=} O\left(\sqrt{\frac{\log\log n}{n}}\right).$$

Also important results in this area is Komlos-Major-Tusnady approximation and Dvoretzky-Kiefer-Wolfowitz inequality

$$\mathsf{P}\Big(\sup_x |F_n(x) - F(x)| > \varepsilon\Big) \leq 2e^{-2n\varepsilon^2} \qquad \forall\, \varepsilon > 0 \tag{1.6}$$

from which Glivenko-Cantelli theorem automatically follow. We even get the rate of convergence to 0. Another result allows one to make uniform bands

$$\sqrt{n}[\widehat{F}_n(s) - F(s)] \xrightarrow{w} B(F(s))$$

namely convergence to a Brownian motion, and where w stands for weak-* topology.

We could think that with all these results, statistics is known and problems are solve before they even arise. However, all the above was based on iid sample. Most of the time, what we have in hand (for example with medical data) contains missing data. We tend to have certain bias; we tend to recruit people that have longer survival. We may have partial observation, which leads to censoring. Let T_i be the time to event and C_i the time to loss to follow-up. Then we may have instead of the ideal (T_i, C_i) the actual observation will be (X_i, δ_i) where

$$X_i = T_i \wedge C_i \quad \delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i \\ 0 & \text{if } T_i > C_i \end{cases}$$

Lots of problems are encountered with our candidates in survival analysis. If the percentage of loss to follow-up is important, one cannot simply exclude them from the inference without introducing bias. You will underestimate if you throw some out since they have potential to survive. The survival $S(t)$

is defined as $1 - F(t) = \mathsf{P}(T > t)$ and for $X \geq 0$, the expectation of X will be

$$\mathsf{E}X = \int_0^\infty [1 - F(t)]dt = \int_0^\infty S(t)dt$$

All of this leads back to the problem of estimating $\widehat{F}_n$, which is the empirical distribution function. The well-known Kaplan-Meier estimator is a good example of function that estimate the fraction of patients that survive (a treatment). We proceed with some more definitions

Definition 1.8 (Point estimate)

Let $X_1, \ldots, X_n$ be a random sample of size n, from F_θ, where $\theta \in \Theta$. A statistic $T(X_1, \ldots, X_n)$ is said to be (point) estimate of θ if $T : \mathbb{R}^n \to \Theta$.

Definition 1.9 (Unbiasedness)

A point estimate of θ is called unbiased if $\mathsf{E}_\theta[T] = \theta$, where

$$\mathsf{E}_\theta[T] = \int t dF_{T,\theta}(t),$$

if $T = T(X_1, \ldots, X_n)$.

$$\mathsf{E}_\theta[T] = \int_{X_1} \cdots \int_{X_n} T(X_1, \ldots, X_n) \underbrace{dF_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}_{\prod_{i=1}^n dF_{\theta,X_i}(x_i)}$$

Note that we must have $\mathsf{E}|T| < \infty$, conditional convergence is not enough: if we do not have absolute convergence, we could rearrange the sum and make it converge to anything we want.

Definition 1.10 (Consistency)

Let $X_1, \ldots, X_n$ be a sequence of iid rv with a common DF, say $F_\theta, \theta \in \Theta$. A sequence of point estimates $T(X_1, \ldots, X_n) = T_n$ is called consistent for θ if $T_n \xrightarrow{\mathsf{p}} \theta$ as $n \to \infty$.

Recall $T_n \xrightarrow{\mathsf{p}} \theta$ iff $\lim_{n\to\infty} \mathsf{P}(|T_n - \theta| > \varepsilon) = 0, \forall \varepsilon > 0$. We are interested by $\mathsf{E}|T_n - \theta|$ as it is not realistic to look at each instance separately (we may have a huge number of points).

Mean Squared Errors (MSE)

$$\begin{aligned}\mathsf{E}[(T_n - \theta)^2] &= \mathsf{E}\Big[(\{T_n - \mathsf{E}_\theta(T_n) + \mathsf{E}_\theta(T_n) - \theta)^2\Big] \\ &= \mathsf{Var}_\theta(T_n) + (\mathsf{Bias}_\theta(T_n))^2\,,\end{aligned}$$

where $\mathsf{Bias}_\theta = \mathsf{E}_\theta(T_n) - \theta$.

We are interested in MSE because we have a good intuition about square norms (as in a Euclidean space), so we can work with concept like orthogonality more easily; moreover, x^2 is a nicer function to work with than $|x|$ (for instance because of smoothness). Finally, squared points translate well to matrix multiplication.

Note that we do not want MSE $= 0$, because in that case we not only don't have any bias (which is good), but we don't have a variance either (which renders the experiment useless). As the theory of statistics developed, statisticians aimed at unbiased estimators and tried to minimize the MSE.

Definition 1.11 (Order statistic)

Consider $X_1, \ldots, X_n$, a random sample, then the order statistic is simply an ordering of the X_i:

- $X_{(1)} = \min_{i=1,\ldots,n}(X_i)$,
- $X_{(2)} =$ second smallest of the X_i,
- $\ldots$
- $X_{(n)} = \max_{i=1,\ldots,n}(X_i)$.

Example 1.2

An example where we use the order statistic, would be in an electrical circuit. If we are interested in the lifetime of the system, then there will be two cases: If the circuit is a parallel system, then we are interested in $X_{(n)} = \max(X_i)$, where each X_i is a resistance and the system will die if all switches died. In

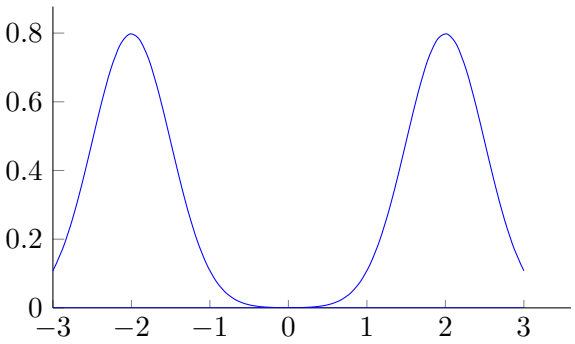


Figure 2: A bimodal distribution

lattice theory notation, we may write this as $X_{(n)} = \vee_{i=1}^{n} X_i$. If it is a series system, then the lifetime will be determined by the shortest lived resistance and this time we look at $X_{(1)} = \min(X_i) = \wedge_{i=1}^{n} X_i$.

Definition 1.12 (Sample Median)

The sample median is a non-parametric statistic. It is defined as follow: Let $X_1, \ldots, X_n$ be a random sample from F, and $X_{(1)}, \ldots, X_{(n)}$ be the order statistic. Then

$$Z_{\frac{1}{2}} = \begin{cases} X_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(X_{n/2} + X_{(n/2)+1}\right) & \text{if } n \text{ is even} \end{cases}$$

is called the sample median.

We must be careful when looking at the density; we can see from the cadlag function behaviour like that of a bimodal distribution for example, the average length of fish can be normally distributed, but with different mean for male or female. Note that if a distribution is symmetric, then $\mathsf{E} = Z_{\frac{1}{2}}$.

Section 1.2. χ^2, t and F-distributions

Say you have treatments $T_1, \ldots, T_k$ (for simplicity let $k = 2$), for instance on mice.

T_1	T_2
$X_{1,1}$	$X_{2,1}$
$\vdots$	$\vdots$
$X_{1,m}$	$X_{2,m}$

Here $X_{1,i}$ and $X_{2,i}$ are trials for the first and second treatment respectively. The biologist Ronald Fisher was the amongst the first to develop a theory of statistics that could deal with similar problems. In his era devoid of computers, an important problem was data compression, as a treatment can have several hundreds if not thousands of trials. It was near impossible to do the computations by hand. Once a table of data is summarized, how do we measure information? Do we use the average, and measure the *distance* of a trial to it? But what is information to begin with?

Fisher's Theory of Information

Suppose we have a density f_θ, for instance $\mathcal{N}(\mu, 1)$. Say that $f_\theta(5)$ is a constant function and that $f_\theta(2)$ resemble a wiggly polynomial. We can say that $f_\theta(5)$ is not sensitive to information as the function is not really dependent on the information θ, whereas $f_\theta(2)$ is. Stemming from this idea, Fisher found a way to measure this sensitivity to information with

$$\mathsf{E}\left[\frac{\partial^2 \log f_\theta(x)}{\partial \theta^2}\right]$$

which measures the *curvature* of $f_\theta(x_0)$ for some fixed value x_0 of x.
If we turn our attention to the summary, we might ask which qualities we want from it. If we were to decompress it, we would maybe want the likeliness of an event, and would want $\overline{X}_i$ to be sufficient and be able to absorb all

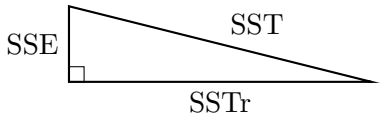


Figure 3: Geometric interpretation of SST, SSE and SSTr

the information about pertinent θ. Suppose that we consider $X_{ij} = p_i + \varepsilon_{ij}$, where $i = 1, \ldots, k$ is the number of treatments and $j = 1, \ldots, n_i$ the number of observations under the first treatment. Fisher assumed that the error term $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ where σ^2 is unknown. We define $\overline{X}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ and $\overline{X}_{..} = \frac{1}{k} \sum_{i=1}^{k} \overline{X}_{i.}$. Then, for our given experimental treatment, we will get the total variation to be the sum of random fluctuations plus the systematic deviation due to treatment, namely what we are interested in. This can be written as

$$\left(X_{ij} - \overline{X}_{..}\right) = \left(X_{ij} - \overline{X}_{i.}\right) + \left(\overline{X}_{i.} - \overline{X}_{..}\right) \tag{1.7}$$

Summing up over all i, j, we have

$$\underbrace{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{..}\right)^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{i.}\right)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\overline{X}_{i.} - \overline{X}_{..}\right)^2}_{\text{SSTr}} \tag{1.8}$$

which stands respectively for total sum of square (STT), sum of squared errors (SSE) and sum of squared treatment (SSTr).[2] If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ and we consider as a specific case $\mathcal{N}(\mu, \sigma^2)$.Then, $\overline{X}_n \perp\!\!\!\perp (X_1 - \overline{X}_n, \ldots, X_n - \overline{X}_n)$ will be our goal if $X_i \sim \mathcal{N}(\mu, \sigma^2)$ entails that $X_i = \mu + \sigma\varepsilon^*$ where $\varepsilon^* \sim \mathcal{N}(0, 1)$. Thus, assuming what σ is known, μ is unknown, we replace

$$X_i = \mu + \sigma\varepsilon^* = \mu + \varepsilon = \overline{X}_n + (X_1 - \overline{X}_n)$$

[2]It is also found under the term SSR under the literature, which stands for sum of squared residual. These notions are crucial when doing regression analysis; they allow you to get the coefficient of determination (R^2) which is an indicator of correlation and with which you can perform hypothesis testing.

for $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Given as seen in the figure that SST=SSE+SSTr, we may ask if $p_1 = \cdots = p_k$ which was Fisher's question: are all treatments working similarly. We can look at this ratio.

$$\frac{\text{SSTr}/(k-1)}{\text{SSE}/k(n-1)}$$

where the $(k-1)$ and $k(n-1)$ are the dimension of the space spanned by our vectors. A note about SSE:

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{i.} \right)^2$$

it expresses the random fluctuations that are intrinsic to our data and that we cannot model. On the other hand, we have SSTr

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\overline{X}_{i.} - \overline{X}_{..} \right)^2$$

is the systematic fluctuation due to the treatment. If SSTr is almost as large as SSE, then we have something not significant.

A word about the constants; they represent the dimension of the spaces. Using linear algebra and looking at (1.7), we can define

$$\begin{aligned} v &= \left\{ \left(X_{ij} - \overline{X}_{..} \right); i = 1, \ldots, k; j = 1, \ldots, n_i \right\} \\ w &= \left\{ \left(\overline{X}_{i.} - \overline{X}_{..} \right); i = 1, \ldots, k \right\} \\ z &= \left\{ \left(X_{ij} - \overline{X}_{i.} \right); i = 1, \ldots, k; j = 1, \ldots, n_i \right\} \end{aligned}$$

and we see that the dimension of w is $k-1$ and z has dimension $(n-1)k$. By looking at span(w) and examining the linear relationship, if we add all terms, we get 0 since $\sum_{i=1}^{k} \left(\overline{X}_{i.} - \overline{X}_{..} \right) = 0$. We need to show that it cannot be less than $k-1$. This is left as an exercise and can be done via direct sums. We have that v, w are subspaces of z and span(v) is actually the biggest space. For we have $\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{..} \right) = 0$ and $\overline{X}_{..} = \frac{1}{k_n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$. Now,

we are assuming $n = n_1 = n_2 = \cdots = n_k$. Since we have a direct sum $\dim[\text{span}(v)] = kn - 1$ and

$$\dim[\text{span}(v)] = \dim[\text{span}(w)] + \dim[\text{span}(z)]$$

and it must be that $(kn-1) - (k-1) = k(n-1) = \dim[\text{span}(z)]$, which indeed justifies the constants; they are the dimension of the spaces spanned by our vectors. If F is too large, we reject the equality hypothesis. We will need to find the PDF of F. We shortly show that for $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, then $X_i^2 \sim \chi^2(1)$ and that $\sum_{i=1}^n X_i^2 \sim \chi^2(n)$.

Recall the moment generating function of a random variable X is $M_X(t) = \mathsf{E}(e^{tX})$ arises from the Laplace transform, which is a smooth one-to-one map. We can get an analog with the Fourier transform, namely the characteristic function, which always exists. If we have a random vector $\boldsymbol{X}$, then $M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathsf{E}(e^{\boldsymbol{t}^\top \cdot \boldsymbol{X}})$ where $\cdot$ is the inner product and where $\boldsymbol{t} = (t_1, \ldots, t_k)$ and $\boldsymbol{X} = (X_1, \ldots, X_k)$. This formula will become $\prod_{i=1}^n e^{t_i \cdot X_i}$ if each X_i are independent since any function is independent and the r.v. will be uncorrelated. We can consider simply the moment generating function of a single r.v. to be $M_{X_1}(t_1) = M_{\boldsymbol{X}}(t_1, 0, \ldots, 0)$. We have independence for $X \perp\!\!\!\perp Y \Leftrightarrow M_{X,Y}(s,t) = M_X(s) M_Y(t)$.

Proposition 1.13 (Chi-square distribution)

We can recover the distribution of the χ^2 random variable through the method of MGF, through a change of variable or with the method of distribution function. Suppose that $Z \sim \mathcal{N}(0,1)$ and define $X = Z^2$. We want to find

$$\begin{aligned} F_X(x) &= \mathsf{P}(X \le x) = \mathsf{P}(|Z| \le \sqrt{x}) \\ &= \mathsf{P}(-\sqrt{x} \le Z \le \sqrt{x}) = \mathsf{P}(-\sqrt{x} < Z \le \sqrt{x}) \\ &= \mathsf{P}(Z \le \sqrt{x}) - \mathsf{P}(Z \le -\sqrt{X}) = F_Z(\sqrt{X}) - F_Z(-\sqrt{X}) \end{aligned}$$

We have made a relationship between the CDF of Z and that of X. In

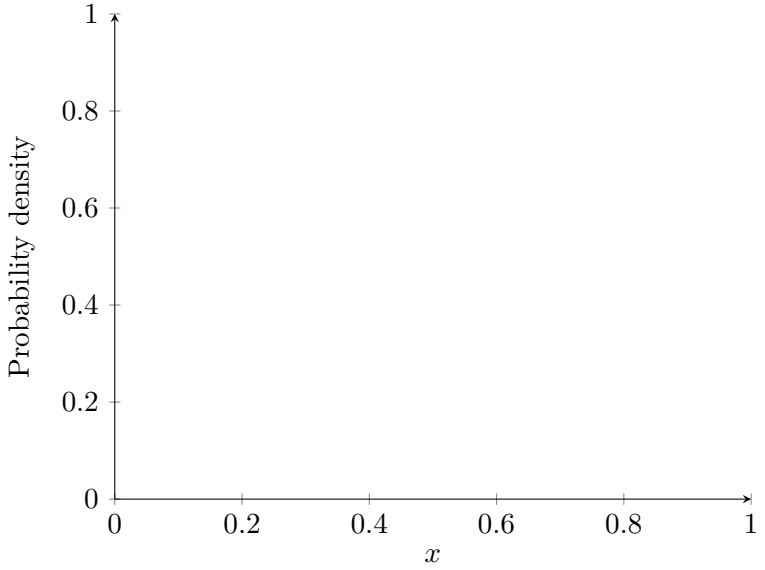


Figure 4: Chi-square random variable
As the number of degrees of freedom increase, the distribution becomes less skewed to the left.

order to work with the PDF, we differentiate

$$\frac{dF_X(x)}{dx} = \frac{d[F_Z(\sqrt{x}) - F_Z(-\sqrt{x})]}{dx}$$

$$f_X(x) = \frac{1}{2\sqrt{x}} f_Z(\sqrt{x}) - \left(-\frac{1}{2\sqrt{x}}\right) f_Z(-\sqrt{x}) = \frac{1}{2\sqrt{x}}[f_Z(\sqrt{x}) + f_Z(-\sqrt{x})]$$

which evaluated at $\sqrt{x}$ is equal to

$$= \frac{1}{2\sqrt{x}}\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{(\sqrt{x})^2}{2}} + \frac{1}{\sqrt{2\pi}} e^{-\frac{(\sqrt{x})^2}{2}}\right] = \frac{1}{\sqrt{x}\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

provided $x \geq 0$. One can rewrite this slightly differently as

$$f_X(x) = \begin{cases} \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} x^{-\frac{1}{2}} e^{-\frac{x^2}{2}}, & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

This is close to the Gamma distribution, which came to existence due to the work of Erlang, a Danish engineer that worked in communication and wanted to study the distribution of the duration of calls using the phone central by looking at the histograms.

Definition 1.14 (Gamma distribution)

We say that $X \sim \Gamma(\alpha, \beta)$ if

$$\begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{2}}, & x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1.9}$$

for $\alpha, \beta > 0$. This is used for example in queueing theory. We have χ^2 in the case $\beta = 2, \alpha = \frac{1}{2}$ and

$$\chi^2(\nu) = \Gamma\left(\alpha = \frac{\nu}{2}, \beta = 2\right) \tag{1.10}$$

with the parameter ν being the number of degrees of freedom.

Proposition 1.15 (Sum of Chi-square r.v.)

If we have that $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, for $i = 1, \ldots, n$, then $X_i^2 \overset{\text{iid}}{\sim} \chi^2(1)$ and

$T = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$. Then, using the moment generating function,

$$\begin{aligned} M_T(s) &= \mathsf{E}(e^{sT}) = \mathsf{E}e^{s\sum_{i=1}^{n} X_i^2} \\ &= \mathsf{E}\left(\prod_{i=1}^{n} e^{sX_i^2}\right) \overset{\perp\!\!\!\perp}{=} \prod_{i=1}^{n} \mathsf{E}\left(e^{sX_i^2}\right) \\ &= \prod_{i=1}^{n} M_{X_i^2}(s) \overset{iid}{=} [M_{X^2}(s)]^n \\ &= \left(\frac{1}{\sqrt{1-2t}}\right)^n \end{aligned}$$

if $t < \frac{1}{2}$ and $(1-2t)^{\frac{n}{2}}$ is χ^2 with n degrees of freedom.

If we go back to Fisher's problem, if we consider $\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \overline{X}_{1.})^2$ is $\mathcal{N}(0, \sigma^2)$; we can divide both numerator and denominators to get it distributed as $\mathcal{N}(0,1)$ random variable. If furthermore $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \ldots, n$, then looking at the difference between our random variables and the sample mean, we obtain

$$X_i - \overline{X}_n = X_i - \left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} a_j X_j$$

where

$$a_j = \begin{cases} -1/n & j \neq 1 \\ 1 - 1/n & j = 1 \end{cases}$$

We are working with a bunch of linearly independent random variables.

$$\begin{aligned}
M_{\Sigma_{j=1}^n a_j X_j}(t) &= \mathsf{E}\left(e^{t\,\Sigma_{j=1}^n a_j X_j}\right) \\
&= \mathsf{E}\left(\prod_{j=1}^n e^{t a_j X_j}\right) \\
&\overset{\perp\!\!\!\perp}{=} \prod_{j=1}^n M_{X_j}(t a_j) \\
&= \prod_{j=1}^n \left(\exp\left(t a_j \mu + \frac{t^2 a_j^2 \sigma^2}{2}\right)\right) \\
&= \exp\left(\mu t \sum_{j=1}^n a_j + \frac{t^2\sigma^2}{2}\sum_{j=1}^n a_j^2\right)
\end{aligned}$$

If we look at

$$\sum_{j=1}^n a_j = -\frac{n-1}{n} + \left(1 - \frac{1}{n}\right) = 0$$

implies

$$\begin{aligned}
\sum_{j=1}^n a_j^2 &= (n-1)\frac{1}{n^2} + \left(1 - \frac{1}{n}\right)^2 \\
&= \frac{n-1}{n^2} + 1 + \frac{1}{n^2} - \frac{2}{n} \\
&= \frac{n}{n^2} + 1 - \frac{2}{n} = 1 - \frac{1}{n}.
\end{aligned}$$

Hence,

$$M_{X_i - \overline{X}_n}(t) = M_{\Sigma_{j=1}^n a_j X_j}(t) = e^{\frac{t^2\sigma^2}{2}\left(1-\frac{1}{n}\right)}$$

and we deduce from the last step that $X_1 - \overline{X}_n \sim \mathcal{N}\left(0, \sigma^2\left(\frac{n-1}{n}\right)\right)$.

Why are we interested in summations? One can think they are important because we are looking at averages; we will see this again with likelihood (which can be though of as the likeliness of the sample at hand), we will have joint distribution and if we have that our random variables are independent of each other, then it will be the product. However, we can easily take logs and recover summations (this is called the log-likelihood). For the moment,

however, the $\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\overline{X}_{1.})^2$ are not independent.

Definition 1.16 (Snedecor-Fisher distribution)

Let X, Y be *independent* χ^2 random variables with respectively m and n degrees of freedom. Then,

$$F = \frac{X/m}{Y/n} \tag{1.11}$$

is said to have a $F-$distribution $F \sim F(m,n)$ with (m,n) degrees of freedom and with probability distribution function

$$F(m,n) = g_F(x) = \frac{\Gamma\left[\frac{m+n}{2}\right]}{\Gamma\left[\frac{m}{2}\right]\Gamma\left[\frac{n}{2}\right]}\frac{m}{n}\left(\frac{m}{n}x\right)^{\frac{m}{2}-1}\left(1+\frac{m}{n}x\right)^{-\frac{m+n}{2}}, \quad x>0$$

and 0 otherwise.

To get the PDF, call the ratio $V = \frac{X/m}{Y/n}$ and introduce an auxiliary transformation $W = Y$, looking at $(X,Y) \to (V,W)$ and applying the change of variable to get the joint density $f_{V,W}$ and we integrate to obtain

$$f_V(v) = \int_w f_{V,W}(v,w)dw$$

Be careful about the range when you do the transformation. Could we just have an injective map between the two? We can have a bijection since the cardinality $\mathbb{R}^k \to \mathbb{R}$ is the same, but we don't have a smooth map.

If the graph of our distribution is not monotone up and has several maxima and minima, then we can partition the graph into parts that are each monotone up or monotone down and apply the Law of total probability to sum all parts. The question one may ask is whether we often run into this type of distribution. In practise, the answer is not really and by Morse-Sard theorem, we know that the set of critical values has measure zero if the function is analytic. (if it is a $\sigma-$compact set, than it is countable).

We will now turn our attention to the work of William Gosset, from Iowa State University, who worked with agricultural studies on fertilizer and we

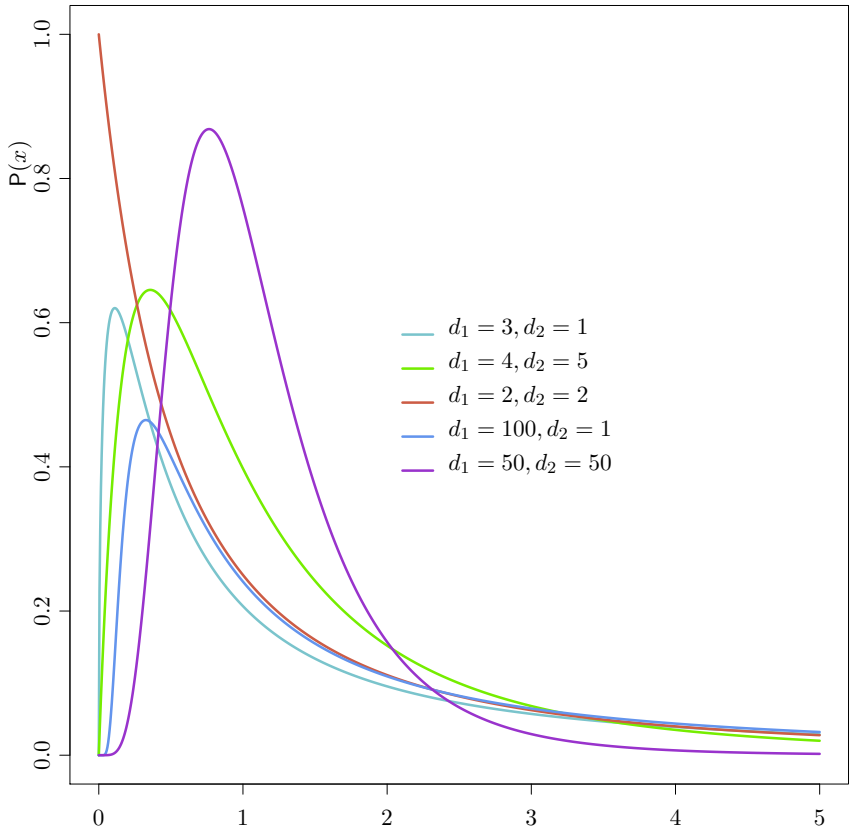


Figure 5: Fisher distribution

carry his idea in the easy case where we have only two different treatments T_1 and T_2 and we have samples that we summarize in a table of the following form:

T_1	X_1	X_2	...	X_n	$\overline{X}_n$
T_2	Y_1	Y_2	...	Y_n	$\overline{Y}_n$

A natural way to make the comparison is to look at the absolute difference between the two means of the treatments and check whether the difference was small or large. However, this approach is not unit free and changing the measurement unit would affect the inference results; it should not be

different if our measurements are in terms of metres or centimetres; it is the reason why we divide by the standard deviation to get a dimensionless measure of distance, that is invariant with respect to scaling. Gosset looked at the estimator

$$\frac{\left|\overline{X}-\overline{Y}\right|}{\sqrt{\overline{\overline{X}-\overline{Y}}}}.$$

If we open this up, we get

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-Y_i)}{\sqrt{\mathsf{Var}(\overline{X}-\overline{Y})}}=\frac{\overline{Z}}{\sqrt{\mathsf{Var}\overline{Z}}}$$

which becomes for 1 sample like a one sample test.

If $X_i \sim \mathcal{N}(\mu_x, \sigma^2), Y_i \sim \mathcal{N}(\mu_y, \sigma^2)$, for $i = 1, \ldots, n$ and $X_i \perp\!\!\!\perp Y_i$ are independent within groups and between groups, then $Z_i \sim \mathcal{N}(\mu_x - \mu_y, 2\sigma^2)$.

Often, we don't have the variance and we have to replace it with S.

$$\frac{\overline{Z}}{\sqrt{\mathsf{Var}\overline{Z}}} \sim \frac{\overline{Z}}{\frac{S_z}{\sqrt{n}}} \qquad \mathsf{Var}\overline{Z}=\frac{\sigma_z^2}{n}$$

Another distribution that we come across quite often in statistics is the t-distribution.

Definition 1.17 (Student t distribution)

If $X \sim \mathcal{N}(0,1)$ and $X \perp\!\!\!\perp Y, Y \sim \chi^2(n)$, then $T = \frac{X}{\sqrt{Y/n}} \sim t(n)$ -the Student-t distribution - with PDF given by

$$f_n(t)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \qquad t\in\mathbb{R}.$$

has a bell-curve with heavier tails.

Note that for $n = 1$, T is a Cauchy random variable. The tail of this distribution die polynomially fast; there only are $n - 1$ existing moments.

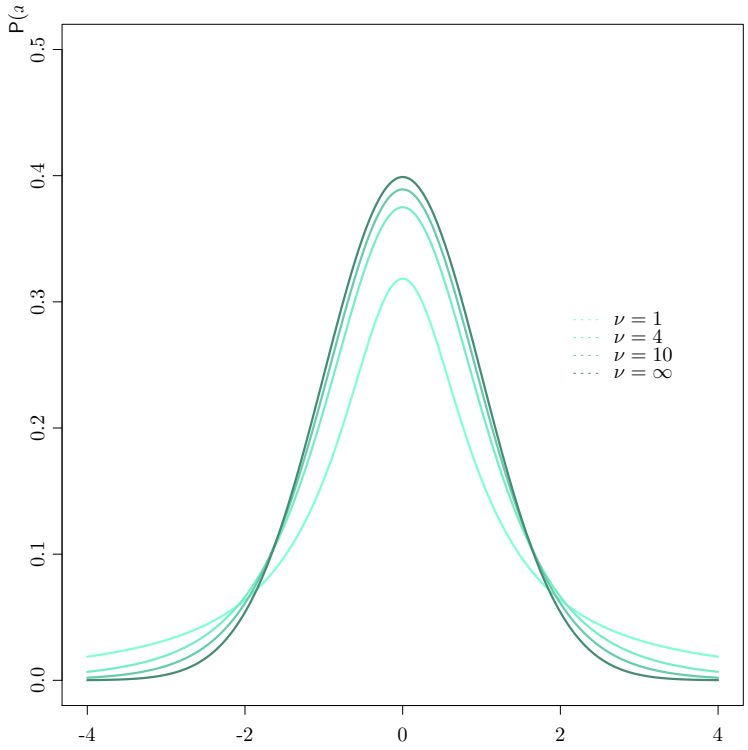


Figure 6: Student t distribution

$f_n(t)$ is symmetric around 0 and as $t \to \pm\infty$, we also note that

$$\lim_{n\to\infty} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} = e^{-\frac{t^2}{2}}.$$

which is the kernel of the PDF of the normal random variable. We only need to show the rest converges to one, *i.e.*

$$\lim_{n\to\infty} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\left(\frac{n}{2}\right)}} = 1.$$

The proof of this is left as an exercise (hint: use the formula $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$). Then $t-$distribution converges to the Normal distribution and is quite often used in finance with the Black-Scholes option-pricing (Lévy pro-

cesses). There is an interesting relationship between the t-distribution and the F-distribution.

$$F(1,n) \overset{\text{dist}}{=} t^2(n)$$

The Fisher distribution would be applicable to the above by raising the t to the power two to get χ^2 and $\frac{Y}{n}$ as a ratio of χ^2 variables. How to do it? As before, we use the transformation theorem using the auxiliary variable $W = Y$ (it will be easier to invert the Jacobian).

Here are a few things we need to know :

Proposition 1.18 (Moments of the t-distribution)

$$\mathsf{E}(X^r) = \begin{cases} n^{\frac{r}{2}} \frac{\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{n-r}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{n}{2}\right)} & \text{if } r < n \text{ and } r \text{ is even,} \\ 0 & \text{otherwise.} \end{cases}$$

In particular, if $n > 2$, $\mathsf{E}(X) = 0$, $\mathsf{E}(X^2) = \mathsf{Var}(X) = \frac{n}{n-2}$ and $\mathsf{E}[X^r] = 0$ if $r < n$ and odd. For the F-distribution,

$$\mathsf{E}[F^k] = \left(\frac{n}{m}\right)^k \frac{\Gamma\left(k + \frac{m}{2}\right)\Gamma\left(\frac{n}{2} - k\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}$$

In particular $\mathsf{E}(X) = \frac{n}{n-2}$ if $n > 2$ and $\mathsf{Var}(X) = \frac{n^2}{m}\frac{(2m-2n-4)}{(n-2)^2(n-4)}$ if $n > 4$. The simple trick to find the moments is to as usual write the definition and reformulate by rearranging the terms to recover the PDF of the F distribution with different parameters.

Section 1.3. Distribution of sample mean and sample variance when sampling from a Normal distribution

We begin with

Theorem 1.19

Let $X_1, \ldots, X_n$ be iid rv $\mathcal{N}(\mu, \sigma^2)$ rv's, then $\overline{X}$ and $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$ are independent.

PROOF We use the joint MGF $M_\zeta(t, t_1, \ldots, t_n)$ for $\zeta \equiv \{\overline{X}, X_1 - \overline{X}, \ldots, X_n - \overline{X}\}$.

$$\begin{aligned} M_\zeta(t, t_1, \ldots, t_n) &= \mathsf{E}\left[\exp\left\{t\overline{X} + \sum_{i=1}^n t_i(X_i - \overline{X})\right\}\right] \\ &= \mathsf{E}\left[\exp\left\{\sum_{i=1}^n t_i X_i - \left(\sum_{i=1}^n t_i - t\right)\overline{X}\right\}\right] \\ &= \mathsf{E}\left[\exp\left\{\sum_{i=1}^n X_i\left(t_i - \frac{t_1 + \ldots + t_n - t}{n}\right)\right\}\right] \end{aligned}$$

Let

$$\bar{t} = \frac{1}{n}\sum_{i=1}^n t_i$$

and continuing from above, we have

$$\begin{aligned} M_\zeta(t, t_1, \ldots, t_n) &= \mathsf{E}\left[\prod_{i=1}^n \exp\left\{\frac{X_i(nt_i - n\bar{t} + t)}{n}\right\}\right] \\ &= \prod_{i=1}^n \mathsf{E}\left[\exp\left\{\frac{X_i\left[t + n(t_i - \bar{t})\right]}{n}\right\}\right] \end{aligned}$$

where the last step follows from independence. Using the fact that $X_i \sim \mathcal{N}(\mu, \sigma^2)$, and adopting the convention

$$t^* = \frac{t + n(t_i - \bar{t})}{n},$$

we obtain

$$\begin{aligned}
&\prod_{i=1}^{n} \exp\left\{\mu t^* + \frac{\sigma^2}{2} t^{*^2}\right\} \\
&= \exp\left\{\frac{\mu}{n}\left[nt + n\sum_{i=1}^{n}(t_i - \bar{t})\right] + \frac{\sigma^2}{2n^2}\sum_{i=1}^{n}[t + n(t_i - \bar{t})]^2\right\} \\
&= \exp\left\{\mu t + \frac{\sigma^2}{2n} t^2\right\} \cdot \exp\left\{\frac{\sigma^2}{2}\sum_{i=1}^{n}(t_i - \bar{t})^2\right\} \\
&= M_{\overline{X}}(t) \cdot M_{X_1 - \overline{X}, \ldots, X_n - \overline{X}}(t_1, \ldots, t_n)
\end{aligned}$$

which in turn implies that $\overline{X}$ and $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$ are independent. □

Corollary 1.20

$\overline{X}$ and S^2 are independent.

Proof If $\overline{X}_n$ is independent from a vector, it is independent of any function of that vector. The result is immediate. □

Corollary 1.21

The ratio of the sample variance statistic with the true variance given by

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

follows a Chi-square distribution with $n - 1$ degrees of freedom.

Proof $X_i \sim \mathcal{N}(\mu, \sigma^2)$ thus $\frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1)$ hence

$$\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(1).$$

The X_i are independent so

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$$

Now, adding and subtracting $\overline{X}$:

$$\begin{aligned}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 &= \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 + n\frac{(\overline{X}-\mu)^2}{\sigma^2} \\ &= \underbrace{\frac{(n-1)S^2}{\sigma^2}}_{V} + \underbrace{\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2}_{W}\end{aligned}$$

since the cross terms vanishes (that is $\sum(X_i - \overline{X}) = 0$).

$$\begin{aligned} M_U(t) &= M_V(t)\cdot M_W(t) \\ (1-2t)^{-\frac{n}{2}} &= M_V(t)\cdot(1-2t)^{-\frac{1}{2}} \\ \Rightarrow \qquad M_V(t) &= (1-2t)^{-\frac{n-1}{2}}, \quad t<\frac{1}{2}\end{aligned}$$

Therefore $V \sim \chi^2(n-1)$. Why does this hold? Note here that

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \qquad \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

as

$$\begin{aligned}\mathsf{E}[e^{t_1 v + t_2 w}] &= \mathsf{E}[e^{t_1 v} e^{t_2 w}] \\ &= \mathsf{E}[e^{t_1 v}]\mathsf{E}[e^{t_2 w}]\end{aligned}$$

since $V \perp\!\!\!\perp W$. □

Corollary 1.22

The scaled difference between the arithmetic mean and the unknown μ given by

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim t(n-1)$$

follows a Student's t distribution.

PROOF Given

$$\sqrt{n}\frac{\overline{X}-\mu}{S} \sim \mathcal{N}(0,1)$$

and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

recall that $\overline{X}$ is independent from S^2, thus

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{(\overline{X}-\mu)/\frac{\sigma}{\sqrt{n}}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} = \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} \sim t(n-1)$$

□

Corollary 1.23

If X_i are iid and $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and Y_j are iid with $Y_j \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and $X_i \perp\!\!\!\perp Y_j$, then

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(m-1, n-1).$$

PROOF

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\left[(m-1)S_1^2/\sigma_1^2\right]/(m-1)}{\left[(n-1)S_2^2/\sigma_2^2\right]/(n-1)} \sim \frac{\frac{\chi^2(m-1)}{m-1}}{\frac{\chi^2(n-1)}{n-1}} \sim F(m-1, n-1).$$

□

Corollary 1.24

Under the conditions of Corollary 1.23, we have

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\left\{(m-1)\frac{S_1^2}{\sigma_1^2}+(n-1)\frac{S_2^2}{\sigma_2^2}\right\}}}\sqrt{\frac{m+n-2}{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}} \sim t(m+n-2)$$

and the difference between our two treatments sample means is given as

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right),$$

and also

$$\frac{(m-1)S_1^2}{\sigma_1^2} + \frac{(n-1)S_2^2}{\sigma_2^2} \sim \chi^2(m+n-2)$$

Chapter 2
Theory of point estimation

Section 2.1. Introduction

We will confine ourselves mostly to parametric settings. This just means that we will assume we know the form of the distribution up to a finite unknown number of parameters? How do we know whether our guess is correct; if we are testing relatively few characteristic, this will be done easily, however we run into trouble if we have a large number of parameters; the number of observations required to perform our analysis will be enormous. We can define residuals and fit a model, observing the residuals and testing for them; first visually -observing if our observed and fitted values are close. Keep in mind that when we are in real life working with real data, things are never perfect and we will look for a theory that can shed some light on the unknown. If we had reason to think for example that $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ has bell curve, but we don't necessarily know the centre nor the dispersion. Sometime, we have mixed knowledge: this leads to semi-parametric models, part of which as the name indicates is parametric while the other is non-parametric. Sir David Cox became famous forty years ago with his proportional hazards model. In such cases, we do not get to observe all observations. We have an initiating event: administrating the drug or stimulus in the context of clinical trial, we follow patients with the two competing risks: time-to-death T_i and time-to-dropout C_i. This relates to the example from the beginning, with $X_i = T_i \wedge C_i$, with the goal that our estimate F_T leads to T_i

A word about the Kaplan-Meier estimates. Suppose that we have some indications of differences between genders or races. We could think of a controversial example (for political issues) about the fact that North American black people have a higher blood pressure. This may be reasoned thinking there was a pre-selection when they were chosen as slave back in Africa and that people with saltier skins were selected, as they would require less water

for the journey. But there may be more covariates in our estimate - we want to know how they affect F_T. If we define

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

respectively the PDF on the numerator and CDF in the denominator, then

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathsf{P}(t \leq T \leq t + \Delta t | T > t)$$

that is given that someone has survived to this point what is the probability that he/she die in a short time (essentially in a very local interval). If we were to work this out, it would yield our ratio, since by definition $\mathsf{P}(A|B) = \mathsf{P}(A \cap B)/\mathsf{P}(B) = \mathsf{P}(A)/\mathsf{P}(B)$ if $A \subset B$ is a small interval. Given the covariates, consider $\lambda(t|Z) = \lambda_0(t)e^{\beta^\top Z}$ and let M be a binary variable for male, F stand for female; we have

$$\frac{\lambda(t|M)}{\lambda(t|F)} = \frac{\lambda_0(t)e^{\beta_1 \cdot 1}}{\lambda_0(t)e^{\beta_1 \cdot 0}} = e^{\beta}$$

if we can make this, than this is robust. There are only finitely many β and this is parametric, while the choice for $\lambda_0(t)$ is infinite and this part is non-parametric; thus we say that this will be as a whole semi-parametric. We pursue now with a

Definition 2.1 (Parameter space)

The set of all admissible values of the parameters of a DF F is called a *parameter space* and denoted (often) by Θ.

Example 2.1

Suppose $X \sim \mathcal{B}(n, p)$, where p is unknown. The best example of this is a referendum with a split between two options; we could have a few million people in our population and we take a sample of size say 2000. Let

$$X_i = \begin{cases} 1 & \text{yes/survive} \\ 0 & \text{no/failure} \end{cases}$$

and we want to predict the result of the referendum. In this case, we want

to find the proportion $p = \mathsf{P}(X_i = 1)$. By admissible, we mean that the parameter will make the above into a distribution function. Therefore,

$$\Theta = \{p \in \mathbb{R} : 0 \leq p \leq 1\}$$

is our parameter space. Another example: suppose $X \sim \mathcal{N}(\mu, \sigma^2)$; then

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\},$$

that is the pairs such that when we write the density function, we get a correct PDF (or PMF). In our case, for

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

will require $\sigma > 0$ and $\mu \in \mathbb{R}$. This is what is meant by admissibility. Another example: for the Gamma distribution, $X \sim \Gamma(\alpha, \beta)$ with α, β unknown. Then $\Theta = \{(\alpha, \beta); \alpha, \beta > 0\}$ otherwise the integral of the PDF will diverge [3]. Recall the PDF was given by (1.9) as

$$\frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{2}}, \quad x > 0$$

Also, we might want to recall definition 1.8 we made of point estimate at this stage.

Example 2.2

Suppose that $X_i \overset{\text{iid}}{\sim} \mathcal{P}(\lambda)$; this is a distribution that arises by taking the limit of a binomial, that is for $np_n \approx \lambda$

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} \to \mathsf{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

for $x = 0, 1, \ldots$ and will be a distribution of rare events. In this case, $\Theta = \{\lambda; \lambda > 0\}$. If we consider $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is a statistic, but could happen to be zero (if all observations are zero), so we will define it as $\mathsf{P}(X = 0) = 1$

[3] One can find more information about these integral in Bartle's book *The elements of real analysis* in the chapter on Euler's functions

if $\lambda = 0$ and if $\lambda > 0$, then $\mathsf{P}(X = x) = e^{-\lambda}\lambda^x/x!$, for $x = 0, 1, \ldots$

But we could also consider $T = 1, T = \sum_{i=1}^n \alpha_i X_i / \sum_{i=1}^n \alpha_i$ for $\alpha_i > 0$, which is as valid as the previous example. We have uncountably many estimators and even with some criteria, we will still have uncountably many despite the additional restrictions. What criteria helps us choosing, even in classes?

Section 2.2. Properties of estimators

The minimal conditions we require from an estimator is

Definition 2.2 (Consistency)

Let $X_1, \ldots, X_n$ be a sequence of iid rv with common DF $F_\theta, \theta \in \Theta$. A sequence of point estimators $T(X_1, \ldots, X_n) = T_n$ is called weakly consistent for θ if $T_n \xrightarrow{\mathsf{P}} \theta$ as $n \to \infty$, $\forall\, \theta \in \Theta$. If $T_n \xrightarrow{\text{a.s.}} \theta$,then we say T_n is strongly consistent

Recall that

$$T_n \xrightarrow{\mathsf{P}} \theta \quad \Leftrightarrow \quad \lim_{n\to\infty} \mathsf{P}(|T_n - \theta| > \varepsilon) \to 0 \tag{2.12}$$

and will differ from $\lim_{n\to\infty} \mathsf{P}\{\omega \in \Omega : |T_n(\omega) - \theta| > \varepsilon\}$ which is a function: to work with it, we need a metric. All topologies in finite-dimensional spaces are equivalent. Also, we introduce here notions that are tentatively covered in Honours Probability, namely Chernoff inequality and Hoeffding's bounds

Theorem 2.3 (Chernoff inequality)

Suppose $X_1, \ldots, X_n$ are independent random variables. Let $X = \sum_{i=1}^n X_i$ We can apply Markov's inequality to get the following inequality

$$\mathsf{P}(X \geq a) \leq \frac{\prod_{i=1}^n \mathsf{E}(e^{tX_i})}{e^{ta}} \tag{2.13}$$

Proof Using Markov inequality and applying the transformation $x \mapsto e^{tx}$

$$\mathsf{P}(X \geq a) \leq \mathsf{P}\left(e^{tX} \geq e^{ta}\right) \leq \frac{\mathsf{E}(e^{tX})}{e^{ta}} = \frac{\prod_{i=1}^n \mathsf{E}(e^{tX_i})}{e^{ta}}$$

since the MGF of the sum is the product of the individual MGF if the variables are independents. □

Theorem 2.4 (Hoeffding's inequality)

Let $X_1, \ldots, X_n$ be independent random variables such that $\mathsf{P}(X_i \in [a_1, b_i]) = 1$, then for $\overline{X}_n$, we have the following inequality

$$\mathsf{P}\left(\overline{X}_n - \mathsf{E}(\overline{X}_n) \geq t\right) \leq \exp\left(-\frac{2t^2n^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \tag{2.14}$$

This theorem is broader and is meant to give bounds for martingales.

Definition 2.5 (Martingale)

First note that a random (stochastic) processes is a family of random variables indexed by a parameter, say t, where $t \in T$. The set T may be countable or uncountable, and of one or several dimensions. In this latter case, we usually say random (stochastic) fields. An interesting example of applications of random fields can be found in statistical theory developed for PET and fMRI in brain mappings.

A martingale is a stochastic process with the following properties:
If $M_1, M_2, \ldots, M_n$ a sequence of random variables, then for all n: $\mathsf{E}|M_n| < \infty$ and

$$\mathsf{E}[M_n | M_1, \ldots, M_{n-1}] = M_{n-1}.$$

A random sum is an example of a martingale: for X_i orthogonal random variables with mean zero, and $M_n = \sum_{i=1}^n X_i$, with a sequence $\{M_t : t \in T\}$. The martingale increments $M_n - M_{n-1}$ will be orthogonal with mean zero. Much of the theory of martingale was developed by Joseph Doob.

Martingales are random processes and they are essentially summation of a bunch of mean zero orthogonal variables. In other words, a martingale process

$$\mathcal{M} = \{M_t : t \in T\}$$

where $M_t = \sum_{s \leq t} m_s$ where $m_s = M(s + \delta s) - M(s)$. The m_s so defined is called the martingale increment at time s. Now $\mathsf{E}[m_s] = 0$ and

$\mathsf{Cov}(m_s, m_u) = 0$ for any s and u. There are, of course, some other technical conditions, such as measurability and existence of moments involved. Because of this construction, you can guess that something like CLT should hold under rather general conditions for martingale processes.

Definition 2.6 (Stochastic integrals)

Suppose you have two random processes $X(t)$ and $Y(t)$. The stochastic integral topic revolves around defining $\int_t x(t,\omega)dy(t,\omega)$.

If one of the two processes has almost sure, on ω, continuous paths and the other one has almost sure, on ω, paths of bounded variation, you can define the aforementioned integral path by path using Stieltjes integral either directly or by invoking the integration by parts. However, the problem starts when these conditions are not met. For instance, if you want to find $\int_t B(t,\omega)dB(t,\omega)$, where $B(t)$ is a Brownian motion. We know that Brownian motions paths are almost surely continuous and almost surely are not of bounded variation. If you check the first chapter of the book by Phillip Protter, you will find a good overview of stochastic integrals. If you are more application oriented you can check Oksenda's book.

Example 2.3

Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{B}(n,p)$. Then,

$$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathsf{p}} p$$

as $n \to \infty$. This procedure mainly tells us that the sample proportion converge to the (population) proportion and allow one to determine how many samples we need to get within say 1% of the actual value. Applying the WLLN to $\{X_n\}_{n=1}^{\infty}$ and noting that $\mu = \mathsf{E}X = p$,

$$\frac{S_n - np}{n} = T_n - p \xrightarrow{\mathsf{p}}_{n\to\infty} 0$$

An important theorem that is used often when establishing consistency is Kolmogorov's Law of Large Numbers.

Theorem 2.7 (Kolmogorov's Strong Law of Large Numbers)

Let $X_1, X_2, \ldots$ be iid random variables with common law $\mathcal{L}(X)$(common cumulative distribution function F). Then

$$\frac{S_n}{n} = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow[n\to\infty]{\text{a.s.}} \mu = \mathsf{E}X \quad \Leftrightarrow \mathsf{E}|X| < \infty. \tag{2.15}$$

We will not attempt to prove this result, as it would require several pages and the use of Kolmogorov's maximal inequality. An easier proof would draw on martingales, but this is beyond the scope of the course.

Using Kolmogorov's theorem, we have the following result

Theorem 2.8

If $X_1, X_2, \ldots$ are iid random variables with common CDF F and $\mathsf{E}|X|^p < \infty$ for some positive integer p, then

$$\frac{1}{n}\sum_{i=1}^n X_i^k \xrightarrow{\text{a.s.}} \mathsf{E}X^k \quad \text{for } 1 \le k \le p$$

and therefore $n^{-1}\sum_{i=1}^n X_i^k$ is a consistent estimator for $\mathsf{E}X^k$ for $1 \le k \le p$. Moreover, if c_n is any sequence of constraints such that $c_n \to 0$ as $n \to \infty$, then

$$\frac{1}{n}\sum_{i=1}^n X_i^k + c_n \xrightarrow{\text{a.s.}} \mathsf{E}X^k$$

for $1 \le k \le p$, *i.e.* $n^{-1}\sum_{i=1}^n X_i^k + c_n$ is also consistent for $\mathsf{E}X^k$.

We see that even with consistency, we cannot narrow the set of estimators; there are still infinitely many.

Example 2.4

Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ and hence

$$\mathsf{E}\left(\frac{S^2}{\sigma^2}\right)(n-1) = n-1 \quad \Rightarrow \quad \mathsf{E}\left(\frac{S^2}{\sigma^2}\right) = 1$$

For the moments of the $\chi^2(\nu)$, we can use that fact that

$$\chi^2(\nu) = \Gamma\left(\alpha = \frac{\nu}{2}, \beta = 2\right)$$

and that $\mathsf{E}(\Gamma) = \alpha\beta$ and $\mathsf{Var}(\Gamma) = \alpha\beta^2$ and our mean is ν and variance 2ν. We know that S^2 is an unbiased estimator for σ^2. Furthermore,

$$\mathsf{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1) \qquad \Rightarrow \qquad \mathsf{Var}\left(\frac{S^2}{\sigma^2}\right) = \frac{2}{n-1}$$

so the variance of the ratio tends to zero as $n \to \infty$, meaning that S^2 is also consistent. It then follows from Chebychev's inequality that

$$\mathsf{P}\left(|S^2 - \sigma^2| > \varepsilon\right) \leq \frac{\mathsf{Var}S^2}{\varepsilon^2} = \frac{2\sigma^4}{(n-1)\varepsilon^2} \xrightarrow{n\to\infty} 0$$

We could also have gotten our result directly from Kolmogorov's theorem.

Theorem 2.9

If T_n is a sequence of estimators for θ such that $\mathsf{E}_\theta[T_n] \to \theta$ and $\mathsf{Var}[T_n] \to 0$ as $n \to \infty$, then T_n is a consistent estimator of θ.

Proof

$$\mathsf{P}(|T_n - \theta| > \varepsilon) \leq \varepsilon^{-2} \underbrace{\mathsf{E}_\theta[(T_n - \theta)^2]}_{\text{MSE}} = \varepsilon^{-2}[\underbrace{\mathsf{Var}(T_n)}_{\to 0} + \underbrace{\text{Bias}_\theta(T_n^2)}_{\to 0}]$$

as $n \to \infty$, thus T_n is a consistent estimator of θ. □

Example 2.5

Say X_i are iid Bernoulli random variables with parameter p.

$$\mathsf{P}(X_i = x) = p^x(1-p)^{1-x}, n = 0, 1, 2, \ldots$$

$$\widehat{p} = \frac{1}{n}\sum X_i, \quad \mathsf{E}\widehat{p} = p, \quad \mathsf{Var}\widehat{p} = \frac{p(1-p)}{n} \to 0$$

as $n \to \infty$, thus $\widehat{p}$ is consistent.

Example 2.6

Suppose we look at the different estimators for the variance, namely

$$S^2 = \frac{1}{n-1}\sum_1^n (x_i - \overline{X})^2, \qquad S_*^2 = \frac{1}{n}\sum_1^n (x_i - \overline{x})^2$$

where X_i are iid $\mathcal{N}(\mu, \sigma^2), i = 1, \ldots, n$ random variables and $\overline{x}_n = \frac{1}{n}\sum_1^n X_i$. We know that S^2 is unbiased: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1), \mathsf{E}\left[\frac{(n-1)}{\sigma^2}S^2\right] = n-1$. Hence $\mathsf{E}S^2 = \sigma^2$ We can show this as long as the X_i's are of the same distribution and that the variance exists.

$$\mathsf{E}[S_*^2] \neq \sigma^2, \qquad \mathsf{E}[S_*^2] = \mathsf{E}\left[\frac{n-1}{n}S^2\right] = \frac{n-1}{n}\sigma^2 \xrightarrow{n\to\infty} \sigma^2.$$

On the other hand,

$$\mathsf{Var}[S_*^2] = \mathsf{Var}\left[\left(1 - \frac{1}{n}\right)S^2\right] = \left(1 - \frac{1}{n}\right)^2 \mathsf{Var}S^2 \xrightarrow{n\to\infty} 0$$

Thus

$$S_*^2 \xrightarrow{\mathsf{P}} \sigma^2$$

that is S_*^2 is a consistent estimator of σ^2.

Exercise 2.1

Suppose $X_1, \ldots, X_n$ are pairwise orthogonal (that is $\mathsf{corr}(X_i, X_j) = 0, i \neq j; i, j = 1, \ldots, n$). Let $\mathsf{Var}(X_i) = \sigma^2 < \infty, i = 1, \ldots, n$ and $\mathsf{E}(X_i) = \mu, i = 1, \ldots, n$. Then $\mathsf{E}[S^2] = \sigma^2$ where

$$S^2 = \frac{1}{n-1}\sum_1^n (X_i - \overline{X})^2$$

To do this, take $X_i - \overline{X} = X_i - \mu + \mu - \overline{X} = X_i - \mu - (\overline{X} - \mu)$. Then raise to the power of 2, and take the expectation.

$$\mathsf{Var}[\overline{X}] = \frac{\sigma^2}{n}$$

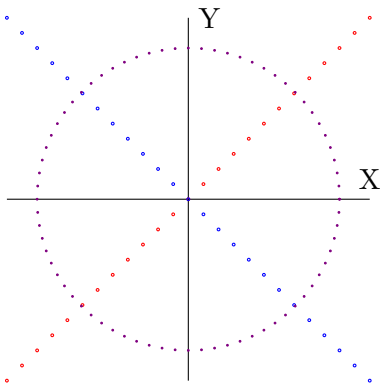


Figure 7: Correlations of -1,0,1 with $\sum X_i Y_i = 0$

Recall the Central limit theorem (CLT) and the following fact: if two variables are jointly normally distributed, then X, Y are independent if and only if they are orthogonal $X \perp Y \Leftrightarrow X \perp\!\!\!\perp Y$. Usually, we only have that independence imply orthogonality, but under normality assumption, the converse holds for some CLT. This was a signal for Doobs that the only type of relationship for Normal random variables is linear.

How did Pearson came up with the concept of correlation? As usual in modelling, we look at extreme cases. Say the data satisfies $\boldsymbol{Y} = \boldsymbol{X}$. We want to summarize the data, say we compute

$$\sum_1^n X_i Y_i$$

This is not very useful as it has not many invariants.

$$\sum_1^n (X_i - \overline{X})(Y_i - \overline{Y})$$

is better because it is translation invariant, but it is not scale invariant yet

$$\mathsf{corr}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 (Y_i - \overline{Y})^2}}$$

is equal to

$$\frac{\sum \frac{1}{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\frac{1}{n}\sum (X_i \overline{X})^2 \frac{1}{n}\sum (Y_i - \overline{Y})^2}} \to \frac{\mathsf{E}[(\boldsymbol{X} - \boldsymbol{\mu}_X)(\boldsymbol{Y} - \boldsymbol{\mu}_Y)]}{\sqrt{\mathsf{Var}(\boldsymbol{X})\mathsf{Var}(\boldsymbol{Y})}}$$

which is the Pearson correlation. A nice geometric interpretation of correlation is that it represents the cosine between $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ since it can be written as

$$\mathsf{corr}(\boldsymbol{X}, \boldsymbol{Y}) \equiv \cos\phi = \frac{\langle \boldsymbol{X}, \boldsymbol{Y} \rangle}{\|\boldsymbol{X}\|\|\boldsymbol{Y}\|}$$

Invariance property

We define the invariance property: let $\mathcal{G}$ be a group of Borel measurable (nice) functions from $\mathbb{R}^n$ to $\mathbb{R}^n$. The group operation is composition, i.e. g_1, g_2 is defined as

$$g_2 g_1(x) = g_2(g_1(x))$$

Definition 2.10 (Invariance under group)

A family of probability distributions $\{\mathsf{P}_\theta : \theta \in \Theta\}$ is said to be invariant under a group $\mathcal{G}$ if $\forall\, g \in \mathcal{G}$ and $\forall\, \theta \in \Theta$. We can find a unique $\theta^* \in \Theta$ such that the distribution of $g(X)$ is given by P_{θ^*} whenever X has the distribution P_θ.

We write $\theta' = \overline{g}\theta$, the above definition is equivalent to

$$\mathsf{P}_{\theta'} = \{g(x) \in A\} = \mathsf{P}_{\overline{g}\theta}\{x \in A\}$$

or equivalently

$$F^*_{\theta'}(X_1, \ldots, X_n) = F_{\overline{g}\theta}(X_1, \ldots, X_n)$$

where F^* is the distribution function of $g(X_1, \ldots, X_n)$.

Example 2.7

Say X is a random variable having binomial distribution $\mathcal{B}(n, p)$, for $0 < p <$

$1, G = \{e, g\}$ where $g(x) = n - x, e(x) = x$. Then $Y = g(x) \sim \mathcal{B}(n, 1-p)$,. If we look at

$$\begin{aligned}\mathsf{P}(Y = y) &= \mathsf{P}(n - x = y) = \mathsf{P}(X = n - y) \\ &= \binom{n}{n-y} p^{n-y}(1-p)^{n-(n-y)} = \binom{n}{y}(1-q)^{n-y}q^y\end{aligned}$$

where $q = 1 - p \; \forall \; y = 0, \ldots, n$ and $\overline{g}(p) = 1 - p$ and set $\overline{\mathcal{G}} = \{e, \overline{g}\}$.

Example 2.8

$$\mathcal{G} = \{g : \mathbb{R}^n \to \mathbb{R}^n | g_{a,b}(X_1, \ldots, X_n) = (ax_i + b_i, \ldots, ax_n + b)\}; \; a > 0; \; b \in \mathbb{R}$$

Let $\boldsymbol{X} = (X_1, \ldots, X_n) \sim f(X_1, \ldots, X_n)$ have distribution

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{\frac{1}{2\sigma^2}\sum_1^n (X_i - \mu)^2\right\}.$$

Then

$$\begin{aligned}g(X) &= (aX_1 + b_1, \ldots, aX_n + b) \sim f^*(X_1, \ldots, X_n) \\ &= \left(\frac{1}{\sqrt{2\pi}(a\sigma)}\right)^n \exp\left\{-\frac{1}{2a^2\sigma^2}\sum (X_i - (a_{\mu_i} + b))^2\right\}\end{aligned}$$

i.e.

$$\overline{g}(\mu, \sigma^2) = (a\mu + b, a^2\sigma^2).$$

Definition 2.11

Let $\mathcal{G}$ be a group of transformation that leaves the family $\{F_\theta : \theta \in \Theta\}$ of distribution functions invariant. An estimate T is said to be invariant under $\mathcal{G}$ if

$$T(g(X_1), \ldots, g(X_n)) = T(X_1, \ldots, X_n.)$$

Example 2.9 (Invariance)

1. Location invariant: $T(X_1 + a, \ldots, X_n + a) = T(X_1, \ldots, X_n)$.

From there, we see that $T_1(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i$ is not location invariant, but $S^2 = T_2(X_1, \ldots, X_n) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ is location invariant.

2. Scale invariant: $T(cX_1, \ldots, cX_n) = T(X_1, \ldots, X_n)$ and remark that neither the sample mean nor the sample variance are fulfilling this property. Some people think we should maybe define it as follows:

$$T(g(X_1), \ldots, g(X_n)) = g(T(X_1, \ldots, X_n))$$

but we must be careful in that case about how our group act (so that we have a well-defined $g : \mathbb{R}^n \to \mathbb{R}$). These things are a bit sour; there is no consensus on a definition. If we looked at the correlation ,

$$\mathsf{corr}(X, Y) = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \overline{X}_n)^2 \sum_{i=1}^n (Y_i - \overline{Y}_n)^2}},$$

we could indeed see that it is scale invariant.

3. Location-scale invariant : $T(aX_1 + b, \ldots, aX_n + b) = T(X_1, \ldots, X_n)$. Look at the so-called signal-to-noise ratio; it is scale invariant, but not location invariant. We could verify this: say $\tilde{X}_i = cX_i, \tilde{\overline{X}}_n = n^{-1}\sum_{i=1}^n cX_i = c\overline{X}_n$ for $c > 0$ and $\tilde{S^2} = (n-1)^{-1}\sum_{i=1}^n (cX_i - \overline{X}_n)^2 = \frac{c^2}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = c^2 S^2$. An example of location-scale invariant statistics is $\mathsf{corr}(X, Y)$. Another example that is scale invariant too, but not location invariant is the reciprocal of the signal-to-noise ratio, the so-called coefficient of variation, defined as $\overline{X}/S$.

4. Permutation invariant: $T(X_{i_1}, \ldots X_{i_n}) = T(X_1, \ldots, X_n)$,
where $\{i_1, \ldots, i_n\}$ is a permutation of $\{1, \ldots, n\}$. All the examples we have seen up to now in this example are permutation invariant. This concept gives birth to the U-statistics. An example of statistic that does not fulfil this requirement is say $T(X_1, \ldots, X_n) = X_1$, or the determinant of a matrix of observations $[X_1, \cdots X_n]$.

Section 2.3. Sufficient statistic

We build a bit on our idea of compression. Say you have data which you collected in an experiment; we want to have a summary, but not to lose information. We would like to be able to decompress this summary, but since we are working with random variables and not deterministic functions, we have to lower our expectations. Two realizations that are equally likely are the same to us: if we could come out with something similar (we cannot regenerate the actual data): $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ are $T-$similar if $\mathsf{P}_{\boldsymbol{X}|T=t,\theta} = \mathsf{P}_{\boldsymbol{Y}|T=t,\theta}$ for all possible values of t. We want $\boldsymbol{X}|T = t \overset{\mathcal{D}}{=} \boldsymbol{Y}|T = t$. We say a realization $\mathbf{x} = (x_1, \ldots, x_n)$ and a realization $\mathbf{y} = (y_1, \ldots, y_n)$ of $\boldsymbol{Y}$ are $T-$similar if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $T-$similar and $T(\mathbf{x}) = T(\mathbf{y})$, that is they fall in the same partition. This is in a sense the compressed values, which are the same.

Definition 2.12 (Sufficiency)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample from F_θ. A statistic $T = T(\boldsymbol{X})$ is sufficient for θ or for the family of distributions $\{F_\theta : \theta \in \Theta\}$ if and only if the conditional distribution of $\boldsymbol{X}$ given $T = t$ does not depend on θ (or any unknown parameter), except perhaps for an null set A, *i.e.* $\mathsf{P}_\theta\{T \in A\} = 0 \; \forall \, t$.

The idea behind this is that you cannot do simulations and generate data if it depends on an unknown parameter. What we need is $\mathsf{P}(X = x|T = t, \theta) = \mathsf{P}(X = x|T = t)$ does not depend on θ and all the relevant information was observed by t.

Example 2.10

Let $X_i \overset{\text{iid}}{\sim} \mathcal{B}(p)$. Consider $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$. The conditional probability $\mathsf{P}(X_i = x_i, i = 1, \ldots, n | \sum_{i=1}^n X_i = t) = P(A|B)$ and note that $A \subset B$, this is $\mathsf{P}(A)/\mathsf{P}(B)$. If we work out

$$\frac{\prod_{i=1}^n \mathsf{P}(X_i = x_i)}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{\prod_{i=1}^n p^{X_i}(1-p)^{1-X_i}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{p^{\Sigma X_i}(1-p)^{n-\Sigma X_i}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{1}{\binom{n}{t}}$$

If we wanted to generated the data, $\mathcal{A}_t = \{(X_1, \ldots, X_n) : \sum_{i=1}^n X_i = t\}$;

all that yield a total of t and we must now pick randomly uniformly one among the set. $\mathcal{A}_t$ collects all $n-$tuples, there are $\binom{n}{t}$ of them. Clearly, these are from the same partition - we need to show that they are the same in probability. It is best to work with *discrete*distributions for practical purposes.

Example 2.11

Let $X_i \overset{\text{iid}}{\sim} \mathcal{P}(\lambda), i = 1, 2, \ldots, n$ and $T(X_1, X_2) = X_1 + X_2$.

$$\begin{aligned}\mathsf{P}(X_1 = x_1, X_2 = x_2 | T = t) &= \frac{\mathsf{P}(X_1 = x_1, X_2 = x_2, T = t)}{\mathsf{P}(T = t)} \\ &= \frac{\mathsf{P}(X_1 = x_1, X_2 = x_2)}{\mathsf{P}(T = t)}\end{aligned}$$

if $x_1 + x_2 = t$ and zero otherwise. Substituting the formula, we get

$$\mathsf{P}(X_1 = x_1, X_2 = x_2 | T = t) = \begin{cases} \dfrac{\dfrac{\lambda^{X_1} e^{-\lambda}}{X_1!} \dfrac{\lambda^{X_2} e^{-\lambda}}{X_2!}}{\dfrac{(2\lambda)^t e^{-2\lambda}}{t!}} & \text{if } X_1 + X_2 = t \\ 0 & \text{otherwise} \end{cases}$$

If we simplify this,

$$\frac{\lambda^{X_1+X_2} e^{-2\lambda}/(X_1! + X_2!)}{2^t \lambda^t e^{-2\lambda}/t!} = \binom{t}{X_1} \left(\frac{1}{2}\right)^t$$

if $x_1 + x_2 = t$ and 0 otherwise. Here, when choosing a pair that satisfies the conditions, we have a different probability for each pair - it is binomial in this case. Let us go for a counter-example: it is easy to see that $X_1 + 2X_2$ constitute one. $\mathsf{P}(X_1 = 0, X_2 = 1 | X_1 + 2X_2 = 2)$ and using again $A \subset B$ and $\mathsf{P}(X_1 = 0, X_2 = 1)/\mathsf{P}(X_1 + 2X_2)$ and working the denominator to

$$\mathsf{P}(X_1 = 0, X_2 = 1) + \mathsf{P}(X_1 = 2, X_2 = 0) = e^{-\lambda} \lambda e^{-\lambda} + \frac{\lambda^2 e^{-\lambda}}{2!} \cdot e^{-\lambda},$$

which can be simplified further as

$$\frac{e^{-\lambda}\lambda e^{-\lambda}}{e^{-\lambda}\lambda e^{-\lambda} + \frac{\lambda^2 e^{-\lambda}}{2!}e^{-\lambda}} = \frac{\lambda}{\lambda + \frac{\lambda^2}{2}} = \frac{1}{1 + \frac{\lambda}{2}}$$

still depends on λ unknown and as such is not sufficient.

Here is an important theorem to find sufficient statistics:

Theorem 2.13 (Fisher-Neyman Factorization theorem)
Let $X_1, \ldots, X_n$ be discrete random variables with PMF $p_\theta(x_1, \ldots, x_n), \theta \in \Theta$. Then $T(X_1, \ldots, X_n)$ is sufficient for θ if and only if we can write

$$p_\theta(x_1, \ldots, x_n) = h(x_1, \ldots, x_n) \cdot g_\theta(T(x_1, \ldots, x_n)) \tag{2.16}$$

where h is a non-negative function of $x_1, \ldots, x_n$ only and doesn't depend on θ and g is a non-negative function of θ and $T(x_1, \ldots, x_n)$ solely (it depends only on observations through T). The statistic $T(X_1, \ldots, X_n)$ and parameter θ may be vectors.

Often, this boils down to taking the joint distribution and trying to factor it.

Proof

Let T be sufficient for θ, then $P(\boldsymbol{X} = \boldsymbol{x}|T = t)$ is independent of θ and we may write

$$\begin{aligned}\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) &= \begin{cases} \mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{x}) = t) & \text{if } T(\boldsymbol{x}) = t \\ 0 & \text{otherwise} \end{cases} \\ &= \underbrace{\mathsf{P}_\theta(T = t)}_{g_\theta(T(\boldsymbol{x}))} \underbrace{\mathsf{P}(\boldsymbol{X} = \boldsymbol{x}|T = t)}_{h(\boldsymbol{x})}\end{aligned}$$

Conversely, suppose (2.16) holds. Then

$$\begin{aligned}\mathsf{P}_\theta(T=t) &= \sum_{\boldsymbol{x}:T(\boldsymbol{x})=t} P_\theta(\boldsymbol{X}=\boldsymbol{x}) \\ &= \sum_{\boldsymbol{x}:T(\boldsymbol{x})=t} g_\theta(T(\boldsymbol{x}))h(\boldsymbol{x}) \\ &= g_\theta(T(\boldsymbol{x})) \sum_{\boldsymbol{x}:T(\boldsymbol{x})=t} h(\boldsymbol{x})\end{aligned}$$

What we want is to find $\mathsf{P}(\boldsymbol{X}=\boldsymbol{x}|T=t)$ and need to divide by $\mathsf{P}(T=t)$. Now suppose that $\mathsf{P}_\theta(T=t)>0$ for some t. Then

$$\begin{aligned}\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x}^*|T=t) &= \frac{P_\theta(\boldsymbol{X}=\boldsymbol{x}^*|T(\boldsymbol{x})=t)}{P_\theta(T(\boldsymbol{x})=t)} \\ \text{if compatibility holds } &= \begin{cases} \dfrac{\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x}^*)}{P_\theta(T(\boldsymbol{x})=t)} & \text{if } T(\boldsymbol{x})=t \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \dfrac{g_\theta(t)h(\boldsymbol{x}^*)}{g_\theta(t)\sum_{\boldsymbol{x}:T(\boldsymbol{x})=t} h(\boldsymbol{x})} & \text{if } T(\boldsymbol{x}^*)=t \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \dfrac{h(\boldsymbol{x}^*)}{\sum_{\boldsymbol{x}:T(\boldsymbol{x})=t} h(\boldsymbol{x})} & \text{if } T(\boldsymbol{x}^*)=t \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

Hence T is sufficient. □

Remark

- The above theorem holds for quite general families of distributions. The absolutely continuous case can be proved similarly. We can have cases such as
$$f_X(x) = \frac{1}{2}\delta_0 + \frac{1}{2}\mathcal{U}(0,1);$$
the above example[4] is continuous, not dominated by Lebesgue measure and has mass at zero.

[4] δ_0 stands for the Dirac delta function.

- The above theorem holds true if $\boldsymbol{\theta}$ is a vector of parameters and T is a random vector. In this case we say T is jointly sufficient for $\boldsymbol{\theta}$.[5] We will see with UMVUE that we have sometimes partial sufficiency. Note that even if $\boldsymbol{\theta}$ is a scalar, T may be a vector if $\boldsymbol{\theta}$ and T are vectors of the same dimension and T is jointly sufficient for $\boldsymbol{\theta}$, it does not follow that the j^{th} component of T is sufficient for the j^{th} component of $\boldsymbol{\theta}$.

- If T is sufficient for θ, then any injective function of T is also sufficient for θ The proof is left to the reader as an exercise.

  It does **not** follow that every function of T is sufficient for θ. One can think of as a trivial example of the constant function. Also, $\overline{X}$ is sufficient for μ if $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$, but $\overline{X}^2$ is not.

Example 2.12

$$f_{X_1,\ldots,X_n|\overline{X}_n}(x_1,\ldots,x_n|\overline{X}_n = t) = \frac{f_{X_1,\ldots,X_n|\overline{X}_n}(x_1,\ldots,x_n,t)}{f_{\overline{X}}(t)}$$

where defined and since $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$, we thus have that $\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{1}{n}\right)$ and

$$\begin{aligned} f_{\boldsymbol{X}}(x_1,\ldots,x_n) &= \prod_{i=1}^{n} f_{X_i}(x_i) \\ &= \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}}{\frac{n}{\sqrt{2\pi}}\exp\left\{-\frac{n}{2}(\overline{X}-\mu)^2\right\}} \\ &= \left[\frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n}{\frac{\sqrt{n}}{\sqrt{2\pi}}}\right] \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\left((x_i-\mu)^2-(\overline{X}-\mu)^2\right)\right]^2\right\} \end{aligned}$$

[5] Be careful with the notation, for example, we have for Normal that $(\overline{X}, S^2)$ is jointly sufficient, but $\overline{X}$ is not sufficient for μ and S^2 is not sufficient for σ^2 for the variance if (μ, σ^2) are unknown.

looking at

$$\sum_{i=1}^{n}(x_i-\mu)^2=\sum_{i=1}^{n}x_i^2-2\mu\sum_{i=1}^{n}x_i+n\mu^2 \tag{2.17}$$

$$\sum_{i=1}^{n}(\overline{X}_n-\mu)=n\overline{X}_n^2-2n\mu\overline{X}_n+n\mu^2 \tag{2.18}$$

and the difference of 2.17 and 2.18 is $\sum_{i=1}^{n}x_i-n\overline{X}_n^2$. Set $Y=\overline{X}_n^2$, we can find directly

$$\begin{aligned}\mathsf{P}(Y\leq y)=\mathsf{P}(\overline{X}_n\leq y)&=\mathsf{P}\left(\left|\overline{X}_n\leq\sqrt{y}\right|\right)\\&=\mathsf{P}\left(-\sqrt{y}\leq\overline{X}_n\leq\sqrt{y}\right)\\&=F_{\overline{X}_n}(\sqrt{y})-F_{\overline{X}_n}(-\sqrt{y})\end{aligned}$$

and see that $\bar{S}^2$ is not sufficient; differentiate $f_Y(y)$ and plug back into the ratio. We see

$$f_{X_1,\ldots,X_n|\overline{X}_n^2}(x_1,\ldots,x_n|t)$$

is not independent of μ. If T is sufficient for $\{F_\theta:\theta\in\Theta\}$ then T is sufficient for $\{F_\theta:\theta\in\omega\}$ where $\omega\subset\Theta$.

Example 2.13

Let X_i be iid $\mathcal{U}(0,\theta)$.

$$f_\theta(x)=\begin{cases}\frac{1}{\theta} & 0\leq x_i,\leq\theta, i=1,\ldots,n\\ 0 & \text{otherwise}\end{cases}$$

Then

$$\begin{aligned}f_\theta(X_1,\ldots,X_n)\overset{\perp\!\!\!\perp}{=}\prod_{i=1}^{n}f_\theta(x_i)&=\begin{cases}\dfrac{1}{\theta^n} & 0\leq x_i\leq\theta\ \forall\ i\\ 0 & \text{otherwise}\end{cases}\\&=\frac{\prod_{i=1}^{n}I_{[0,\theta]}(x_i)}{\theta^n}\end{aligned}$$

We want to this by factorization. If we write this

$$= \frac{I_{[0,\theta]}(\bigwedge_{i=1}^n x_i) I_{[0,\theta]}(\bigvee_{i=1}^n x_i)}{\theta^n}$$

where $\bigwedge_{i=1}^n x_i = \min_{1\leq i\leq n} x_i$ and $\bigvee_{i=1}^n x_i = \max_{1\leq i\leq n} x_i$. But maybe we can rewrite this only in terms of the maximum as

$$= \frac{I_{[0,\vee x_i]}(\bigwedge_{i=1}^n x_i) I_{[\wedge x_i,\theta]}(\bigvee_{i=1}^n x_i)}{\theta^n}$$

which is of the form $h(x_1,\ldots,x_n)g_\theta(T(x_1,\ldots,x_n))$ and $g_\theta(T(x_1,\ldots,x_n))$ is called the kernel. We have $T(x_1,\ldots,x_n) = \bigwedge_{i=1}^n x_i$. Here, we note that $h(x_1,\ldots,x_n) = I_{[0,\theta]}(\min x_i)$ and $g_\theta(T(x_1,\ldots,x_n)) = \frac{I_{[0,\theta]}(\max x_i)}{\theta^n}$. Why do we care about the Uniform distribution? It happens that it is really useful in simulation. It is also useful for inverse probability integral transform. For example, one can generate $X \sim F$, $Y = F(X) \sim \mathcal{U}(0,1)$ and thus generate easily uniform and compute the inverse distribution by finding $F^{-1}(u)$.

Example 2.14

Let $X_i \overset{\text{iid}}{\sim}$ discrete $\mathcal{U}\{1,\ldots,N\}$, $i \in \{1,\ldots,n\}$ where $N \in \mathbb{N}$ is unknown parameter.

$$\mathsf{P}_N(X_i = K) = \begin{cases} \frac{1}{N} & \text{if } k = 1,2,\ldots,N \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \mathsf{P}_N(X_i = k_i, i = 1,\ldots,n) &= \frac{1}{N^n} \text{ if } 1 \leq k_i \leq N \\ &= \frac{1}{N^n} I_{k_{(1)}} I_{k_{(n)}} \\ &= \frac{\mathcal{E}\left(N - \bigvee_{i=1}^n k_i\right) \mathcal{E}\left(\bigwedge_{i=1}^n k_i - 1\right)}{N^n} \end{aligned}$$

We can define

$$g_N\left(\bigvee_{i=1}^n k_i\right) = \frac{\mathcal{E}\left(N - \bigvee_{i=1}^n k_i\right)}{N^n}$$

and

$$h(k_1,\ldots,k_n) = \mathcal{E}\left(\bigwedge_{i=1}^n k_i - 1\right)$$

using the Fisher-Neyman theorem, $\bigvee_{i=1}^{n} X_i$ is sufficient for the family $\{\mathsf{P}_N : N \in \mathbb{N}\}$, note that $T = \max X_i$

$$\begin{aligned} \mathsf{P}_N(T = t) &= \mathsf{P}_N(T \leq t) - \mathsf{P}_N(T \leq t-1) \\ &= [\mathsf{P}(X_i \leq t)]^n - [\mathsf{P}(X_i \leq t-1)]^n \\ \text{where } T = \bigvee_{i=1}^{n} X_i \quad &= \left[\frac{t}{N}\right]^n - \left[\frac{t-1}{N}\right]^n \end{aligned}$$

noting that the max is less than t if and only if all observations are less than of equal since they are independently and identically distributed. It is easy to find the joint distribution of order statistic, (although messier in the discrete case). One can always the above trick for maximums and minimums. Noting that $\{\boldsymbol{X} = \mathbf{k}\} \subseteq \{T = t\}$, we get

$$\begin{aligned} \mathsf{P}_N(\boldsymbol{X} = \mathbf{k} | T = t) &= \begin{cases} \dfrac{\mathsf{P}_N(\boldsymbol{X} = \mathbf{k})}{\mathsf{P}_N(T = t)} & \text{if } T(k) = t \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{\dfrac{1}{N^n}}{\dfrac{1}{N^n}[t^n - (t-1)^n]} \text{ if } T(k) = t \end{aligned}$$

and zero elsewhere, so it does not depend on our unknown parameter N. In the above formula, we should have carefully added indicator functions, but they would be cancelling out so not doing so was harmless. We do one more example on the Fisher-Neyman Factorization theorem, this time with a Normal distribution.

Example 2.15

Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. In this case,

$$\begin{aligned} f_{\mu,\sigma^2}(\boldsymbol{x}) &= \prod_{i=1}^{n} f_{\mu,\sigma^2}(\boldsymbol{x}) \\ &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(u_i-\mu)^2}{2\sigma^2}} \right] \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right) \exp\left\{ -\frac{1}{2\sigma^2} \left[\sum_{n=1}^{\infty} x_i^2 - 2\mu \sum_{n=1}^{\infty} x_i + \mu^2 \right] \right\} \end{aligned}$$

and we can take $h(\boldsymbol{x}) = 1$ and $T(\boldsymbol{x}) = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$ are the jointly sufficient statistics. Since there exists a one-to-one correspondence between the vectors

$$\left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2 \right) \Leftrightarrow \left(\bar{x}_n, S^2 \right),$$

by our previous remark $(\bar{x}_n, S^2)$ is also a sufficient statistic.

We will go with a few more remarks about sufficiency and examples, than go to completeness (a notion due to Lehmann-Scheffé) to get uniformly minimum variance unbiased estimator (UMVUE). If we consider our mean squared error, where we look at the Euclidean distance between our estimator and θ. In general, we cannot minimize this[6];

$$\mathsf{E}[T(\boldsymbol{x}) - \theta]^2 = \ \mathrm{MSE}_\theta(T) = \mathsf{Var}T(X) + (\mathrm{Bias}_\theta T(\boldsymbol{x}))^2 .$$

We will confine our attention for now to all unbiased estimators. If we find an estimator $\mathsf{E}_\theta\left(\widehat{\theta}(T)\right) = \theta$, hence we get another unbiased estimator by finding the conditional. Furthermore, as we will see with Rao-Blackwell theorem, $\mathsf{Var}(\widehat{\theta}(T)) \leq \mathsf{Var}(S(\boldsymbol{X}))$. The question now is how many time must we do this? It appears that if the family of estimators is complete, we only need to do this once.

[6] This leads to paradox in $\infty-$dimensions, see Galton

Fact 2.14 (Stein's paradox)

As a side remark, here is an interesting paradox, due to Stein. Let us introduce the notion of admissibility. If we have two estimators, T_1 and T_2, we say that T_1 is inadmissible if $\text{MSE}_\theta(T_2) \leq \text{MSE}_\theta(T_1)\ \forall\ \theta$ with strict inequality holding for at least one. Suppose we have a multivariate $\boldsymbol{\theta}$. The paradox arises for dimension greater or equal to 3; say $\boldsymbol{X}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, I_n)$. Then $\overline{\boldsymbol{X}}_n$ is not admissible for 3 or more dimensions. This suggests that naive statistics can be misleadings. We can do better estimates (in regression), namely shrinkage estimators.

Remark

As mentioned before, note that $\overline{X}$ is not sufficient for μ if σ^2 is unknown, nor is S^2 for σ^2. This means we cannot regenerate data from this distribution (we say in such case that the statistic is partially sufficient as $f(x_1, \ldots, x_n|\bar{x})$ will not depend on μ, but still on σ^2). However, $\overline{X}$ is sufficient for μ if σ^2 is *known* and so is $\sum_{i=1}^n (X_i - \mu)^2$ for σ^2 if μ is known.

Example 2.16

Let $X_i \overset{\text{iid}}{\sim} \mathcal{U}(-\frac{\theta}{2}, \frac{\theta}{2})$ for $\theta > 0, i = 1, \ldots, n$, $\boldsymbol{x} = (x_1, \ldots, x_n)$ and

$$f_\theta(\boldsymbol{x}) = \begin{cases} \dfrac{1}{\theta^n} & \text{if } |x_i| \leq \frac{\theta}{2} \\ 0 & \text{otherwise} \end{cases}.$$

We can rewrite this as

$$= \frac{1}{\theta^n} \mathcal{E}\left(\bigwedge_{i=1}^n x_i + \frac{\theta}{2}\right) \mathcal{E}\left(\frac{\theta}{2} - \bigvee_{i=1}^n x_i\right).$$

Then, using Fisher-Neyman factorization theorem, $(\bigwedge_{i=1}^n X_i, \bigvee_{i=1}^n X_i)$ is a sufficient statistic for θ. Remark that our statistic can have dimension higher than that of the parameter we are trying to estimate.

Another fundamental concept that deals with uniqueness is

Section 2.4. Completeness

Definition 2.15 (Completeness)

Let $\{f_\theta(x); \theta \in \Theta\}$ be a family of PDF's (or PMF's). We say that this family is complete if for any measurable function g (almost continuous), $\mathsf{E}_\theta[g(X)] = 0 \; \forall \; \theta \in \Theta$ implies that $\mathsf{P}\{g(X) = 0\} = 1 \; \forall \; \theta \in \Theta$.

Definition 2.16 (Complete statistic)

A statistic $T(X)$ is said to be complete if the family of distributions of T is complete.

Example 2.17

Let $X_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. Then $T = \sum_{i=1}^n X_i$ is a sufficient statistic for p. Look at the likelihood

$$\begin{aligned}
\mathsf{P}_p(X_i = x_i, i = 1, \ldots, n) &= \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \\
&= p^{\Sigma x_i}(1-p)^{n-\Sigma x_i} \\
&= (1-p)^n \left[\frac{p}{1-p}\right]^{\Sigma x_i}
\end{aligned}$$

where the term between brackets is the kernel; namely the part for which we cannot separate observables from unobservables. We now claim that T is also complete; using the method of moment generating functions, if $T = \sum_{i=1}^n X_i \sim \mathcal{B}(n, p)$, then $\mathsf{E}_p\Big[g(T)\Big] = 0 \; \forall \; p \in (0,1) \Rightarrow g = 0$. We see here what is meant by uniqueness; look at the definition of

$$\mathsf{E}_p\Big[g(T)\Big] = \sum_{t=0}^n g(T)\mathsf{P}(T = t) = \sum_{t=0}^n g(t)\binom{n}{t}\left(\frac{p}{1-p}\right)^t = 0 \; \forall \; p \in (0,1).$$

We have $n+1$ unknowns with infinitely many equations. Suppose $p \in \{p_1, \ldots, p_{n+1}\}$ for some fixed values of p_i. We can certainly plug all these values and get a system of equations. To get zero, we need the matrix to be invertible; this is what is meant by uniqueness. Factoring a term in our last

equation and letting $\frac{p}{1-p} = \theta$, we have

$$(1-p)^n \sum_{t=0}^{n} g(t)\binom{n}{t}\left(\frac{p}{1-p}\right)^t = 0 \;\forall\; p \in (0,1) \Rightarrow \sum_{t=0}^{n} a_t \theta^t = 0 \;\forall\; \theta > 0$$

and using the Fundamental theorem of algebra, we know that this can have at most n solutions, but it happens for uncountably many values of n, so a_t must be zero and $g(t)\binom{n}{t} = 0$, for $t = 0, \ldots, n \Rightarrow g(t) = 0 \;\forall\; t$ and g is the zero function. As an immediate consequence of Rao-Blackwell theorem, we have

$$\widehat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{T}{n} \tag{2.19}$$

is UMVUE since $\mathsf{E}(g(T)|T) = g(T)$. If our unbiased estimator is already a function of the sufficient statistic, it is UMVUE.

Most of the time, completeness boils down to Fourier transform, Laplace transform, the Fundamental theorem of algebra or otherwise some trick.

Example 2.18

Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\theta)$ for $\theta > 0$. Then $\{\mathcal{N}(0,\theta) : \theta > 0\}$ is not complete since $\mathsf{E}_t(X) = 0 \;\forall\; \theta > 0$ and $g(X) = X$ is not identically zero. However, note that $T(X) = X^2$ is complete. If $X \sim \mathcal{N}(0,\theta)$, then $X^2/\theta \sim \chi^2(1)$ or $X^2 \sim \theta\chi^2(1)$ and

$$f_T(t) = \frac{e^{-\frac{t}{2\theta}}}{\sqrt{2\pi t\theta}} \quad t > 0$$

and zero otherwise. Rewriting the above, we get

$$\mathsf{E}_t[g(T)] = \frac{1}{\sqrt{2\pi\theta}} \int_0^\infty g(t) t^{-\frac{1}{2}} e^{-\frac{t}{2\theta}} dt = 0 \quad \forall\; \theta > 0$$

so the integral must vanish and be equal to zero. Using uniqueness of the Laplace transform, we have $g(t)t^{-\frac{1}{2}} = 0 \;\forall\; t > 0$ implying that $g(t) = 0$ almost surely with respect to Lebesgue measure.

Example 2.19

Let $X \sim \mathcal{U}(0,\theta)$ for $\theta > 0$. Then X is complete since

$$\mathsf{E}_t\Big[g(X)\Big] = \int_0^\theta \frac{1}{\theta} g(x)dx = 0 \quad \forall\, \theta > 0 \Rightarrow \frac{1}{\theta}\int_0^\theta g(x)dx = 0$$

is zero if the integral is zero. Differentiating both sides with respect to θ (since we assume g is continuous). We can also use Lusin's theorem in this case; if we can do this for any subinterval, we can make a $\sigma-$field on these interval and generate the Borel $\sigma-$field.

Now let $X_i \overset{\text{iid}}{\sim} \mathcal{U}(0,\theta)$, for $i = 1,\ldots,n$. Then $M_n = \bigvee_{i=1}^n X_i$ is a sufficient and complete statistic since $f_{M_n}(x;\theta)$. Before we look more closely at this example, we need to investigate the

Proposition 2.17 (Distribution of maximum of a random variable)

Looking at

$$\begin{aligned} F_{M_n}(x) &= \mathsf{P}(M_n \le x) = \mathsf{P}(X_i \le x_i, i = 1,\ldots,n) \\ &\overset{\perp\!\!\!\perp}{=} \prod_{i=1}^n \mathsf{P}(X_i \le x) = \Big[F_X(x)\Big]^n \end{aligned}$$

and differentiating both sides

$$\begin{aligned} f_{M_n}(x) = \frac{df_{M_n}(x)}{dx} &= n f_X(x)\Big[F_X(x)\Big]^{n-1} \\ &= n\left(\frac{1}{\theta}\right)\left(\frac{x}{\theta}\right)^{n-1} \quad \text{for } 0 < x < \theta \text{ and} \\ f_{M_n}(x,\theta) &= \begin{cases} n\theta^{-n}x^{n-1} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Example 2.20

We have that

$$\begin{aligned} \mathsf{E}_t\Big[g(M_n)\Big] = 0 \Rightarrow \int_0^\theta g(x) n\theta^{-n}x^{n-1}dx = 0 \;\forall\, \theta > 0 \\ \Rightarrow \int_0^\theta x^{n-1}g(x)dx = 0 \quad \forall\, \theta > 0. \end{aligned}$$

Assuming that g is continuous, $\theta^{n-1}g(\theta) = 0 \quad \forall\, \theta > 0$ and this entails that $g(\theta) = 0 \;\forall\; \theta > 0$. Recall that we had proved earlier that M_n was sufficient for θ.

Example 2.21

As always, consider $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \theta^2)$ for $i = 1, \ldots, n$ and $\theta > 0$. Then $T = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is sufficient for θ. T is however not complete since $\mathsf{E}_\theta\Big[g(T)\Big] = 0 \;\forall\; \theta$ if

$$g(t) = 2\left(\sum_{i=1}^n X_i\right)^2 - (n+1)\sum_{i=1}^n X_i^2$$

is not identically zero.

Example 2.22

Let $X_i \overset{\text{iid}}{\sim}$ discrete $\mathcal{U}\{1, \ldots, N\}$. $M_n = \bigvee_{i=1}^n X_i$ was shown to be sufficient. We also want to show it is complete. In this case,

$$P_{M_n}(x) = \frac{x^n}{N^n} - \frac{(x-1)^n}{N^n}, \quad x = 1, \ldots, N$$

and for the expectation,

$$\begin{aligned}
\mathsf{E}_N\big[g(M_n)\big] &= \sum_{x=1}^N g(x)\left[\frac{x^n}{N^n} - \frac{(x-1)^n}{N^n}\right] = 0 \\
&= \frac{1}{N^n}\underbrace{\sum_{x=1}^N g(x)\left[x^n - (x-1)^n\right]}_{\phi(N)} = 0 \quad \forall\, N \in \mathbb{N} \qquad (2.20)
\end{aligned}$$

We can look at differencing $\phi(N)$ for $\frac{f(N+1)-f(N)}{(N+1)-N}$ and

$$0 = \phi(N) = \left[N^n - (N-1)^n\right]$$

and so $g(N) = 0 \;\forall\; N \geq 2$ and $g(1) = 0$ by 2.20. Therefore, $g(N) = 0 \;\forall\; N \geq 1$ and as such M_n is also complete.

It is easy to see that if we exclude the value $N = n_0$ from the parameter space, $\mathbb{N} = \{1, 2, \ldots\}$, the family $\{f_{M_n,N} : N \in \mathbb{N}\}$ is not complete any longer. This is so since

$$E_N[g(M_n)] = 0$$

if

$$g(x) = \begin{cases} 0 & \text{if } x \neq n_0, n_0 + 1 \\ c & \text{if } x = n_0 \\ -c & \text{if } x = n_0 + 1 \end{cases}$$

for some $c > 0$.

To see this, write the definition of expected value. Suppose that $n = 1$ (we take only one sample) and

$$\begin{aligned} \mathsf{E}_N[g(M_n)] &= \sum_{x=1}^{N} g(x)\frac{1}{n}, && \forall N \\ &= \frac{1}{N}[\underbrace{g(n_0)}_{c} + \underbrace{g(n_0+1)}_{-c}], && \forall\, N \in \mathbb{N} \setminus \{n_0\} \end{aligned}$$

Remark

The above observation shows that the exclusion of even one member from the family $\{P_N : N \in \mathbb{N}\}$ destroys completeness. Recall that if a statistic is sufficient for a class of distributions, it is sufficient for any subclass of those distributions. Completeness works in the opposite way.

Exponential family of distributions

The exponential family of distributions includes distributions such as Binomial, Bernoulli, Gamma, Poisson, Normal, χ^2, Geometric among many others. The class arises from maximal entropy in statistical mechanics, which is an extension of Laplace's insufficient reasoning principle. We could classify entropy as the negative of information. The exponential family is used in generalized linear models - although another absent class here is the mixture family. All exponential family distributions are unimodal.

Let $\Theta \subset \mathbb{R}^k$ and $\{x : f_{\boldsymbol{\theta}}(x) > 0\}$ does not depend on $\boldsymbol{\theta}$.

Definition 2.18

If there exists real valued nice functions $Q_1, \ldots, Q_k$ and D defined on Θ, and $T_1, \ldots, T_k$ and S on $\mathbb{R}^m$ such that

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp\left\{\sum_{i=1}^{k} Q_i(\boldsymbol{\theta}) T_i(\boldsymbol{x}) + D(\boldsymbol{\theta}) + S(\boldsymbol{x})\right\}; \tag{2.21}$$

we say that the family $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is a k-parameter exponential family (where $\boldsymbol{x} = (x_1, \ldots, x_m)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$).

Example 2.23

Let $X \sim \mathcal{B}(n, p), 0 < p < 1$; then

$$\begin{aligned}
\mathsf{P}(X = x) &= \binom{n}{x} p^x (1-p)^{n-x} \\
&= \binom{n}{x} (1-p)^n \left(\frac{p}{1-p}\right)^x \\
&= \exp\Bigg\{ \underbrace{x}_{T(x)} \underbrace{\log \frac{p}{1-p}}_{Q(p)} + \underbrace{n \log(1-p)}_{D(p)} + \underbrace{\log \binom{n}{x}}_{S(x)} \Bigg\},
\end{aligned}$$

where $p \in \Theta = (0, 1)$ and the support does not depend on an unknown parameter.

Example 2.24

$X \sim \mathcal{U}(0, \theta), \theta > 0$

$$f_\theta(x) = \begin{cases} \dfrac{1}{\theta} & \text{if } 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

This is not exponential family. Another example is again binomial, but with n unknown. We can also think of the Negative binomial without knowing the number of successes.

Example 2.25

$X \sim \mathcal{N}(\mu, \sigma^2)$, where $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in R^+\}$

$$
\begin{aligned}
f_{\mu,\sigma^2}(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R} \\
&= \exp\left\{ -\frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2} x - \frac{1}{2}\left[\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2) \right] \right\}
\end{aligned}
$$

Thus

$$
\begin{aligned}
&Q(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}, \quad Q_2(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2} \quad D(\theta) = -\frac{1}{2}\left[\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2) \right] \\
&T_1(x) = x^2, \qquad T_2(x) = x, \qquad S(x) = 0, \;\; k = 2, \; m = 1
\end{aligned}
$$

Theorem 2.19 (Complete sufficient statistics for exponential family)
Let $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ be a k-parameter exponential family given by

$$
f_{\boldsymbol{\theta}}(x) = \exp\left\{ \sum_{i=1}^{k} Q_i(\boldsymbol{\theta}) T_i(\boldsymbol{x}) + D(\boldsymbol{\theta}) + S(\boldsymbol{x}) \right\} \tag{2.22}
$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta$ an interval in $\mathbb{R}^k$ (namely an hypercube). $T_1, \ldots, T_k$ and S are defined on $\mathbb{R}^n$, $T = (T_1, \ldots, T_k)$, and $x = (x_1, \ldots, x_n)$, for $k \leq n$. Let $\mathbf{Q} = (Q_1, \ldots, Q_k)$ and suppose that the range of $\mathbf{Q}$ contains an open set in $\mathbb{R}^k$. Then $T = (T_1(\boldsymbol{x}), \ldots, T_k(\boldsymbol{x})$ is a complete sufficient statistic.

PROOF Let $k = 1$ and X be a discrete random variable. To simplify life suppose wlog that $Q(\theta) = \theta$ and that Θ contains an interval (α, β)

$$
\begin{aligned}
\mathsf{E}_\theta[g(T(x))] &= \sum_t g(t) \mathsf{P}_\theta(T(x) = t) \\
&= \sum_t g(t) \exp\left\{ \theta t + D(\theta) + S^*(t) \right\} \qquad (2.23) \\
&= 0, \qquad \forall\, \theta \in \Theta.
\end{aligned}
$$

Now

$$\begin{aligned}\mathsf{P}_\theta(T=t) &= \sum_{x:T(x)=t} \mathsf{P}_\theta(X=x) \\ &= \sum_{x:T(x)=t} \exp\left\{\sum_{i=1}^{k} Q_i(\theta)T_i(\boldsymbol{x}) + D(\bar{Q}) + S(\boldsymbol{x})\right\} \\ &= \sum_{x:T(x)=t} \exp\left\{\theta t + D(\theta) + S(x)\right\} = \exp\{\theta t + D(\theta)\} \sum_{x:T(x)=t} \exp\{S(x)\}\end{aligned}$$

where we defined

$$S^*(t) = \log\left[\sum_{x:T(x)=t} \{S(x)\}\right]$$

It tells us that $T \sim \exp\{\theta t + D(\theta) + S^*(t)\}$ and follows the same path for $k > 1$. Define

$$x^+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad x^- = \begin{cases} -x & \text{if } x < 0 \\ 0 & \text{otherwise} \end{cases}$$

Then

$$g(t) = g^+(t) - g^-(t), \tag{2.24}$$

where both g^+ and g^- are non-negative functions.

Using (2.23) and (2.24), we have

$$\sum_t g^+ e^{\theta t + S^*(t)} = \sum_t g^-(t) e^{\theta t + S^*(t)}, \quad \forall\ \theta \in \Theta \tag{2.25}$$

and $e^{D(\theta)}$ gets cancelled from both sides by factorizing. Let $\theta_0 \in (a, b)$ be fixed and write

$$p^{\pm}(t) = \frac{g^{\pm}(t) e^{\theta_0 t + S^*(t)}}{\sum_t g^{\pm}(t) e^{\theta_0 t + S^*(t)}}; \tag{2.26}$$

we can make this into a probability distribution and use Laplace transform. Then both p^+ and p^- are PMF and is follows from (2.25) that [7]

$$\sum_t e^{\delta t} p^+(t) = \sum_t e^{\delta t} p^-(t), \forall \delta \in (\alpha - \theta_0, \beta - \theta_0) \tag{2.27}$$

By the uniqueness of MGF, (2.27) implies that $p^+(t) = p^-(t)$ for all t.

This in turn implies that $g^+(t) = g^-(t)$ and hence $g \equiv 0$. Using Fisher-Neyman factorization theorem T is also sufficient. Thus T is a complete sufficient statistic. □

Definition 2.20 (Unbiased estimators)

Recall that $T : \mathbb{R}^n \to \Theta$ is called unbiased if $\mathsf{E}_\theta(T(X)) = \theta$ for all $\theta \in \Theta$. A function $\psi(\theta)$ is called *estimable function* if there exists a T such that

$$\mathsf{E}_\theta[T(\boldsymbol{X})] = \psi(\theta) \qquad \forall\, \theta \in \Theta.$$

Example 2.26

Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \ldots, n$ and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \quad \text{with} \quad \mathsf{E}(S^2) = \sigma^2$$

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{with} \quad \mathsf{E}(\overline{X}_n) = \mu;$$

thus μ and σ^2 are estimable functions.

Example 2.27

Let $X \sim \mathcal{P}(\lambda)$ and take $\psi(\lambda) = e^{-3\lambda}$ and $T(X) = (-2)^x$. Then, in this case

$$\mathsf{E}_\lambda[T(X)] = \sum_{x=0}^{\infty} (-2)^x \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(-2\lambda)^x}{x!} = e^{-\lambda} e^{-2\lambda} = e^{-3\lambda}.$$

[7] Plug (2.26) in (2.27) and use the fact that the denominators are equal by getting back here. We now have the Laplace transform; work it for all θ.

This is nonsense, since we can use a negative function to approximate something positive.

Section 2.5. Minimum variance unbiased estimators

Definition 2.21

Let $\mathcal{U}$ be the set of all unbiased estimators T of $\theta \in \Theta$ such that $\mathsf{E}_\theta[T^2(X)] < \infty \ \forall \ \theta \in \Theta$. An estimator $T_0 \in \mathcal{U}$ is called uniformly minimum variance unbiased estimator (UMVUE) of θ if $\mathsf{E}_\theta(T_0 - \theta)^2 \leq \mathsf{E}_\theta(T - \theta)^2 \ \forall \ \theta \in \Theta$ or

$$\mathsf{Var}_\theta(T_0) \leq \mathsf{Var}_\theta(T) \quad \forall \ T \in \mathcal{U}.$$

We now tackle existence and uniqueness of UMVUE.The proof provided for the next theorem is analytic, however you can come out with a geometric proof using projections and Hilbert spaces.

Theorem 2.22

Let $\mathcal{U}$ be the class of all unbiased estimators T of a parameter $\theta \in \Theta$ with $\mathsf{E}_\theta T^2 < \infty \ \forall \ \theta \in \Theta$ and suppose $\mathcal{U}$ is non-empty (*i.e.* θ is estimable). Let $\mathcal{U}_0$ be the set of all unbiased estimators ν of 0, that is

$$\nu_0 = \{\nu : \mathsf{E}_\theta \nu = 0, \mathsf{E}_\theta \nu^2 < \infty \ \forall \ t\}. \tag{2.28}$$

Then $T_0 \in \mathcal{U}$ is UMVUE[8] if and only if

$$\mathsf{E}_\theta \nu T_0 = 0 \text{ for all } \theta, \text{ for all } \nu \in \mathcal{U}_0.$$

Proof The conditions of the theorem guarantee the existence of $\mathsf{E}_\theta(\nu T_0) \leq [\mathsf{E}_\theta \nu^2]^{\frac{1}{2}} [\mathsf{E}_\theta T_0^2]^{\frac{1}{2}}$ (using Cauchy-Schwartz, since both second moment exist, the inner product is finite). for all θ and $\nu \in \mathcal{U}_0$. Suppose that $T_0 \in \mathcal{U}$ is a UMVUE and $\mathsf{E}_{\theta_0}(\nu_0 T_0) \neq 0$ for some θ_0 and some $\nu_0 \in \mathcal{U}_0$. Then $T_0 + \lambda v_0 \in \mathcal{U}$ for all real λ (since $\mathsf{E}T_0 + \lambda \mathsf{E}\nu_0 = \mathsf{E}T_0 \in \mathcal{U}$). If $\mathsf{E}_{\theta_0} \nu_0^2 = 0$, then

[8] Note that we don't require sufficiency or completeness here. The proof is however not constructive.

$\mathsf{E}_t(\nu_0 T_0) = 0$ (using the above inequality) must hold since [9] $\mathsf{P}_{\theta_0}\{\nu_0 = 0\} = 1$. Let $\mathsf{E}_{\theta_0}\nu_0^2 = 0$. Choose $\lambda_0 = -\frac{\mathsf{E}_{\theta_0}(\nu_0 T_0)}{\mathsf{E}_{\theta_0}\nu_0^2}$. Then $\mathsf{E}_{\theta_0}(T_0 + \lambda_0\nu_0)^2$ is equal to

$$\mathsf{E}_{\theta_0} T_0^2 - \frac{[\mathsf{E}_{\theta_0}\nu_0 T_0]^2}{\mathsf{E}_{\theta_0}\nu_0^2} < \mathsf{E}_{\theta_0} T_0^2 \tag{2.29}$$

Since $T_0 + \lambda_0\nu_0 \in \mathcal{U}$ and $T \in \mathcal{U}$, it follows from (2.29) that $\mathsf{Var}_{\theta_0}(T_0 + \lambda\nu_0) < \mathsf{Var}_{\theta_0}(T_0)$, which is a contradiction.

Conversely, let $\mathsf{E}_\theta(\nu T_0) = 0$ for all θ and $\nu \in \mathcal{U}_0$ for some $T_0 \in \mathcal{U}$, all $\theta \in \Theta$ and all $\nu \in \mathcal{U}_0$ and let $T \in \mathcal{U}$. Then $T_0 - T \in \mathcal{U}_0$ and for every θ, we have $\mathsf{E}_\theta\{T_0(T_0 - T)\} = 0$ since T_0 is orthogonal to $T_0 - T$ (since $\mathsf{E}_\theta(\nu T_0) = 0$). We then have

$$\mathsf{E}_\theta T_0^2 = \mathsf{E}_\theta(TT_0) \leq \left[\mathsf{E}_\theta T_0^2\right]^{\frac{1}{2}} \left[\mathsf{E}_\theta T^2\right]^{\frac{1}{2}}$$

by Cauchy-Schwartz inequality. If $\mathsf{E}_\theta T_0^2 = 0$, then $\mathsf{P}(T_0 = 0) = 1$ and there is nothing to prove. Otherwise,

$$\left(\mathsf{E}_\theta T_0^2\right)^{\frac{1}{2}} \leq (\mathsf{E}_\theta T^2)^{\frac{1}{2}}$$

or $\mathsf{Var}(T_0) \leq \mathsf{Var}(T)$. Since T is arbitrary, the proof is complete. □

Theorem 2.23

Let $\mathcal{U}$ be the non-empty class of unbiased estimators as defined in the previous theorem. Then, there exists at most one UMVUE for θ.

Theorem 2.24 (Rao-Blackwell)

Let $\{F_\theta : \theta \in \Theta\}$ be a family of CDF and h be any statistic in $\mathcal{U}$, where $\mathcal{U}$ is the (non-empty) class of all unbiased estimators of θ with $\mathsf{E}_\theta h^2 < \infty$. Let T be a sufficient statistic for the family $\{F_\theta : \theta \in \Theta\}$. Then, the conditional expectation $\mathsf{E}_\theta\{h|T\}$ is independent of θ and is an unbiased estimator of θ. Moreover,

$$\mathsf{E}_\theta[\mathsf{E}(h|T) - \theta]^2 \leq \mathsf{E}(h - \theta)^2, \quad \forall\ \theta \in \Theta \tag{2.30}$$

[9] We can also use that if $\mathsf{E}_{\theta_0}\nu_0^2 = 0$, then ν_0 is 0 almost everywhere since the integral under that measure is zero.

and the equality in (2.30) holds if and only if $h = \mathsf{E}(h|T)$, that is

$$\mathsf{P}\{h = \mathsf{E}(h|T)\} = 1, \quad \forall\, \theta \in \Theta.$$

PROOF We have $\mathsf{E}_\theta\{\mathsf{E}(h|T)\} = \mathsf{E}_\theta h = \theta$ by the iterated law of expectation and $\mathsf{E}(h|T) = \psi(T)$ is an unbiased estimator of θ. It is therefore sufficient to show that

$$\mathsf{E}_\theta\{\mathsf{E}(h|T)\}^2 \leq \mathsf{E}_\theta h^2, \quad \forall\, \theta \in \Theta \tag{2.31}$$

given that the conditional expectation is a projection n the σ-field of a Borel-function and has smaller variance. Note that (2.30) was a mean squared error, now we can use the variance since both are unbiased and the square of the first moments cancel out. Note that $\mathsf{E}_\theta h^2 = \mathsf{E}_t\{\mathsf{E}(h^2|T)\}$, so that it suffices to show that

$$(\mathsf{E}(h|T))^2 \leq \mathsf{E}\left(h^2|T\right) \tag{2.32}$$

Recall that $\mathsf{E}(\mathsf{E}(X|Y)) = \mathsf{E}X$ and $\mathsf{Var}X = \mathsf{Var}(\mathsf{E}(X|Y)) + (\mathsf{Var}(X|Y))$. Using Cauchy-Schwartz inequality,

$$\mathsf{E}^2(h|T) \leq \mathsf{E}(h^2|T)\mathsf{E}(1|T)$$

and the inequality (2.32) follows. Remark that we implicitly used sufficiency here to obtain independence of T and θ so that we could us a statistic.

The equality holds in (2.30) if and only if

$$\underbrace{\mathsf{E}_\theta[\mathsf{E}(h|T)]^2}_{\psi(T)} = \underbrace{\mathsf{E}_\theta h^2}_{\mathsf{E}_\theta[\mathsf{E}(h^2|T)]},$$

that is $\mathsf{E}_\theta[\mathsf{E}(h^2|T) - \mathsf{E}^2(h|T)] = 0$ which is the same as $\mathsf{E}_\theta\{\mathsf{Var}(h|T)\} = 0$. This happens if and only if $\mathsf{Var}(h|T) = 0$, that is if and only if $\mathsf{E}(h^2|T) = \mathsf{E}^2(h|T)$ as will be the case if and only if h is a function of T. Thus $h = \mathsf{E}(h|T)$ □

Theorem 2.25 (Lehmann-Scheffé)

If T is a complete sufficient statistic and there exists an unbiased estimator h of θ, there exists a unique UMVUE of θ which is given by $\mathsf{E}(h|T)$.

Corollary 2.26

Let $X_1, \ldots, X_n$ denote a random sample from a distribution that has pdf $f(x\theta), \theta \in \Theta$, let $T = u(X_1, \ldots, X_n)$ be a sufficient statistic for θ and let the family $\{g_1(t_1; \theta) : \theta \in \Theta\}$ of probability density function be complete. If there is a function of T that is an unbiased estimator of θ, then this function of T is the unique unbiased minimum variance estimator of θ.

PROOF If $h_1, h_2 \in \mathcal{U}$, then $\mathsf{E}(h_1|T)$ and $\mathsf{E}(h_2|T)$ are both unbiased and

$$\mathsf{E}_\theta\{\underbrace{\mathsf{E}(h_1|T)}_{\psi_1(T)} - \underbrace{\mathsf{E}(h_2|T)}_{\psi_2(T)}\} = 0, \qquad \forall\, \theta \in \Theta.$$

Since T is a complete sufficient statistic, it follows that $\mathsf{E}(h_1|T) = \mathsf{E}(h_2|T)$ and hence by Rao-Blackwell theorem, $\mathsf{E}(h|T)$ is the UMVUE. □

Example 2.28

Let $X_i \overset{\text{iid}}{\sim} \mathcal{B}(p), i = 1, \ldots, n$ and take $\widehat{p} = \frac{1}{n}\sum_{i=1}^n X_i$. We showed $\sum_{i=1}^n X_i$ is sufficient (so is thus any one-to-one function). We also showed completeness using the fundamental theorem of algebra. Now

$$\mathsf{E}_p(\widehat{p}) = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}np = p$$

is the UMVUE for p.

Example 2.29

Let $X_i \overset{\text{iid}}{\sim} \mathcal{P}(\lambda), i = 1, \ldots, n$. For sufficiency, we can use Fisher-Neyman factorization to get

$$\mathsf{P}_\lambda(X_i = x_i, i = 1, \ldots, n) = \prod_{i=1}^n \mathsf{P}(X_i = x_i) = \frac{e^{-n\lambda}\lambda^{\Sigma x_i}}{\prod_{i=1}^n x_i}$$

and since the denominator is of the form $h(x_1, \ldots, x_n)$ and the numerator of the form $g_\lambda(T(x_1, \ldots, x_n))$, $T = \sum_{i=1}^n X_i$ is sufficient. Now $T \sim \mathcal{P}(n\lambda)$ using the method of moment generating functions; the mgf of T is given by

$$M_T(s) = \mathsf{E}_\lambda\left(e^{sT}\right) = \mathsf{E}_\lambda\left(e^{s\Sigma X_i}\right) = \left[\mathsf{E}_\lambda(e^{sX}\right]^n$$

and

$$\begin{aligned}\mathsf{E}_\lambda(e^{sx} &= \sum_{x=0}^{\infty} e^{sx}\frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=0}^{\infty}\frac{e^{-\lambda}\left(\lambda e^s\right)^x}{x!} \\ &= e^{-\lambda}e^{\lambda e^s} = \exp(\lambda(e^s - 1))\end{aligned}$$

so

$$M_T(s) = \left[e^{\lambda(e^s-1)}\right]^n = e^{n\lambda(e^s-1)}$$

so $T \sim \mathcal{P}(n\lambda)$ and

$$\mathsf{E}_\lambda[g(X)] = \sum_{x=0}^{\infty} g(x)\frac{e^{-\lambda}x^\lambda}{x!} = 0 \;\forall\; \lambda > 0.$$

To show that the Poisson family is complete, write down the proof showing that it belongs to the exponential family. Finally, $T = \sum_{i=1}^n X_i$ and $\mathsf{E}_\lambda(T) = n\lambda$ since $T \sim \mathcal{P}(n\lambda)$. Thus, the UMVUE is $\frac{T}{n}$ since $\mathsf{E}_\lambda\left(\frac{T}{n}\right) = \lambda$.

Example 2.30 (UMVUE for Normal: $\mu, \sigma^2, \sigma$and p^{th} quantile)

Suppose $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Using the exponential family theorem, $(\overline{X}_n, S^2)$ is complete sufficient for (μ, σ^2). $\overline{X}_n$ is UMVUE for μ and S^2 is UMVUE for σ^2. Also, $K(n)S$ is the UMVUE for σ, where

$$K(n) = \sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.$$

Let $Y = \frac{(n-1)S^2}{\sigma^2}$ and recall this is distributed as $\chi^2(n-1)$ random variable.

We can now recover K using that $\left(\sigma^2 Y/(n-1)\right)^{\frac{1}{2}} = S$ and so we have

$$\begin{aligned}
\mathsf{E}(S) &= \frac{\sigma}{\sqrt{n-1}}\mathsf{E}\left[Y^{\frac{1}{2}}\right] \\
&= \frac{\sigma}{\sqrt{n-1}}\int_0^\infty y^{\frac{1}{2}} \underbrace{f_Y(y)}_{\chi^2(n-1)}\, dy \\
&= \frac{\sigma}{\sqrt{n-1}}\int_0^\infty y^{\frac{1}{2}} \frac{y^{\frac{n-1}{2}-1}e^{-\frac{1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-1}{2}}} dy \\
&= \frac{\sigma}{\sqrt{n-1}}\frac{1}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-1}{2}}}\int_0^\infty y^{\frac{n}{2}-1}e^{-\frac{1}{2}} dy \\
&= \frac{\sigma}{\sqrt{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)2^{\frac{n}{2}}}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-1}{2}}}\int_0^\infty \frac{y^{\frac{n}{2}-1}e^{-\frac{1}{2}}}{\Gamma\left(\frac{n}{2}\right)2^{\frac{n}{2}}} dy
\end{aligned}$$

and thus

$$\mathsf{E}(S) = \frac{\sigma}{\sqrt{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)2^{\frac{n}{2}}}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-1}{2}}}$$

Finally, we can tackle the problem of finding the UMVUE of the p^{th} quantile $\mathfrak{z}$, where $p = \mathsf{P}\left(X \leq \mathfrak{z}_p\right)$ equals

$$\mathsf{P}\left(\frac{X-\mu}{\sigma} \leq \frac{\mathfrak{z}_p - \mu}{\sigma}\right) = \mathsf{P}\left(Z \leq \frac{\mathfrak{z}_p - \mu}{\sigma}\right) = \mathsf{P}(Z \leq Z_{1-p})$$

where $Z \sim \mathcal{N}(0,1)$ is the standard normal distribution. We can go in the table to look at Z_{1-p} and we can recover $\mathfrak{z}_p = Z_{1-p}\sigma + \mu$.

Now $\widehat{\mathfrak{z}}_p = Z_{1-p}\widehat{\sigma} + \widehat{\mu}$. We thus only need to find the UMVUE of $\widehat{\sigma}$ and of $\widehat{\mu}$ and since this is a linear combination, we have a UMVUE of $\mathfrak{z}_p$, that is $Z_{1-p}K(n)S + \overline{X}_n$.

Example 2.31 (UMVUE for discrete uniform parameter N)

Suppose that a sample of n iid observations X_i are drawn from the discrete uniform $(1, \ldots, N)$, where N is unknown. We wish to find the UMVUE of N. First, we can use a result proved earlier on, namely that $M_n = \max_{1 \leq i \leq n}$ is complete sufficient. Next, using Rao-Blackwell and Lehmann-Scheffé, it suffices to find $g(.)$, where $\mathsf{E}_N(G(M_n)) = N \;\forall\; N \in \mathbb{N}$. The distribution of

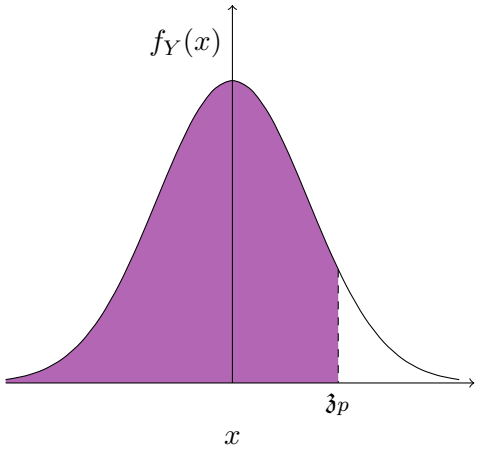


Figure 8: p^{th} quantile of the Normal distribution.

This can be used for example when our data is sensitive to extreme observations (for example salaries). We often are given quartiles: median and medians for both the upper half of the data and the lower half.

the n^{th} order statistic was given as

$$\mathsf{P}(M_n = y) = \mathsf{P}(M_n \leq y) - \mathsf{P}(M_n \leq y-1) = \left(\frac{y}{N}\right)^n - \left(\frac{y-1}{N}\right)^n$$

and

$$\mathsf{E}_N\left[g(M_n)\right] = \sum_{y=1}^{N} g(y)\mathsf{P}(M_n = y) = N \qquad \forall\, N \in \mathbb{N}$$

and this if and only if

$$\begin{aligned} & \sum_{y=1}^{N} g(y)\frac{y^n - (y-1)^n}{N^n} = N \\ \Leftrightarrow \quad & \sum_{y=1}^{N} g(y)(y^n - (y-1)^n) = N^{n+1} \end{aligned}$$

and denoting the left hand side of the equation by $\varphi(N)$, we can see that

we can recover a difference equation

$$\varphi(N) - \varphi(N-1) = N^{n+1} - (N-1)^{n+1}$$

and we have that

$$g(N)\left[N^n - (N-1)^n\right] = N^{n+1} - (N-1)^{n+1}$$

and thus

$$g(N) = \frac{N^{n+1} - (N-1)^{n+1}}{N^n - (N-1)^n}.$$

Section 2.6. Lower bound for variance

Recall that we have defined in the beginning of the course Fisher's information

Definition 2.27 (Fisher's information)

The quantity

$$I(\theta) = \mathsf{E}_\theta\left[\frac{\partial \log f_\theta(X)}{\partial \theta}\right]^2 \tag{2.33}$$

and we have that information increase linearly; if we have a random sample of size n, then

$$I_n(\theta) = \left[\frac{\partial \log f_\theta(\boldsymbol{X})}{\partial \theta}\right]^2 = nI(\theta)$$

The next result has loads of implications and applications and will be followed by many examples.

Theorem 2.28 (Fréchet-Cramer-Rao inequality)

Let Θ be an open interval of the interval of the real line ($\Theta \subset \mathbb{R}$) and let $\{f_\theta : \theta \in \Theta\}$ be a family of PDF or PMF. Assume that the set$\{x : f_\theta(x) = 0\}$ is independent of θ [10]. For every θ, let $\frac{\partial f_\theta(x)}{\partial \theta}$ be defined. Suppose that for

[10] This is the assumption of same support; examples where this conditions fail include the uniform on $(0, \theta)$; it holds for the exponential family. The class of probability measure are equivalent if $\mu \ll \nu, \nu \ll \mu$.

every $\theta \in \Theta$,[11]

$$\frac{\partial}{\partial \theta} \int_{\boldsymbol{x}} f_\theta(\boldsymbol{x}) d\boldsymbol{x} = \int_{\boldsymbol{x}} \frac{\partial}{\partial \theta} f_\theta(\boldsymbol{x}) d\boldsymbol{x} = 0. \tag{2.34}$$

Let ψ be defined on θ and differentiable there and let T be an unbiased estimator of ψ such that $\mathsf{E}_\theta T^2 < \infty \; \forall \; \theta \in \Theta$. Assume that

$$\frac{\partial}{\partial \theta} \int_{\boldsymbol{x}} T(\boldsymbol{x}) f_\theta(\boldsymbol{x}) d\boldsymbol{x} = \int_{\boldsymbol{x}} T(\boldsymbol{x}) \frac{\partial}{\partial \theta} f_\theta(\boldsymbol{x}) d\boldsymbol{x} \qquad \forall \; \theta \in \Theta. \tag{2.35}$$

Let φ be any function of $\Theta \to \mathbb{R}$. Then

$$[\psi'(\theta)]^2 \leq \mathsf{E}_\theta \left\{T - \varphi(t)\right\}^2 \mathsf{E}_\theta \left\{ \frac{\partial \log f_\theta(\boldsymbol{x})}{\partial \theta} \right\}^2 \tag{2.36}$$

for all $\theta \in \Theta$. This resembles to a large extend Heisenberg uncertainty principle; the rightmost term is the Fisher's information that was introduced earlier on. The interesting case will be when ψ and φ are the same: we then get a lower bound for the variance. If $[\psi'(\theta)]^2 = 1$, then the variance will be inversely proportional to the information. We cannot get zero if the regularity conditions hold; there is an intrinsic uncertainty that we cannot make disappear and our inference won't be better than some level. Here are some corollary to the main result.

For any $\theta_0 \in \Theta$, wither $\psi'(\theta_0) = 0$ and equality holds in (2.36) for $\theta = \theta_0$ or we have

$$\mathsf{E}_{\theta_0} \left\{T - \varphi(\theta)\right\}^2 \geq \frac{[\psi'(\theta_0)]^2}{\mathsf{E}_{\theta_0} \left\{ \frac{\partial \log f_\theta(\boldsymbol{x})}{\partial \theta} \right\}^2} \tag{2.37}$$

If, in the later case, equality holds in (2.37), then there exists a real number $K(\theta_0) \neq 0$ such that

$$T(\boldsymbol{x}) - \varphi(\theta_0) = K(\theta_0) \frac{\partial \log f_{\theta_0}(\boldsymbol{x})}{\partial \theta} \tag{2.38}$$

[11] The equivalent statement for (2.34) will see the integrals replaced with sums if we have a PMF.

with probability 1 provided that T is not a constant.

Remark

- Conditions (2.34) and (2.35) are frequently refered to as the *regularity conditions*.

- It is assumed implicitly that

$$0 < \mathsf{E}_\theta \left(\frac{\partial}{\partial \theta} \log f_\theta(\boldsymbol{x}) \right) < \infty \quad \forall\, \theta \in \Theta.$$

  There has been a lot of literature on identification problem and the case of the term being zero; this is a big concern in econometrics. For that, see the work of Franklin Fisher or Thomas Rothenberg.

One big question is identifiability, which we define as

Definition 2.29 (Identifiability)

A model $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ is said to be identifiable if the map from $\Theta \xrightarrow{\Pi} \mathcal{A}$, where $\mathcal{A}$ is the space of probability measures, is one-to-one. That is

$$\mathsf{P}_{\theta_1} = \mathsf{P}_{\theta_2} \Rightarrow \theta_1 = \theta_2 \qquad \forall\, \theta_1, \theta_2 \in \Theta.$$

Say that we have an identified model; we know the shape up to finitely many unknown parameters. Econometricians have show that having an injective map Π is like $\mathsf{E}_\theta \left\{ \frac{\partial}{\partial \theta} \log f_\theta(\boldsymbol{x}) \right\}$ being positive definite matrix. Admissibility can hold if this is zero at some parameter value (identification is not violated). Some results, including Sard-Smale-Morse theorem in differential topology, allow one to capture how often does positive definitiveness fails if it is admissible. There is a close tie between the two notions, but it is not perfect. The set of θ such that

$$\mathsf{E}_\theta \left\{ \theta \in \Theta : \mathsf{E}_\theta \left\{ \frac{\partial}{\partial \theta} \log f_\theta(\boldsymbol{x}) \right\}^2 = 0 \right\}$$

is a nowhere dense set[12] and under certain conditions has Lebesgue measure zero[13]

PROOF If follows from (2.34) that $\mathsf{E}_\theta \left\{ \frac{\partial}{\partial \theta} \log f_\theta(\boldsymbol{x}) \right\} = 0$

$$\int_{\boldsymbol{x}} \frac{\frac{\partial}{\partial \theta} f_\theta(\boldsymbol{x})}{f_\theta(\boldsymbol{x})} f_\theta(\boldsymbol{x}) d\boldsymbol{x} = \frac{\partial}{\partial \theta} \int_{\boldsymbol{x}} f_\theta(\boldsymbol{x}) d\boldsymbol{x} = 0$$

and from (2.35) that

$$\mathsf{E}_\theta \left\{ T(\boldsymbol{x}) \frac{\partial}{\partial \theta} \log f_\theta(\boldsymbol{x}) \right\} = \psi'(\theta).$$

To see this, write the definition of expectation

$$\begin{aligned} \int_{\boldsymbol{x}} T(\boldsymbol{x}) \frac{\frac{\partial}{\partial \theta} f_\theta(\boldsymbol{x})}{f_\theta(\boldsymbol{x})} f_\theta(\boldsymbol{x}) d\boldsymbol{x} &= \int_{\boldsymbol{x}} T(\boldsymbol{x}) \frac{\partial}{\partial \theta} f_\theta(\boldsymbol{x}) d\boldsymbol{x} \\ &= \frac{\partial}{\partial \theta} \int_{\boldsymbol{x}} T(\boldsymbol{x}) f_\theta(\boldsymbol{x}) d\boldsymbol{x} \\ &= \frac{\partial}{\partial \theta} \mathsf{E} T(\boldsymbol{x}) = \frac{\partial}{\partial \theta} \psi(\theta) = \psi'(\theta) \end{aligned}$$

so that

$$\mathsf{E}_\theta \left\{ [T - \varphi(\theta)] \frac{\partial}{\partial \theta} \log f_\theta(\boldsymbol{x}) \right\} = \psi'(\theta)$$

using $\mathsf{E}_\theta \left\{ \frac{\partial}{\partial \theta} \log f_\theta(\boldsymbol{x}) \right\} = 0$ and $\mathsf{E}_\theta \left\{ T(\boldsymbol{x}) \frac{\partial}{\partial \theta} \log f_\theta(\boldsymbol{x}) \right\} = \psi'(\theta)$. By Cauchy-Schwartz, (2.36) follows immediately. To prove (2.37), it suffices to consider the case where $\psi'(\theta_0) \neq 0$ or the one where in (2.36) the equality sign does not hold for $\theta = \theta_0$.

To see this, consider the following set an

$$A = \psi'(\theta_0) = 0; \qquad B = \text{ equality holds in (2.36)}; \qquad C = (2.37) \text{ hold.}$$

[12] That is, int(cl(A))=$\emptyset$; the closure does not have any open sets.

[13] Here, we require finite dimensionality, a continuous function of θ and admissibility. Even C^∞ is not good enough: we need analytic function in order to have a set of measure zero. See the work of Oxtoby and Ulam.

Using basic logic, the following are equivalent: $(A \vee B) \wedge C = D$ and

$$\neg D = (\neg A \wedge \neg C) \vee (\neg B \wedge \neg C)$$

In either case it follows from (2.36) that $\mathsf{E}_{\theta_0}\left\{\frac{\partial \log f_\theta(x)}{\partial \theta}\right\}^2 > 0$ and hence (2.37) follows (since $0 < [\psi'(\theta_0)]^2$, then the two terms must be positive.)

$$(\psi'(\theta_0))^2 \leq \mathsf{E}_{\theta_0}\{T - \varphi(\theta)\}\mathsf{E}_\theta\left\{\frac{\partial \log f_\theta(x)}{\partial \theta}\right\}^2$$

If the equality holds in (2.37) then necessarily $\psi'(\theta_0) \neq 0$.

Remark

Why is it necessary? Since the support of the probability measures and the probability measures are equivalent. Thus, $T = \varphi(\theta_0)$ almost surely. P_{θ_0} implies $T = \varphi(\theta_0)$ almost surely P_θ, $\forall\ \theta$ and therefore T is constant.

Therefore, from Cauchy-Schwartz inequality, there exists a real number $k(\theta_0)$ such that

$$T(X) - \varphi(\theta_0) = k(\theta_0)\left.\frac{\partial \log f_{\theta_0}(x)}{\partial \theta}\right|_0$$

and (2.38) holds. Since □

Since T is not a constant, it follows that $k(\theta_0) \neq 0$.

Remark

If we take $\varphi = \psi$ in (2.37), we find

$$\mathsf{Var}(T(X)) \geq \frac{(\psi'(\theta))^2}{\mathsf{E}_\theta\left\{\frac{\partial \log f_\theta(x)}{\partial \theta}\right\}^2} \tag{2.39}$$

In particular, if $\psi(\theta) = \theta$ (2.39) reduces to[14]

$$\mathsf{Var}(T(X)) \geq (\mathsf{E}_\theta \left\{ \frac{\partial \log f_\theta(x)}{\partial \theta} \right\}^2)^{-1} = \frac{1}{I(\theta)}.$$

Remark

If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a sample from a PDF (PMF) $f_\theta(x), \theta \in \Theta$, then

$$\begin{aligned}
\mathsf{E}_\theta \left\{ \frac{\partial \log f_\theta(\bar{x})}{\partial \theta} \right\}^2 &\overset{\perp\!\!\!\perp}{=} \mathsf{E}_\theta \left\{ \frac{\partial \log \prod_1^n f_\theta(x_i)}{\partial \theta} \right\}^2 \\
&= \sum_{i=1}^n \mathsf{E}_\theta \left\{ \frac{\partial \log f_\theta(x_i)}{\partial \theta} \right\}^2 \\
&= n\mathsf{E}_\theta \left\{ \frac{\partial \log f_\theta(x)}{\partial \theta} \right\}^2
\end{aligned}$$

Note that the cross terms are zero since the functions are independent and the product of expectations is the expectation of the product; the correlation terms are zero from independence of the random variables. We also have $\mathsf{E}\left(\frac{\partial \log f_\theta(x)}{\partial \theta}\right) = 0$. Fisher though that information should increase linearly, this is why he worked with the log function, because he wanted to force his information function to vary linearly with respect to the data.

Example 2.32

$X \sim \mathcal{B}(n, \theta), \theta \in \Theta = (0, 1)$. First, we check the regularity conditions:

(a) $\{x : f_\theta(x) = 0\}$ is independent of θ and $f_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ for $x = 0, 1, \ldots, n$.

(b) We can interchange differential and integral operators since we are working with a discrete distribution.

(c) We have only one sample: suppose $T(X)$ is an unbiased estimate of θ,

[14] The information about what your observables carry about the unknown parameter tell you that the uncertainty from the variance cannot be below the uncertainty from the data, which is intuitive.

i.e. $\mathsf{E}_\theta[T(X)] = \theta$, then

$$\log f_\theta(x) = \log \binom{n}{x} + x \log \theta + (n-x) \log(1-\theta)$$

and taking derivatives

$$\left\{ \frac{\partial \log f_\theta(x)}{\partial \theta} \right\}^2 = \left\{ \frac{x}{\theta} - \frac{n-x}{1-\theta} \right\}^2$$

and $I(\theta)$ is equal to

$$\begin{aligned} I(\theta) = \mathsf{E}_\theta \left\{ \frac{\partial}{\partial \theta} \log f_\theta(x) \right\}^2 &= \mathsf{E}_\theta \left\{ \frac{x}{\theta} - \frac{n-x}{1-\theta} \right\}^2 \\ &= \mathsf{E}_\theta \left\{ \frac{x^2}{\theta^2} \right\} + \mathsf{E}_\theta \left\{ (\frac{n-x}{1-\theta})^2 \right\} - 2\mathsf{E} \left\{ \frac{x(n-x)}{\theta(1-\theta)} \right\} \end{aligned}$$

Recall that $\mathsf{Var}(X) = n\theta(1-\theta), \mathsf{E}(x) = n\theta$, hence

$$\begin{aligned} \mathsf{E}_\theta(x^2) &= \mathsf{Var}(x) + (\mathsf{E}(x))^2 \\ &= n\theta(1-\theta) + (n\theta)^2 \\ &= n\theta - n\theta^2 + n^2\theta^2 \\ &= n\theta(1 - \theta + n\theta) \\ &= n\theta((n-1)\theta + 1) \\ &= \frac{1}{\theta^2}\mathsf{E}_\theta(x^2) + \frac{1}{(1-\theta)^2}\mathsf{E}((n-x)^2) - \frac{\theta}{\theta(1-\theta)}(n\mathsf{E}(x) - \mathsf{E}(x^2)) \end{aligned}$$

Note that $n - x \sim \mathcal{B}(n, 1-\theta)$, hence

$$\mathsf{E}[(n-x)^2] = n(1-\theta)[(n-1)(1-\theta) + 1]$$

and after completion of the missing steps, we get

$$\mathsf{E}_\theta \left\{ \frac{\partial \log f_\theta(x)}{\partial \theta} \right\}^2 = \frac{n}{\theta(1-\theta)}.$$

There is however an easier way to show this. If the regularity conditions hold, then we can show that

$$\mathsf{E}_\theta\left(\frac{\partial \log f_\theta(x)}{\partial \theta}\right)^2 = -\mathsf{E}_\theta\left(\frac{\partial^2 \log f_\theta(x)}{\partial \theta^2}\right) \tag{2.40}$$

which is an important property of the information. Recall that

$$\frac{\partial \log f_\theta(x)}{\partial \theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

and so

$$\frac{\partial^2 \log f_\theta(x)}{\partial \theta^2} = \frac{-x}{\theta^2} - \frac{n-x}{(1-\theta)^2}.$$

Note that we only need to take expectation once

$$-\mathsf{E}\left\{\frac{\partial^2 \log f_\theta(x)}{\partial \theta^2}\right\} = \frac{\mathsf{E}(x)}{\theta^2} + \frac{\mathsf{E}(n-x)}{(1-\theta)^2} = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)}$$

Using Fréchet-Cramer-Rao, we have $\mathsf{Var}(T(x)) \geq \frac{\theta(1-\theta)}{n}$. Note that $\widehat{\theta} = x/n$ is an unbiased estimator of θ and $\mathsf{Var}(\widehat{\theta}) = \mathsf{Var}\left(\frac{x}{n}\right) = \frac{\theta(1-\theta)}{n}$. The sample proportion attains the lower bound, so this is as good as it gets in terms of variance. There are many cases of UMVUE which don't attain the FCR bound, but this one does. We can use this bound as a golden rule.

Example 2.33

Say $X_i \overset{\text{iid}}{\sim} \mathcal{U}(0,\theta)$. Here, the support depend on the parameter. We can however still calculate Fisher's information and see that we cannot reach FCR lower bound with the UMVUE. We have

$$f_\theta = \begin{cases} 1/\theta & \text{if } 0 < x < \theta \\ o & \text{otherwise} \end{cases},$$

$\log f_\theta(x) = -\log\theta$, and $\frac{\partial}{\partial\theta}\log f_\theta(x) = -\frac{1}{\theta}$. We cannot use the second derivative here (because of the lack of regularity), therefore we calculate

$$I(\theta) = \mathsf{E}_\theta\left\{\frac{\partial}{\partial\theta}\log f_\theta(x)\right\}^2 = \frac{1}{\theta^2}$$

We learnt that $M_n = \max x_i$ is a complete sufficient statistic and it is enough to show that $\mathsf{E}[M_n] = \frac{n}{n+1}\theta$, hence

$$\begin{aligned} f_{M_n}(x) &= \frac{d}{dx}F_{M_n}(x) = \frac{d}{dx}\mathsf{P}(M_n \le x) \\ &= \frac{d}{dx}\mathsf{P}(X_i \le x_i; i = 1, 2, \ldots, n) \\ &= \frac{d}{dx}\left(\prod_1^n \mathsf{P}(X_i \le x)\right) = \frac{d}{dx}F_X^n(x) \\ &= n f_X(x) F_X^{n-1}(x) \end{aligned}$$

if $0 < x < \theta$ and thus

$$f_{M_n}(x) = \begin{cases} n\frac{1}{\theta}\left(\frac{x}{\theta}\right)^{n-1} & \text{if } 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

so

$$\mathsf{E}_\theta(M_n) = \int_0^\theta x\frac{nx^{n-1}}{\theta^n}dx = \frac{n}{\theta^n}\int_0^\theta x^n dx = \frac{n}{\theta^n}\frac{\theta^{n+1}}{n} = \frac{n}{n+1}\theta$$

is a consistent estimator, but not unbiased. We can cook up an unbiased estimator: $\mathsf{E}_\theta\left(\frac{n+1}{n}M_n\right) = \theta$, thus $\frac{n+1}{n}M_n$ is the UMVUE for θ. By definition,

$$\mathsf{Var}(T) = \mathsf{E}(T^2) - (\mathsf{E}(T))^2$$

and therefore

$$\begin{aligned} \mathsf{E}_\theta(T^2) &= \mathsf{E}\left(\left(\frac{n+1}{n}\right)^2 M_n^2\right) \\ &= \left(\frac{n+1}{n}\right)^2 \mathsf{E}_\theta(M_n^2) \end{aligned}$$

and

$$\begin{aligned}\mathsf{E}(M_n^2) &= \int_0^\theta \frac{x^2 n x^{n-1}}{\theta^n} dx \\ &= \frac{n}{\theta^n}\int_0^\theta x^{n+1} dx \\ &= \frac{n}{\theta^2}\frac{\theta^{n+2}}{n+2} = \frac{n}{n+2}\theta^2 \\ &= \left(\frac{n+1}{n}\right)^2 \frac{n}{n+1}\theta^2\end{aligned}$$

so

$$\begin{aligned}\mathsf{Var}_\theta(T) &= \frac{(n+1)^2}{(n+2)n}\theta^2 - \theta^2 \\ &= \theta^2\left(\frac{(n+1)^2}{n(n+2)} - 1\right) \\ &= \frac{\theta^2}{n(n+2)}\end{aligned}$$

for the UMVUE. Compare this with $nI(\theta) = n\frac{1}{\theta^2}$, FCR lower bound gives us $\frac{\theta^2}{n}$ while $\mathsf{Var}(T) = \frac{\theta^2}{n(n+2)}$.

Theorem 2.30 (Chapmans, Robbins & Kiefer inequality)

Let $\Theta \subset \mathbb{R}$ and $\{f_\theta(x) : \theta \in \Theta\}$ be a class of PDF. Let ψ be defined on Θ and let T be an unbiased estimator of $\psi(\theta)$ with $\mathsf{E}_\theta T^2 < \infty$, $\forall\ \theta \in \Theta$. If $\theta \neq \varphi$, assume that f_θ and f_φ are different and assume further that $\exists \varphi \in \Theta$ such that $\theta \neq \varphi$ and

$$S(\theta) = \{x : f_\theta(x) > 0\} \supset S(\varphi) = \{x : f_\varphi(x) > 0\}.$$

Then

$$\mathsf{Var}_\theta(T(x)) \geq \sup_{\left\{\substack{\varphi : S(\varphi \subseteq S(\theta)) \\ \varphi \neq \theta\ \forall\ \theta \in \Theta}\right\}} \frac{(\psi(\varphi) - \psi(\theta))^2}{\mathsf{Var}\{f_\varphi(x)/f_\theta(x)\}} \tag{2.41}$$

PROOF The proof is for the continuous case; it is similar for discrete random variables. First, notice that

$$\begin{aligned}\mathsf{E}_\theta T(X)\left(\frac{f_\varphi - f_\theta}{f_\theta}\right) &= \int_{S(\theta)} T(x)\frac{f_\varphi - f_\theta}{f_\theta} f_\theta dx \\ &= \int_{S(\theta)} T(x) f_\varphi - f_\theta dx \\ &= \int_{S(\varphi)} T(x) f_\varphi dx - \int_{S(\theta)} T(x) f_\theta dx \\ &= \psi(\varphi) - \psi(\theta)\end{aligned}$$

first by cancelling the densities f_θ, then by using linearity and noting that f_φ is zero evaluated outside of $S(\varphi)$. Also, $T(X)$ is unbiased and we recover the last term.

If we look at the covariance between $T(X)$ and $f_\varphi - f_\theta$, we observe from properties of covariance that

$$\mathsf{Cov}\left(T(X), \frac{f_\varphi}{f_\theta}\right) = \mathsf{Cov}\left(T(X), \frac{f_\varphi}{f_\theta} - 1\right) = \mathsf{Cov}\left(T(X), \frac{f_\varphi - f_\theta}{f_\theta}\right).$$

We can write the above as

$$\mathsf{E}_\theta\left(T(X)\frac{f_\varphi - f_\theta}{f_\theta}\right) - \mathsf{E}_\theta T(X)\mathsf{E}_\theta\left(\frac{f_\varphi - f_\theta}{f_\theta}\right)$$

where the rightmost term is zero since it involves a difference of distributions functions; write this as

$$\mathsf{E}_\theta\left(\frac{f_\varphi - f_\theta}{f_\theta}\right) = \int_{S(\varphi)} f_\varphi dx - \int_{S(\theta)} f_\theta dx = 1 - 1 = 0.$$

Finally, using the Cauchy-Schwartz inequality, we get that $\mathsf{Cov}^2(Z, Y) \leq \mathsf{Var}Z\mathsf{Var}Y$ for given random variables Z, Y. Using this, we get that

$$\mathsf{Cov}^2\left(T(X), \frac{f_\varphi}{f_\theta}\right) = [\psi(\varphi) - \psi(\theta)]^2 \leq \mathsf{Var}_\theta T(X)\mathsf{Var}_\theta\left(\frac{f_\varphi}{f_\theta}\right) \tag{2.42}$$

and in particular, if $S(\varphi) \subset S(\theta)$ and the rightmost term of (2.42) is non-vanishing, then we can divide through by $\mathsf{Var}_\theta\left(\frac{f_\varphi}{f_\theta}\right)$ to get (2.41). □

Section 2.7. Maximum likelihood estimates

We now present an approach which is quite intuitive (compared to the method of moment which lacks justification). The maximum likelihood estimate (or MLE) is the most commonly used approach in estimation. First, we have observables and unobservables, which are linked in the parametric setting by $f_\theta(x)$. $\boldsymbol{x}$ tells me about the likeliness under θ of observing this value. As a function of $\theta, \mathcal{L}_\theta = \mathsf{P}_\theta(X_i = x_i, i = 1, \ldots, n)$ is called the likelihood function . If we are not superstitious, what we have observed is something that we should have expected to happen: we want to fit as best as possible or observations and get good results, looking at $\widehat{\theta}_n = \text{argmax } \mathcal{L}(\theta)$ and we have shown convergence in many different cases

1. $\widehat{\theta}_n \xrightarrow{\text{a.s.}} \theta$ is in particular of interest to us.

   We also have under rather general conditions

2. $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I^{-1}(\boldsymbol{\theta}))$, where $I^{-1}(\theta)$ is the lower bound of FCR inequality.

   Maximum likelihood estimates will be asymptotically most efficient. The more information you have, the lower the variance gets. It is important because as statisticians, we want to make confidence intervals as we care about margins of error (which is not the case in fields like machine learning, who want to hit the target, period.)

3. If $\widehat{\theta}_n$ is the MLE of θ, then $\psi(\widehat{\theta}_n)$ is the MLE of $\psi(\theta)$.

If the likelihood is continuously differentiable, then we can find the max using the first derivative.

Example 2.34 (MLE of Bernoulli)

Let $X_i \overset{\text{iid}}{\sim} \mathcal{B}(p)$ for $i = 1, \ldots, n$. First, we set up the likelihood

$$\begin{aligned}\mathcal{L}(p) &= \mathsf{P}_p(X_i = x_i, i = 1, \ldots, n) \\ &= \prod_{i=1}^{n} \mathsf{P}(X_i = x_i) \\ &= \prod_{i=1}^{n} \left[p^{x_i}(1-p)^{1-x_i}\right] \\ &= p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}.\end{aligned}$$

Let now $t = \sum_{i=1}^{n} x_i$. We can look at the log-likelihood since the logarithm is a monotonic transformation, it does not change the maximum and is easier. We get

$$\frac{d \log \mathcal{L}(p)}{dp} = \frac{d}{dp}\left[t \log p + (n-t)\log(1-p)\right] = \frac{t}{p} - \frac{n-t}{1-p}.$$

Therefore, $\widehat{p}$ should satisfy $\frac{t}{\widehat{p}} - \frac{n-t}{1-\widehat{p}} = 0$ and $(1-\widehat{p})t = (n-t)\widehat{p}$, so $t - t\widehat{p} - \widehat{p}n + t\widehat{p} = 0$ which imply that $\widehat{p} = \frac{t}{n}$ so the sample proportion is the estimate of the MLE, which coincides in this case with the UMVUE.

Example 2.35 (MLE of Normal for μ, σ^2)

Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \ldots, n$. The joint PDF or PMF will be our likelihood

$$\mathcal{L}(\mu, \sigma^2) = f_{\mu,\sigma^2}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right).$$

Looking at the log-likelihood $\ell(\mu, \sigma^2) = \log \mathcal{L}(\mu, \sigma^2)$, we have

$$\sum_{i=1}^{n}\left[-\log(\sqrt{2\pi}\sigma) - \frac{(x_i-\mu)^2}{2\sigma^2}\right] = -n\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}.$$

Let $\sigma^2 = \theta$ and write the previous as

$$-\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

We are interested in $\frac{\partial \ell(\mu,\theta)}{\partial \mu}=0$ and $\frac{\partial \ell(\mu,\theta)}{\partial \theta}$, then solve the resulting system of equation and look at the second derivative to verify indeed that it is negative and we obtain a maximum. In the multivariate case, we must verify that the Hessian matrix is negative semi-definite. Thus

$$\begin{aligned}\frac{\partial \ell(\mu,\theta)}{\partial \mu}=0 &\Leftrightarrow \frac{1}{2\sigma^2}\cdot 2\sum_{i=1}^{n}(X_i-\mu)=0\\ \frac{\partial \ell(\mu,\theta)}{\partial \theta}=0 &\Leftrightarrow -\frac{n}{2}\frac{1}{\theta}+\frac{1}{2\theta^2}\sum_{i=1}^{n}(X_i-\mu)^2\end{aligned}$$

which yields $\widehat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}_n$ and for the second, assuming the distribution is non degenerate (that is $\sigma^2\neq 0$), then

$$\frac{1}{\theta}\sum_{i=1}^{n}(X_i-\mu)^2=n\Leftrightarrow\widehat{\theta}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\widehat{\mu})^2=\widehat{\sigma^2}.$$

We talked about this a while ago; S^2 is divided by $n-1$ while the MLE is divided by n and is thus biased.

We could compare the MSE of the two, knowing that $\widehat{\sigma^2}=\left(\frac{n-1}{n}\right)S^2$, so

$$\mathsf{E}\widehat{\sigma^2}=\mathsf{E}\left[\left(\frac{n-1}{n}S^2\right)\right]=\left(\frac{n-1}{n}S^2\right)\sigma^2=\left(1-\frac{1}{n}\right)\sigma^2$$

and furthermore

$$\mathsf{Var}\widehat{\sigma^2}=\mathsf{Var}\left(\frac{n-1}{n}S^2\right)=\frac{\sigma^4}{n^2}\mathsf{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right).$$

Note that $\frac{(n-1)S^2}{\sigma^2}$ is distributed as $\chi^2(n-1)$ so the variance is $2(n-1)$ which entail $\mathsf{Var}\left(\widehat{\sigma^2}\right)=2(n-1)\frac{\sigma^4}{n^2}$. The mean square error is as usual given by

$$\mathsf{MSE}\left(\widehat{\sigma^2}\right)=\mathsf{Var}\left(\widehat{\sigma^2}\right)+\mathsf{Bias}^2\left(\widehat{\sigma^2}\right)$$

where the square of the bias is $\left[\sigma^2-\left(1-\frac{1}{n}\right)\sigma^2\right]^2=\frac{\sigma^4}{n^2}$ so that the mean

square error

$$\mathsf{MSE}\left(\widehat{\sigma^2}\right) = 2(n-1)\frac{\sigma^4}{n^2} + \frac{\sigma^4}{n^2} = \frac{2n-1}{n^2}\sigma^4.$$

What about the variance of S^2? We get

$$\mathsf{Var}(S^2) = \mathsf{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right)\frac{\sigma^4}{(n-1)^2} = 2(n-1)\frac{\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1}$$

and comparing the two

$$\frac{2n-1}{n^2}\sigma^4 < \frac{2\sigma^4}{n-1} \quad \Leftrightarrow \quad (2n-1)(n-1) < 2n^2 \quad \Leftrightarrow 2n^2 - 3n + 1 < 2n^2$$

happens for all $n \in \mathbb{N}$. Surprisingly enough, $\mathsf{MSE}\left(\widehat{\sigma^2}\right) < \mathsf{MSE}(S^2)$ and the MLE estimator has a smaller mean square error.

Example 2.36 (MLE of Uniform)

Let $X_i \overset{\text{iid}}{\sim} \mathcal{U}(0,\theta)$. If I look at

$$\mathcal{L}(\theta) = \frac{1}{\theta^n}\mathbf{1}_{(0,\theta)}\left(\max_{1\le i\le n} x_i\right)\mathbf{1}_{(0,\max x_i)}\left(\min_{1\le i\le n} x_i\right),$$

but the last term does not play any role for the sufficiency of the MLE since it does not depend on any unknown parameters which yield

$$\mathcal{L}(\theta) = \begin{cases} \frac{1}{\theta^n} & \text{if } 0 < \max_{1\le i\le n} x_i \le \theta \\ 0 & \text{otherwise} \end{cases}$$

is a decreasing function of θ. We need to find the minimum value of the domain. So

$$\frac{1}{\left(\max_{1\le i\le n} x_i\right)^n} = \sup_{\theta > \max_{1\le i\le n} x_i} \frac{1}{\theta^n}.$$

Example 2.37 (MLE of Hypergeometric (N))

This is used to determine the population size in fish ponds using capture-recapture method. if we condition on n, the inference does not depend on that parameter and all of $n-x$, x, n, M are known. We us the MLE to get

a fast estimate. Recall that

$$P_N(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad \text{if } \max(0, n-N+M) \leq x \leq \min(n, M)$$

and zero otherwise. This is the likelihood since we are dealing with one sample from hypergeometric (where there is no replacement) compared to the n samples of the binomial (with replacement).

In this case, the maximimizer needs not to be unique, as long as it satisfies

$$\frac{\mathsf{P}_{N^*}(x)}{\mathsf{P}_{N^*+1}(x)} \geq 1 \qquad \frac{\mathsf{P}_{N^*}(x)}{\mathsf{P}_{N^*-1}(x)} \geq 1$$

Another example which can be even more illustrative is the uniform, where this time we have

Example 2.38

$X_i \overset{\text{iid}}{\sim} \mathcal{U}\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$ for a sample of size n. If we write down the likelihood, we see that

$$\mathcal{L}(\theta) = \begin{cases} 1 & \text{if } \theta - \frac{1}{2} \leq \min x_i \leq \max x_i \leq \theta + \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

and whatever value you choose between the two does the job, that is take θ such that $\theta \leq \min_{1\leq i\leq n} x_i + \frac{1}{2}$ and $\max_{1\leq i\leq n} x_i - \frac{1}{2} \leq \theta$. Indeed, if we look at T_α for $0 < \alpha < 1$, then

$$T_\alpha(x_1, \ldots, x_n) = \max_{1\leq i\leq n} x_i - \frac{1}{2} + \alpha[1 + \min_{1\leq i\leq n} x_i - \max_{1\leq i\leq n} x_i].$$

and T_α is the MLE for any $\alpha \in (0, 1)$. In particular, $\alpha = \frac{1}{2}$ gives us

$$T_{\frac{1}{2}}(x_1, \ldots, x_n) = \frac{\max x_i + \min x_i}{2}.$$

We now proceed with an academic example, which is nevertheless instructive.

Example 2.39 (Oliver)

This example illustrates a distribution for which an MLE is necessarily an actual observation, but not necessarily any particular observation. Suppose $X_1, \ldots, X_n$ is a random sample from PDF

$$f_\theta(x) = \begin{cases} \frac{2}{\alpha}\frac{x}{\theta} & \\ \frac{2}{\alpha}\frac{\alpha - x}{\alpha - \theta} & \text{if } \theta < x \leq \alpha \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha > 0$ is a (known) constant.

First, we want to show using the fact that the likelihood is continuous and the parameter space is a compact set that it has a maximizer. New, we show that it is differentiable and then that it cannot be that the MLE is between two points.

The likelihood function is

$$\mathcal{L}(\theta, x_1, \ldots, x_n) = \left(\frac{2}{\alpha}\right)^n \prod_{i:x \leq \theta} \left(\frac{x_i}{\theta}\right) \prod_{i:x_i > \theta} \left(\frac{\alpha - x_i}{\alpha - \theta}\right) \tag{2.43}$$

Assume furthermore that the observations are arranged in increasing order of magnitude, *i.e.* $0 \leq x_1 < x_2 < \ldots, < x_n \leq \alpha$. This function $\mathcal{L}(\theta, \boldsymbol{x})$ is continuous in θ except if $\theta \neq 0, \theta \neq \alpha$, in which case we have to delete one of the branches if $\theta = 0$ or if $\theta = \alpha$. Assume $0 < \theta < \alpha$. We define f_θ by (2.43). If $\theta = 0$,then

$$f_\theta(x) = \begin{cases} \frac{2}{\alpha}\frac{\alpha - x}{\alpha} & \text{if } 0 \leq x \leq \alpha \\ 0 & \text{otherwise} \end{cases}$$

or if $\theta = \alpha$

$$f_\theta(x) = \begin{cases} \frac{2x}{\alpha^2} & \text{if } 0 \leq x \leq \alpha \\ 0 & \text{otherwise} \end{cases}$$

We first need to check that the above are density functions; this is left as an exercise. The parameter space is $\Theta = [0, \alpha]$ and the likelihood in the more

general case is

$$\mathcal{L}(\theta, x_1, \ldots, x_n) = \begin{cases} \left(\frac{2}{\alpha}\right)^n \prod_{i:x \leq \theta} \left(\frac{x_i}{\theta}\right) \prod_{i:x_i > \theta} \left(\frac{\alpha - x_i}{\alpha - \theta}\right) & \text{if } 0 < \theta < \alpha \\ \left(\frac{2}{\alpha^2}\right)^n \prod_{i=1}^n x_i & \text{if } \theta = \alpha \\ \left(\frac{2}{\alpha^2}\right)^n \prod_{i=1}^n (\alpha - x_i) & \text{if } \theta = 0 \end{cases}$$

We need to check continuity at the endpoints

$$\mathcal{L}(\theta, x_1, \ldots, x_n) = \begin{cases} \left(\frac{2}{\alpha}\right)^n \left(\frac{1}{\theta}\right)^a \left(\frac{1}{\alpha - \theta}\right)^b \prod_{i:x \leq \theta} x_i \prod_{i:x_i > \theta} \alpha - x_i & \text{if } 0 < \theta < \alpha \\ \left(\frac{2}{\alpha^2}\right)^n \prod_{i=1}^n x_i & \text{if } \theta = \alpha \\ \left(\frac{2}{\alpha^2}\right)^n \prod_{i=1}^n (\alpha - x_i) & \text{if } \theta = 0 \end{cases}$$

where $a(\theta)$ is the cardinality of $\{i : x_i \leq \theta\}$ and $b(\theta)$ the cardinality of $\{i : x_i > \theta\}$. This condition is clear given $x_1, \ldots, x_n$. Indeed, for $\theta < x_1, a(\theta) = 0$. As such, $a(\theta) \to 0, b(\theta) \to n$ as $\theta \to 0^+$. Then $\mathcal{L}(\theta, x_1, \ldots, x_n)$ is a continuous function of θ on $[0, \alpha]$ such that

$$\mathcal{L}(\hat{\theta}, x_1, \ldots, x_n) \geq \mathcal{L}(\theta, x_1, \ldots, x_n) \; \forall \; \theta \in [0, \alpha].$$

For $x_j, \theta < x_{j+1}$, we have

$$\mathcal{L}(\theta) = \left(\frac{2}{\alpha}\right)^n \theta^{-j} (\alpha - \theta)^{-(n-j)} \prod_{i=1}^{j} x_j \prod_{i=j+1}^{n} (\alpha - x_i)$$

since we have assumed earlier wlog that the observations were ordered in an increasing fashion. Thus,

$$\frac{\partial \ell}{\partial \theta} = -\frac{j}{\theta} + \frac{n-j}{\alpha - \theta}$$

and

$$\frac{\partial^2 \ell}{\partial \theta^2} = \frac{j}{\theta^2} + \frac{n-j}{(\alpha - \theta)^2} > 0.$$

it follows that any stationary value that exists must be a minimum. Hence

$\hat{\theta}$ is not $\hat{\theta} \notin (x_j, x_{j+1}) \; \forall \; j = 1, \ldots, n-1$. If $0 \leq \theta \leq x_i$, then

$$\mathcal{L} = \left(\frac{2}{\alpha}\right)^n (\alpha - \theta)^{-n} \prod_{i=1}^{n} (\alpha - x_i)$$

which is a strictly increasing function of θ. Likewise, one can show that $\mathcal{L}(\theta)$ is a strictly decreasing function of θ in $x_n < \theta \leq \alpha$. Then $\hat{\theta}$ should be one of the observations $x_1, \ldots, x_n$.

In particular, let $\alpha = 5$ and $n = 3$. If the observations arranged in an increasing order of magnitude are $1, 2, 4$, in this case, the MLE can be show to be $\hat{\theta}_{\text{ML}} = 1$. If the sample values are $2, 3, 4$, then $\hat{\theta}_{\text{ML}} = 4$. This is bad, as the MLE does not work well in this case and switch from the smallest to the largest observation; it exhibits quite a bit of instability. We now prove a few important theorem about MLE.

Principle of Sufficiency

Theorem 2.31

Let T be a sufficient statistic for the family of PDF's (PMF's) $f_\theta(x), \theta \in \Theta$. If a MLE of θ exists, it is a function of T.

PROOF Since T is sufficient, we can write

$$\mathcal{L}(\theta, x_1, \ldots, x_n) = h(x_1, \ldots, x_n) g_\theta(T(x_1, \ldots, x_n))$$

using the factorization criterion. Maximization of the likelihood function with respect to θ is therefore equivalent to maximization of $f_\theta(T(x_1, \ldots, x_n))$ which is a function of T also. □

Remark

Although the previous theorem states that the MLE is a function of the sufficient statistic, the MLE itself may not be a sufficient statistic.

Example 2.40

Let $X_i \overset{\text{iid}}{\sim} \mathcal{U}\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$. We learned that any $T(\boldsymbol{x})$ such that

$$\max_{1 \leq i \leq n} \left(x_i - \frac{1}{2}\right) \leq T(x_1, \ldots, x_n) \leq min_{1 \leq i \leq n} \left(x_i + \frac{1}{2}\right)$$

is a MLE. In particular

$$T(x_1, \ldots, x_n) = \min_{1 \leq i \leq n} + \frac{1}{2}$$

is MLE. It is not however a sufficient statistic. Recall that $(\min_{1 \leq i \leq n} x_i, \max_{1 \leq i \leq n} x_i)$ is sufficient in this case.

Section 2.8. Consistency of the MLE

Lemma 2.32

Suppose $X_i \overset{\text{iid}}{\sim} f(x; \theta)$ for $i = 1, 2, \ldots, n$ and $\theta \in \Theta \subseteq \mathbb{R}^p$ is a bounded open set. Define

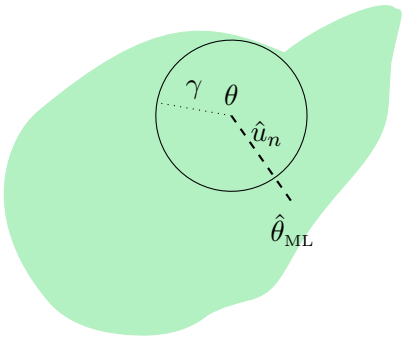
$$R_{n,\theta}^{\frac{1}{2}}(u) = \left[\frac{\mathcal{L}(\theta + u; x_1, \ldots, x_n)}{\mathcal{L}(t; x_1, \ldots, x_n)}\right]^{\frac{1}{2}} = \prod_{i=1}^{n} \frac{f^{\frac{1}{2}}(x_i; \theta + u)}{f^{\frac{1}{2}}(x_i; \theta)} \tag{2.44}$$

where $u \in \mathcal{U} = \Theta - \theta$ and we understand the latter as being the set of all points of Θ shifted by θ . Then

$$\hat{\theta}_{\text{ML}} \xrightarrow{\mathsf{P}} \theta \quad \text{if} \quad \mathsf{P}\left\{\sup_{\|u\| > \gamma} R_{n,\theta}^{\frac{1}{2}} \geq 1\right\} \xrightarrow{n \to \infty} 0 \ \ \forall\, \gamma > 0.$$

If the last bit holds true uniformly in $\theta \in K \subseteq \Theta$, then $\hat{\theta}_{\text{ML}}$ is uniformly consistent in K.

Proof The idea of the proof is simple. We want to show that $\hat{\theta}_{\text{ML}}$ should be in the vicinity of θ. In other words, the chance that $\hat{\theta}_{\text{ML}}$ falls outside of a neighbourhood around θ tends to zero, no matter how small is the



neighbourhood. Set $\hat{u}_n = \hat{\theta}_{\mathrm{ML}} - \theta$ so that $R_{n,\theta}(\hat{u}_n) = \sup_u R_{n,\theta}(u)$. Then

$$\begin{aligned}\mathsf{P}^{(*)_t}\left\{\|\hat{\theta}_{\mathrm{ML}} - \theta\| > \gamma\right\} &= \mathsf{P}_\theta\left\{\|\hat{u}_n\| > \gamma\right\} \\ &\leq \mathsf{P}\left\{\sup_{\|u\|>\gamma} R_{n,\theta}^{\frac{1}{2}} \geq 1\right\} \xrightarrow{n\to\infty} 0\end{aligned}$$

□

Theorem 2.33 (Consistency of the Maximum Likelihood Estimators)
Suppose Θ is a bounded open set of $\mathbb{R}^p$ and $f(x;\theta)$ is the PDF of P_θ with respect to the σ-finite measure ν. Let $f(x;\theta)$ be a continuous function of θ on $\overline{\Theta}$ [15] for almost all $x \in \mathcal{X}$ and let the following conditions be satisfied

1. For all $\theta \in \Theta$ and all $\gamma > 0$,

   $$0 < K_\theta(\gamma) = \inf_{\|\theta-\theta'\|>\gamma} r_2^2(\theta,\theta') = \inf_{\|\theta-\theta'\|>\gamma} \int_x \left[f^{\frac{1}{2}}(x;\theta) - f^{\frac{1}{2}}(x;\theta')\right]\nu(dx)$$

   is positive and thus distinguishable.

2. For all $\theta \in \overline{\Theta}$[16]

   $$\left\{\int_x \sup_{\|t\|\leq\delta}\left[f^{\frac{1}{2}}(x;\theta) - f^{\frac{1}{2}}(x;\theta+t)\right]^2\right\}^{\frac{1}{2}} = \omega_\theta(\delta) \xrightarrow{\delta\to 0} 0$$

[15] $\overline{\Theta}$ denotes the closure of the set Θ
[16] This condition is some type of modulus of continuity condition.

PROOF In view of the above lemma and Markov's inequality, it suffices to find an upper bound for the expectation $\mathsf{E}_\theta[\sup_\Gamma R_n(u)]$ which tends to zero as $n \to \infty$. This is the core of the proof which is given below.

Suppose θ is fixed. Consider $R_{n,\theta}(u)$ as a function of u. Let Γ be a sphere of small radius δ situated in its entirety in the region $\|u\| > \frac{1}{2}\gamma$. We shall bound the expectation $\mathsf{E}_\theta[\sup_\Gamma R_n(u)]$. If u_0 is the center of Γ, then

$$\begin{aligned}\sup_\Gamma R_n^{\frac{1}{2}}(u) &= \sup_\Gamma \prod_{i=1}^n \left[\frac{f(x_i;\theta+u)}{f(x_i;\theta)}\right]^{\frac{1}{2}} \\ &\le \prod_{i=1}^n f^{-\frac{1}{2}}(x_i;\theta)\left[f^{\frac{1}{2}}(x_i;\theta+u_0)\right. \\ &\quad \left. + \sup_{\|t\|\le\delta}\left|f^{\frac{1}{2}}(x_i;\theta+u_0+t) - f^{\frac{1}{2}}(x_i;\theta+u_0)\right|\right]\end{aligned}$$

for $t = u - u_0$. Note that $f(x;\theta+u) = f(x;\theta+u_0) + [f(x;\theta+u_0+t) - f(x;\theta+u_0)]$. Therefore, we obtain by Markov inequality

$$\begin{aligned}\mathsf{E}_\theta[\sup_\Gamma R_n(u)] \le \prod_{i=1}^n \Big[&\int_x f^{\frac{1}{2}}(x_i;\theta) f^{\frac{1}{2}}(x_i;\theta+u_0)\nu(dx) \\ &+ \int_x \sup_{\|t\|\le\delta}\left|f^{\frac{1}{2}}(x_i;\theta+u_0+t) - f^{\frac{1}{2}}(x_i;\theta+u_0)\right| f^{\frac{1}{2}}(x_i;\theta)\nu(dx)\Big]\end{aligned}$$

Now, it is easy to see that

$$\begin{aligned}&\int_x f^{\frac{1}{2}}(x_i;\theta) f^{\frac{1}{2}}(x_i;\theta+u_0)\nu(dx) \\ &\quad = \frac{1}{2}\left\{\int_x f(x;\theta)d\nu + \int_x f(x;\theta+u_0)d\nu\right. \\ &\qquad \left. - \int_x \left[f^{\frac{1}{2}}(x;\theta+u_0) - f^{\frac{1}{2}}(x;\theta)\right]^2 d\nu\right\} \\ &\quad = \frac{1}{2}[2 - r_2^2(\theta,\theta+u_0)] \\ &\quad \le 1 - \frac{\kappa_\theta\left(\frac{\gamma}{2}\right)}{2}\end{aligned}$$

On the other hand, using the Cauchy-Schwartz inequality

$$\int_x \sup_{\|t\|\leq\delta} \left| f^{\frac{1}{2}}(x_i;\theta+u_0+t) - f^{\frac{1}{2}}(x_i;\theta+u_0) \right| f^{\frac{1}{2}}(x_i;\theta) d\nu \leq \omega_{\theta+u_0}(\delta)$$

using condition 2. Taking this into account together with the elementary inequality $1+a \leq e^a$ for $a \in \mathbb{R}$, we obtain

$$\mathsf{E}_\theta \left[\sup_{\Gamma} R_n^{\frac{1}{2}}(u) \right] \leq \exp\left\{ -n \left[\frac{\kappa_\theta\left(\frac{\gamma}{2}\right)}{2} - \omega_{\theta+u_0}(\delta) \right] \right\}. \tag{2.45}$$

It follows from (2.45) that to each point ξ of the set $\overline{\mathcal{U}} \setminus \{\|u\| \leq \gamma\}$ there corresponds a sphere $\Gamma(\xi)$ with center ξ such that $\mathsf{E}_\theta \left[\sup_{\Gamma(\xi)} R_n^{\frac{1}{2}}(u) \right] \to 0$ in $\mathsf{P}_\theta-$probability as $n \to \infty$. Using compactness of $\overline{\Theta}$, select a finite cover $\Gamma(\xi_q)$ for $q = 1, 2, \ldots, N$ of the set $\overline{\mathcal{U}} \setminus \{\|u\| \leq \gamma\}$ from the collection $\{\Gamma(\xi)\}$. Then

$$\sup_{\|u\|\geq\gamma} R_n^{\frac{1}{2}}(u) \leq \sum_{q=1}^{N} \sup_{\Gamma(\xi_q)} R_n^{\frac{1}{2}}(u) \xrightarrow{n\to\infty} 0 \text{ in } \mathsf{P}_\theta.$$

Since each of the expectation of each term on the right hand side tends to zero, the proof is complete.[17] □

Proposition 2.34 (Asymptotic Normality of the MLE)

Let $X_i \overset{\text{iid}}{\sim} f_\theta$ where $f \in C^3$, then[18]

$$\frac{\partial}{\partial t} \log f_\theta(x_i) = \frac{\partial}{\partial \theta} \log f_{\hat{\theta}}(x_i) + (\theta - \hat{\theta}) \frac{\partial^2}{\partial \theta^2} f_{\hat{\theta}}(x_i + \epsilon_i$$

and since $\hat{\theta}$ is the solution of the first derivative of the likelihood, we therefore have

$$\sum_{i=1}^{n} \frac{\partial}{\partial t} \log f_\theta(x_i) = 0 + (\hat{\theta} - \theta) \left[-\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f_{\hat{\theta}}(x_i) \right] + \varepsilon_n$$

[17] Note that κ is positive and ω goes to zero, so $-n\kappa$ is negative and $e^{-n\kappa} \to 0$ as $n \to \infty$.

[18] C^3 stands for three times continuously differentiable functions space.

where the left hand side is the so-called score function and thus

$$\frac{\sum_{i=1}^n \frac{\partial}{\partial t} \log f_\theta(x_i)}{\sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log f_{\hat\theta}(x_i)} = (\hat\theta - \theta) + \frac{\varepsilon_n}{\sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log f_{\hat\theta}(x_i)}.$$

This then implies that

$$\begin{aligned}
&\frac{1}{\sqrt{\sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log f_{\hat\theta}(x_i) n}} \cdot \frac{\sum_{i=1}^n \frac{\partial}{\partial t} \log f_\theta(x_i)/n}{\sqrt{\sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log f_{\hat\theta}(x_i) n}} / \sqrt{n} \\
&= A \cdot B \\
&= \sqrt{n}(\hat\theta - \theta) + \frac{\sqrt{n}\varepsilon_n}{\sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log f_{\hat\theta}(x_i)}.
\end{aligned}$$

We now note using WLLN that $a \xrightarrow{\mathsf{P}} \left(-\mathsf{E}\left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(x)\right]\right)^{-\frac{1}{2}}$ and using the CLT that $B \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$. The second term on the right hand side is equal to

$$\frac{\varepsilon_n}{\sqrt{n}} \cdot \left(-\sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log f_{\hat\theta}(x_i)/n\right) \to \frac{\varepsilon_n}{\sqrt{n}} \mathcal{I}(\theta)^{-1}.$$

It then suffices to show that $\frac{\varepsilon_n}{\sqrt{n}} \xrightarrow{\mathsf{P}} \mathcal{O}_p$ or equivalently $\varepsilon_n = \mathcal{O}_p(\sqrt{n})$.

◇◇◇

Section 2.9. Invariance property of MLE

Let $\{f_\theta : \theta \in \Theta\}$ be a family of PDF's (PMF's) and let $\mathcal{L}(\theta)$ be the likelihood function. Suppose that $\Theta \subset \mathbb{R}^k$, for $k \geq 1$[19].

Before we present Zehna's theorem, here is the foreword of the original article that illustrate the need to develop results for maps that are not injective.

> *One of the distinguishing features of the method of maximum likelihood in statistical estimation is the fact that it enjoys a cer-*

[19] This precision is important as we usually don't have this property in ∞-dimension

tain invariance property. likelihood estimator for $u(\hat{\theta})$ where u is some function of $u(\theta)$. Some textbooks on the subject avoid any explicit mention of properties that u must possess in order for invariance to hold. When a proof of the property is given, it is at least assumed, either explicitly or implicitly that u is 1-1 thereby defining a unique inverse.

Now if the assumption that u be 1-1 is really necessary, then the invariance principle could not be invoked to find the maximum likelihood estimator for even as common a case as the variance, $p(1-p)$, of a Bernoulli random variable.

Indeed, there may be some doubt as to the meaning of maximum likelihood in such a case. The purpose of this note is to point out that the notion of a maximum likelihood estimator for $u(\theta)$ when u is not 1-1 can and should be made explicit.

The method used for accomplishing this task has the desirable feature that it coincides with the usual method employed when u is 1-1.

Theorem 2.35 (Zehna's theorem)

Let $h : \Theta \to \Lambda$ be a mapping of Θ onto Λ. If $\hat{\theta}$ is an MLE of θ, then $h(\hat{\theta})$ is the MLE of $h(\theta)$.

PROOF For each $\lambda \in \Lambda$, let us define

$$\Theta_\lambda = \{\theta : \theta \in \Theta, h(\theta) = \lambda\} \text{ and } M(\lambda; \boldsymbol{x}) = \sup_{\theta \in \Theta_\lambda} \mathcal{L}(\theta, \boldsymbol{x})$$

Then M defined on Λ is called the likelihood function induced by h.

Remark

One might legitimately ask why we define it with the sup instead of the inf. Think about it and look at consistency.

If $\hat{\theta}$ is any MLE of θ, then $\hat{\theta}$ belongs to one and only one set $\Theta_{\hat{\lambda}}$ say. Since

$\hat{\theta} \in \Theta_{\hat{\lambda}}, \hat{\lambda} = h(\hat{\theta})$. Now

$$M(\hat{\lambda}, \boldsymbol{x}) = \sup_{\theta \in \Theta_{\hat{\lambda}}} \mathcal{L}(\theta, \boldsymbol{x}) \geq \mathcal{L}(\hat{\theta}, \boldsymbol{x})$$

On the other hand

$$M(\hat{\lambda}, \boldsymbol{x}) \leq \sup_{\lambda \in \Lambda} m(\lambda, \boldsymbol{x}) = \sup_{\theta \in \Theta} \mathcal{L}(\theta, \boldsymbol{x}) = \mathcal{L}(\hat{\theta}, \boldsymbol{x})$$

Thus

$$M(\hat{\lambda}, \boldsymbol{x}) = \sup_{\lambda \in \Lambda} M(\lambda, \boldsymbol{x})$$

It then follows that $\hat{\lambda} \leq h(\hat{\theta})$ is the MLE of $h(\theta)$. □

Chapter 3
Testing Statistical Hypothesis (TSM) and confidence intervals

In this section, we will develop the theory of hypothesis testing, which allows one to check the validity of assertions, build confidence intervals and check validities of models. We will define $H_0 : \theta \in \Theta_0$ to be the null hypothesis, corresponding to the *status quo*. We could think for example for a crime that the null hypothesis is innocence, against the alternative hypothesis (guilt in this case), which we denote by $H_A : \theta \in \Theta_A$. We will want to build tests that allow the rejection of the null; however, we will also want this test to prevent from rejecting the hypothesis.

Section 3.1. Hypothesis tests

Let $X_1, \ldots, X_n$ be $f_\theta, \theta \in \Theta$, where $\Theta_0 \cup \Theta_A = \Theta$ and $\Theta_0 \cap \Theta_A = \emptyset$.

Definition 3.1 (Parametric hypothesis)

A parametric hypothesis is an assertion about the unknown parameter θ.

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \qquad \text{null hypothesis} \\ H_A &: \theta \in \Theta_A \qquad \text{alternative hypothesis} \end{aligned}$$

We set up our problem so that we don't want to make a false call on H_0. For instance, if H_0 is the hypothesis that we bury someone only if the person is dead, we don't want to give that up easily and falsely condemn an innocent person.

Given a sample point $\boldsymbol{x}$, we want to find a decision rule that will lead to the rejection or failure to reject the null hypothesis. It is important to notice that failing to reject does not guarantee that H_0 be true, but rather that no strong enough evidence against the alternative are present.

If $\Theta_0(\Theta_1)$ contains only one point we say $\Theta_0(\Theta_1)$ is a simple hypothesis otherwise the hypothesis is said to be composite. Notice that if $\Theta_0(\Theta_1)$ is

		True	
		H_0	H_1
Accepted	H_0	Correct	Type II error
	H_1	Type I error	Correct

Table 1: Decision table for hypothesis test errors

simple, then the distribution of the observation is completely specified under $\Theta_0(\Theta_1)$.

Example 3.1

Say X_i are iid $\mathcal{N}(\mu, \sigma^2)$. Then $H_0 : \mu = 0, \sigma^2 = 2$ is a simple hypothesis, while $H_0 : \mu = 0$ is composite.

Definition 3.2

Let X be f_θ, $\theta \in \Theta$. a subset C of $\mathbb{R}_n$ such that if $\boldsymbol{x} \in C$ then H_0 is rejected (with probability 1) is called the critical region (set)

$$C = \{x \in \mathbb{R}_n : H_0 \text{ is rejected}\}$$

Definition 3.3 (Test function)

Every "nice" function $\varphi : \mathbb{R}_n \to [0, 1]$ is known as a test function (probability of rejection). φ is a test of hypothesis $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ at α level of significance or (we simply say at level α) if

$$\mathsf{E}_\theta[\varphi(X)] \leq \alpha, \forall \theta \in \Theta_0$$

in short, φ is a test for the problem (α, H_0, H_1).

If $\varphi : \mathbb{R}_n \to \{0, 1\}$ then φ is called a non-randomized test. Otherwise, it is a randomized test.

Let φ be a non-randomized test. Then

$$\mathsf{E}_\theta[\varphi(X)] = P_\theta[\varphi(X) = 1]$$

and "$\varphi(x) = 1$" means rejecting with probability 1.

A few remarks before we go on with other definitions. Note that we want often our null hypothesis to be specific, so that it can lead to rejection in some cases. For example, we may be interested in the context of linear regression by verifying whether coefficients given from least squares is significantly different from zero. Note that an hypothesis is a conjecture about a population parameter rather than on a sample which is observed. After stating the null and hypothesis alternatives, we will want to determine the significance (the power of the test).

Definition 3.4 (Power function)

We define

$$\beta_\varphi(\theta) = \mathsf{E}_\theta[\varphi(\boldsymbol{X})] = \mathsf{P}_\theta\{ \text{ reject } H_0\},$$

where β_φ is called the power function of φ. Then, the probability of Type II error is

$$\mathsf{P}_\theta\{ \text{ Type II error } \} = 1 - \beta_\varphi(\theta),$$

for $\theta \in \Theta_1$,

$$1 - \beta_\varphi(\theta) = 1 - \mathsf{P}_\theta\{ \text{ reject } H_0\} = \mathsf{P}_\theta\{ \text{ accept } H_0\}.$$

In other words, the power of a test is the probability that it will correctly lead to the rejection of a false null hypothesis.

Definition 3.5 (Most Powerful tests)

Let Φ_α be the class of all tests for the problem $(\alpha, \Theta_0, \Theta_1)$. A test $\varphi_0 \in \Phi_\alpha$ is said to the most powerful (MP) test against an alternative $\theta^* \in \Theta$; if

$$\beta_{\varphi_0}(\theta^*) \geq \beta_\varphi(\theta^*), \forall \varphi \in \Phi_\alpha$$

Definition 3.6 (Uniformly Most Powerful tests)

A test $\varphi_0 \in \Phi_\alpha$ for the problem (α, H_0, H_1) is said to be uniformly most powerful (UMP) if

$$\beta_{\varphi_0}(\theta) \geq \beta_\varphi(\theta), \forall \varphi \in \Phi_\alpha, \forall \theta \in \Theta_1$$

Example 3.2

$\Theta = \{\theta_0.\theta_1\}, H_0 : \theta = \theta_0, H_1 : \theta = \theta_1$. Let x_i be iid $f_\theta(x)$ and

$$\frac{\mathcal{L}(\theta_0, \boldsymbol{x})}{\mathcal{L}(\theta_1, \boldsymbol{x})} > 1$$

Often, we cannot attain the size α chosen, which is arbitrarily chosen. A common practise in the literature is to report the P-value, which is the smallest level α at which the sample statistic observed is significant. Formally, we define the P-value as

Definition 3.7 (P-value)

The probability of observing under H_0 a sample outcome at least as extreme as the one observed is called the P-value. The smaller the P-value, the more extreme the outcome and the stronger the evidence against H_0.

If α is given, we reject H_0 if $P \leq \alpha$ and do not reject H_0 if $P > \alpha$. In the two-sided case when the critical region is of the form $C = \{|T(\boldsymbol{x}) > k\}$, the one-sided P-value is doubled to obtain the P-value. If the distribution of T is not symmetric, the P-value is not well defined in the two-sided case, although many authors recommend doubling the one-sided P-value.

We now present a famous result due to Neyman and Pearson that gives a general recipe for Most Powerful tests.

Theorem 3.8 (Neyman-Pearson fundamental lemma)

1. Any test φ of the form

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{f_1(x)}{f_0(x)} > k \\ \gamma(x) & \text{if } \frac{f_1(x)}{f_0(x)} = k \\ 0 & \text{if } \frac{f_1(x)}{f_0(x)} < k \end{cases}$$

for some $k \geq 0$ and $0 \leq \gamma(\boldsymbol{x}) \leq 1$ is the most powerful of its size for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. If $k = \infty$, the test

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } f_0(\boldsymbol{x}) \leq 0 \\ 0 & \text{if } f_0(\boldsymbol{x}) > 0 \end{cases}$$

is the most powerful of size 0 for testing H_0 versus H_1.

2. Given $\alpha \in [0,1]$, there exists a test of form 1) or 2) with $\varphi(\boldsymbol{x}) \leq \gamma$ (a constant) for which $\mathsf{E}_\theta[\varphi(X)] = \alpha$.

PROOF We first begin with part a. Let φ be a test satisfying (1) and φ^* be any test with $\mathsf{E}_{\theta_0}[\varphi^*(X)] \leq \mathsf{E}_{\theta_0}[\varphi(X)]$. If the distributions are absolutely continuous, the equality holds with probability zero, so we can ignore the middle term of (1). Recall that $\varphi(x) = 1$ if $f_1(x)/f_0(x) > k$ is the probability of rejection.

We will assume that we are in the continuous case (the discrete case can be treated likewise). We have

$$\begin{aligned} A &= \int_x [\varphi(x) - \varphi^*(x)][f_1(x) - kf_0(x)]dx \\ &= \left(\int_{x:f_1(x)>kf_0(x)} + \int_{x:f_1(x)<kf_0(x)} \right) [\varphi(x) - \varphi^*(x)][f_1(x) - kf_0(x)]dx \end{aligned}$$

For any $x \in \{x : f_1(x) > kf_0(x)\}$, $\varphi(x) - \varphi^*(x) = 1 - \varphi(x) \geq 0$ so that the integrand is positive. For $x \in \{x : f_1(x) < kf_0(x)\}$, $\varphi(x) - \varphi^*(x) = -\varphi(x) \leq 0$ so that the integrand is again ≥ 0. It follows that

$$A = \mathsf{E}_{\theta_1}[\varphi(X)] - \mathsf{E}_{\theta_1}[\varphi^*(X)] - k\left\{\mathsf{E}_{\theta_0}[\varphi(X)] - \mathsf{E}_{\theta_0}[\varphi^*(X)]\right\} \geq 0$$

which in turns implies that $\mathsf{E}_{\theta_1}[\varphi(X)] - \mathsf{E}_{\theta_1}[\varphi^*(X)] \geq k\mathsf{E}_{\theta_0}[\varphi(X)] - \mathsf{E}_{\theta_0}[\varphi^*(X)] \geq 0$ and therefore $\varphi(x)$ is more powerful as $\mathsf{E}_{\theta_1}[\varphi(X)] \geq \mathsf{E}_{\theta_1}[\varphi^*(X)]$.

If $k = \infty$, any test φ^* of size 0 must vanish on the set $\{x : f_0(x) > 0\}$. We have

$$\mathsf{E}_{\theta_1}[\varphi(X)] - \mathsf{E}_{\theta_1}[\varphi^*(X)] = \int_{x:f_0(x)=0} [1 - \varphi^*(x)]f_1(x)dx \geq 0$$

□

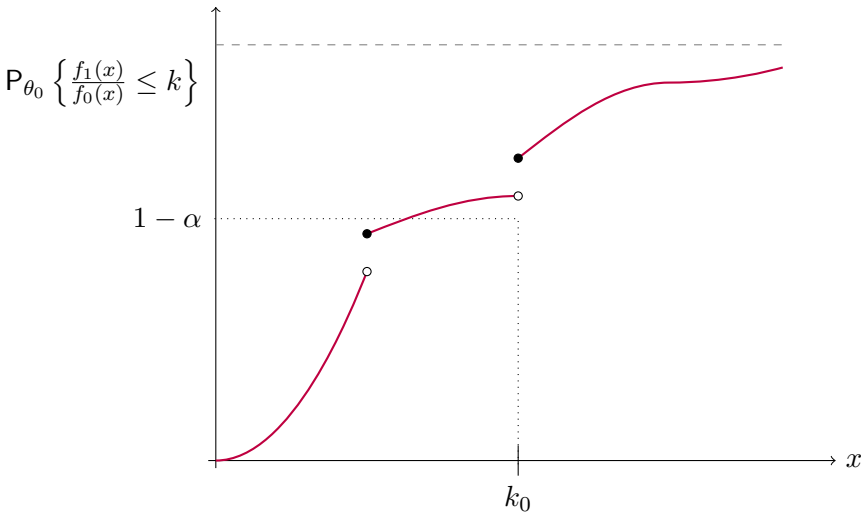


Figure 9: Choosing a value of k_0 for a MP test.

Remark

We look at tests of the same size (that is, they are optimal over tests of same size) as φ. If $k = \infty$, the size of $\mathsf{E}_{\theta_0}(\varphi) = 0$ so the term vanishes. Think of it and write the definition of $\mathsf{E}_{\theta_0}\varphi^*(x) = \int_x \varphi^*(x) f_0(x) dx = 0$ since $\varphi^*(x) \geq 0, f_0(x) \geq 0$ almost surely under P_{θ_0}. We can easily extend this result to more than two categories (finite number).

We now prove the second part of the Neyman-Pearson lemma.

Proof We confine ourselves to the case $0 < \alpha \leq 1$. Let $\gamma(x) = \gamma$. The size of a test of form (1) is

$$\begin{aligned}\mathsf{E}_{\theta_0}[\varphi(X)] &= \mathsf{P}_{\theta_0}\{f_1(X) > kf_0(X)\} + \gamma \mathsf{P}_{\theta_0}\{f_1(X) = kf_0(X)\} \\ &= 1 - \mathsf{P}_{\theta_0}\{f_1(X) \leq kf_0(X)\} + \gamma \mathsf{P}_{\theta_0}\{f_1(X) = kf_0(X)\}\end{aligned}$$

and since $\mathsf{P}_{\theta_0}\{f_0(X)=0\}=0$[20], we may write

$$\mathsf{E}_{\theta_0}[\varphi(X)] = 1 - \mathsf{P}_{\theta_0}\left\{\frac{f_1(X)}{f_0(X)} \leq k\right\} + \gamma \mathsf{P}_{\theta_0}\left\{\frac{f_1(X)}{f_0(X)} = k\right\}.$$

Given $0 < \alpha \leq 1$, we wish to find k and γ such that $\mathsf{E}_{\theta_0}[\varphi(X)] = \alpha$, that is

$$\mathsf{P}_{\theta_0}\left\{\frac{f_1(X)}{f_0(X)} \leq k\right\} - \gamma \mathsf{P}_{\theta_0}\left\{\frac{f_1(X)}{f_0(X)} = k\right\} = 1 - \alpha. \tag{3.46}$$

Note that $\mathsf{P}_{\theta_0}\left\{\frac{f_1(X)}{f_0(X)} \leq k\right\}$ is a CDF and hence it is a non-decreasing and right-continuous function of k. If there exists a k_0 such that $\mathsf{P}_{\theta_0}\left\{\frac{f_1(X)}{f_0(X)} \leq k\right\} = 1-\alpha$, then we choose $\gamma = 0$ and $k = k_0$. Otherwise, there exists a k_0 such that

$$\mathsf{P}_{\theta_0}\left\{\frac{f_1(X)}{f_0(X)} < k\right\} \leq 1 - \alpha < \mathsf{P}_{\theta_0}\left\{\frac{f_1(X)}{f_0(X)} \leq k\right\},$$

that is [21] there is a jump at k_0. In this case, we choose $k = k_0$ and solve (3.46) for γ, and

$$\gamma = \frac{\mathsf{P}_{\theta_0}\left\{\frac{f_1(X)}{f_0(X)} \leq k\right\} - (1-\alpha)}{\mathsf{P}_{\theta_0}\left\{\frac{f_1(X)}{f_0(X)} = k_0\right\}}.$$

□

Theorem 3.9

If a sufficient statistic T exists for the family $\{f_\theta : \theta \in \Theta\}, \Theta = \{\theta_0, \theta_1\}$, the Neyman-Pearson MP test is a function of T. Indeed,

$$\frac{f_1(x)}{f_0(x)} = \frac{h(x)g_1(T(x))}{h(x)g_0(T(x))} = \frac{g_1(T(x))}{g_0(T(x))}.$$

This is in agreement with the sufficient principle.

We now present some examples that illustrate the MP tests

[20] This is true since for $A = \{x : f_0(x) = 0\}$, then $\mathsf{P}_{t_0}(A) = \int_A f_0(x)dx = 0$.

[21] This doesn't happen if we are in the absolutely continuous case. We have continuous cases which are not and the measure induced by the variable for example $f(x) = \frac{1}{2}$ if $x = 0$, $\frac{1}{2}\mathcal{U}(0,1)$ if $0 < x < 1$ is not dominated by Lebesgue measure.

Example 3.3

Suppose that we want to test the hypothesis $H_0 : X \sim \mathcal{N}(0,1)$ versus $H_1 : X \sim \mathcal{C}(0,1)$. According to Neyman-Pearson, we only need to look at the ratio

$$\frac{f_1(x)}{f_0(x)} = \frac{\pi^{-1}(1+x^2)^{-1}}{(2\pi)^{-\frac{1}{2}}\exp\left(-\frac{x^2}{2}\right)} = \frac{\sqrt{2}}{\sqrt{\pi}}\frac{\exp\left(\frac{x^2}{2}\right)}{1+x^2}.$$

The MP test is of the form

$$\varphi(x) = \begin{cases} 1 & \text{if } \sqrt{\frac{2}{\pi}}\frac{\exp(x^2/2)}{1+x^2} > k \\ 0 & \text{otherwise} \end{cases}$$

where k is determined so that $\mathsf{E}_{\theta_0}[\varphi(X)] = \alpha$. Note that $f_1(x)/f_0(x)$ is a non-decreasing [22] function of $|x|$. It then follows that

$$\varphi(x) = \begin{cases} 1 & \text{if } |x| > k_1 \\ 0 & \text{if } |x| < k_1 \end{cases}.$$

Thus,

$$\mathsf{P}_{\theta_0}\left(\sqrt{\frac{2}{\pi}}\frac{\exp(x^2/2)}{1+x^2} > k\right) = \alpha,$$

knowing that $X \sim \mathcal{N}(0,1)$. Finding the distribution of this creature is not easy. It is much easier to see $\mathsf{P}_{\theta_0}\{|X| > k_1\} = \alpha \Rightarrow 2\mathsf{P}_{\theta_0}\{x > k_1\} = \alpha$; we only need to go to the table and find the value of k for $\alpha/2$. If we have more than one sample, the problem is harder to do. We can reduce the form, but even then problems arise. In such cases, we can try simulations (which is cheap). We generated data which is plugged in the distribution and look at the histogram with the quantiles.

Next example is a bit more involved

Example 3.4

Suppose we have a $n-$sample of draws from Bernoulli distribution where $X_i \overset{\text{iid}}{\sim} \mathcal{B}(p)$ and we test $H_0 : p < p_0$ vs $H_1 : p = p_1$ where $p_1 > p_0$. The MP

[22] Looking at the derivative or using the fact that for a one sample, the observation is sufficient

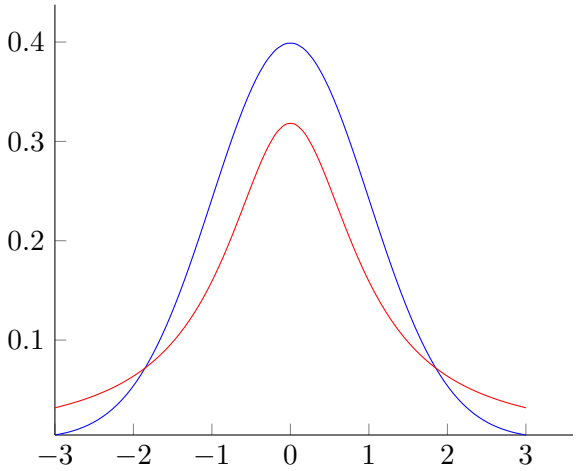


Figure 10: Test for the identification of the random variable distribution Comparison between Cauchy density function versus Normal density.

test in this case is

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \lambda(\boldsymbol{x}) = \frac{p_1^t(1-p_1)^{n-t}}{p_0^t(1-p_0)^{n-t}} > k \\ \gamma & \text{if } \lambda(\boldsymbol{x}) = k \\ 0 & \text{if } \lambda(\boldsymbol{x}) < k \end{cases}$$

where k and γ are determined from $\mathsf{E}_{\theta_0}[\varphi(X)] = \alpha$. Now, $\lambda(\boldsymbol{x}) = \left(\frac{p_1}{p_0}\right)^t \left(\frac{1-p_1}{1-p_0}\right)^{n-t}$ and since $p_1 > p_0$, $\lambda(\boldsymbol{x})$ is an increasing function of t. We can see this by differentiating the function (taking logarithms first). It follows that $\lambda(\boldsymbol{x}) > k$ if and only if $t > k_1$ for some constant k_1. Then

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } t > k_1 \\ \gamma & \text{if } t = k_1 \\ 0 & \text{if } t < k_1 \end{cases}$$

and

$$\alpha = \mathsf{E}_{\theta_0}[\varphi(\boldsymbol{x})] = \mathsf{P}_{\theta_0}\{T > k_1\} + \gamma \mathsf{P}_{\theta_0}\{T = k_1\}$$

where $T = \sum_{i=1}^{n} X_i \sim \mathcal{B}(n, p)$. The above is equal to

$$= \sum_{r=k_1+1}^{n} \binom{n}{r} p_0^r (1-p_0)^{n-r} + \gamma \binom{n}{k_1} p_0^{k_1} (1-p_0)^{n-k_1}$$

which can be solved using the table for the CDF of the binomial distribution. Note that the most powerful size α test is independent of p_1 as long as $p_1 > p_0$, that is the test remains most powerful of size α against $p > p_0$ and is therefore a uniformly most powerful test (or UMP) test of null hypothesis $p = p_0$ against the alternative $p > p_0$.

Section 3.2. Families with the Monotone Likelihood Ratio Property

Definition 3.10 (Monotone likelihood ratio property)

Let $\{f_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}$ be a family of PDF's (PMF's). We say that $\{f_\theta\}$ has an MLR property $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ in $T(\boldsymbol{x})$ if for $\theta_1 < \theta_2$, whenever $f_{\theta_1}, f_{\theta_2}$ are distinct, the ratio $f_{\theta_2}(\boldsymbol{x})/f_{\theta_1}(\boldsymbol{x})$ is a non-decreasing function of $T(\boldsymbol{x})$ for the set of values $\boldsymbol{x}$ for which at least one of f_{θ_1} and f_{θ_2} is greater than 0.

Example 3.5

Let X_i be iid $\mathcal{U}(0, \theta), \theta > 0$. Note that for the purpose of this example, we will inverse the roles of θ_1 and θ_2. The likelihood function is

$$\mathcal{L}(\theta, \boldsymbol{x}) = f(\boldsymbol{x}) = \begin{cases} \frac{1}{\theta^n} & \text{if } 0 < \max x_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

If $\theta_1 > \theta_2$,

$$\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} = \frac{\frac{1}{\theta_1^n} \mathbf{1} \max_{(0,\theta_1)} x_i}{\frac{1}{\theta_2^n} \mathbf{1} \max_{(0,\theta_2) x_i}} = \left(\frac{\theta_2}{\theta_1}\right)^n \frac{\mathbf{1} \max_{(0,\theta_1)} x_i}{\mathbf{1} \max_{(0,\theta_2) x_i}}$$

Thus

$$R(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \max x_i \in [0, \theta_2] \\ \infty & \text{if } \max x_i \in (\theta_2, \theta_1] \end{cases}$$

We adopted this usage of θ_1, θ_2 to have the bigger term in the nominator and the smaller term in the denominator.

Then the family of uniform $\mathcal{U}(0, \theta), \theta > 0$ has the MLR property in $\max x_i$. We have a general result which tells us more about the class of distributions which have this property.

Theorem 3.11

The exponential 1-parameter family $f_\theta(\boldsymbol{x}) = \exp\{Q(\theta)T(\boldsymbol{x}) + S(\boldsymbol{x}) + D(\theta)\}$ where $Q(\theta)$ is a non-decreasing function of θ has the MLR property in $T(\boldsymbol{x})$.

PROOF Left as an exercise: choose 2 values with $\theta_2 > \theta_1$ and look at

$$\frac{f_{\theta_2}(\boldsymbol{x})}{f_{\theta_1}(\boldsymbol{x})} = \exp\{[Q(\theta_2) - Q(\theta_1)]T(\boldsymbol{x}) + D(\theta_2) - D(\theta_1)$$

and take the derivative to show that the ratio is non-decreasing. □

Example 3.6

Let $X_i \overset{\text{iid}}{\sim} \mathcal{B}(m, p)$, $i = 1, \ldots, n$. Then the likelihood function is

$$\begin{aligned}
\mathcal{L}(p, \boldsymbol{x}) &= \prod_{i=1}^{n} \binom{m}{x_i} p^{x_i}(1-p)^{m-x_i} \\
&= \prod_{i=1}^{n} \binom{m}{x_i} p^{\Sigma_{i=1}^n x_i}(1-p)^{mn-\Sigma_{i=1}^n x_i} \\
&= \prod_{i=1}^{n} \binom{m}{x_i} \left(\frac{p}{1-p}\right)^t (1-p)^{mn}
\end{aligned}$$

where $T(\boldsymbol{x}) = \sum_{i=1}^n X_i, \sum_{i=1}^n x_i = t$ and $\theta = \frac{p}{1-p}$.

$$\exp\left\{\log\theta t + \sum_{i=1}^{n} \log\binom{m}{x_i} + mn\log(1-p)\right\}$$

and so we have $Q(\theta) = \log\theta; T(\boldsymbol{x}) = \sum_{i=1}^n \log\binom{m}{x_i}; D(\theta) = mn\log\left(\frac{1}{1+\theta}\right)$. Since log is a monotone function, we get a non-decreasing function and the result holds.

Theorem 3.12

Let $X \sim f_\theta, \theta \in \Theta$, where $\{f_\theta\}$ as the MLR property in $T(\boldsymbol{x})$ for testing

$$\begin{cases} H_0 & \theta \leq \theta_0 \\ H_1 & \theta > \theta_1 \end{cases}$$

where $\theta_0 \in \Theta$. Any test of the form

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } T(\boldsymbol{x}) > t_0 \\ \gamma & \text{if } T(\boldsymbol{x}) = t_0 \\ 0 & \text{if } T(\boldsymbol{x}) < t_0 \end{cases}$$

has non-decreasing power function and is the UMP test of its size $\mathsf{E}_\theta\, \varphi(\boldsymbol{x}) = \alpha$ provided that $\alpha \neq 0$.

For the proof, we know how to test $\theta = \theta_0$ vs $\theta \neq \theta_0$. The likelihood ratio being a monotone function, then since the MLR has a power function which is monotone, then MLR has monotone power function, that is $\mathsf{E}_{\theta_0}\varphi(\boldsymbol{x})$ is a power function. The supremum is the same for other values for H_1, since finding t_0 and γ does not depend on the value of θ_0. In general, there are no UMP test for

$$\begin{cases} H_0 & : \theta \in (a, b) \\ H_1 & : \theta \notin (a, b) \end{cases} \quad \text{or} \quad \begin{cases} H_0 & : \theta = \theta_0 \\ H_1 & : \theta \neq \theta_0 \end{cases}$$

In many cases, the UMP test may not even exist; this is why we need to develop new approaches. We had earlier the likelihood ratio for comparison; can we develop something similar, maybe the ratio of integrals. The criterion should be meaningful: the ratio means large values are good for the null hypothesis and this should be reflected by our test. Ratio of integrals $\int_{\theta_0} f_\theta(x) / \int_{\theta_1} f_\theta(x)$ yields some problems. Fisher developed another approach, using $\mathcal{N}(\mu, \sigma^2)$ as a motivating example. How would we construct tests to verify the two previous or other similar situations? No unique answer

exists, but here is one:

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\theta\in\Theta_0} \mathcal{L}(\theta, \boldsymbol{x})}{\sup_{\theta\in\Theta} \mathcal{L}(\theta, \boldsymbol{x})} \leq 1$$

Large values of $\lambda(\boldsymbol{x})$ are good for H_0 and small values are bad for H_0. How small is the value for a given size $\alpha = \mathsf{P}_{H_0}(\lambda(\boldsymbol{x}) \leq C)$ is a question we will need to take care of. In general, its is not easy to find the exact distribution of $\lambda(\boldsymbol{x})$; we need to rely on large-sample theory and on the fact that $-2\log\lambda(\boldsymbol{x}) \to \chi^2(\nu)$ where $\nu = \dim\Theta - \dim\Theta_0$. The problem with the integrals is that we would recover the average, and that the numerator's average could be higher. It may be hard then to find directions for acceptance or rejection.

We reject H_0 if $\lambda(\boldsymbol{x}) \leq c$ otherwise we fail to do so.

Example 3.7

Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and suppose that we wish to test

$$\begin{cases} H_0 & \mu = \mu_0 \\ H_1 & \mu \neq \mu_0 \end{cases}; \qquad \Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^2\}$$

and notice that both are composite hypothesis. We present the calculations step by step for finding the ratio of MLE.

Step 1: Calculate the likelihood function:

$$\begin{aligned} \mathcal{L}(\theta = (\mu, \sigma^2), \boldsymbol{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \end{aligned}$$

Step 2: Find the supremum on Θ:

$$\sup_{\theta\in\Theta} \mathcal{L}(\theta, \boldsymbol{x}) = \mathcal{L}((\bar{x}, \widehat{\sigma^2}_{\text{ML}}), \boldsymbol{x})$$

where

$$\widehat{\sigma^2}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Step 3: Find the supremum on Θ_0:

$$\sup_{\theta\in\Theta_0} \mathcal{L}(\theta, \boldsymbol{x}) = \mathcal{L}((\mu, \widehat{\sigma^2}_*), \boldsymbol{x})$$

where

$$\widehat{\sigma^2}_* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Step 4: We are now set to calculate $\lambda(\boldsymbol{x})$:

$$\begin{aligned}
\lambda(\boldsymbol{x}) &= \frac{\sup_{\theta\in\Theta_0} \mathcal{L}(\theta, \boldsymbol{x})}{\sup_{\theta\in\Theta} \mathcal{L}(\theta, \boldsymbol{x})} \\
&= \frac{\mathcal{L}((\mu, \hat{\sigma}^2_*), \boldsymbol{x})}{\mathcal{L}((\mu, \hat{\sigma}^2_{\mathrm{ML}}, \boldsymbol{x})} \\
&= \frac{\left(\sqrt{2\pi}\hat{\sigma}_*\right)^{-n} \exp\left\{-\frac{1}{2\hat{\sigma}^2_*}\sum_1^n (x_i - \mu_0)^2\right\}}{\left(\sqrt{2\pi}\hat{\sigma}_{\mathrm{ML}}\right)^{-n} \exp\left\{-\frac{1}{2\hat{\sigma}^2_{\mathrm{ML}}}\sum_1^n (x_i - \bar{x}_0)^2\right\}}
\end{aligned}$$

and noticing that $\widehat{\sigma^2}_* \sum_{i=1}^n (x_i - \mu_0)^2 = n$ and similarly for the denom-

inator, we have

$$\begin{aligned}
\lambda(\boldsymbol{x}) &= \left(\frac{\hat{\sigma}_{\mathrm{ML}}}{\hat{\sigma}_*}\right)^n \frac{\exp\left(-\frac{n}{2}\right)}{\exp\left(-\frac{n}{2}\right)} \\
&= \left(\frac{\hat{\sigma}_{\mathrm{ML}}}{\hat{\sigma}_*}\right)^n \\
&= \left(\frac{\hat{\sigma}^2_{\mathrm{ML}}}{\hat{\sigma}^2_*}\right)^{\frac{n}{2}} \\
&= \left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \mu_0)^2}\right)^{\frac{n}{2}} \\
&= \left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}\right)^{\frac{n}{2}} \\
&= \left(1 + \left(\frac{1}{n-1}\right) \frac{n(\bar{x} - \mu_0)^2}{\left(\frac{1}{n-1}\right) \sum_{i=1}^n (x_i - \bar{x})^2}\right)^{-\frac{n}{2}}
\end{aligned}$$

which is like a t distribution. Let

$$F = \frac{n(\bar{x} - \mu_0)^2}{\left(\frac{1}{n-1}\right) \sum_{i=1}^n (x_i - \bar{x})^2}$$

is a ratio of $\chi^2(1)/\chi^2(n-1)$. The above then reduces to

$$\lambda(\boldsymbol{x}) = \left(\frac{1}{1 + \frac{F}{n-1}}\right)^{-\frac{n}{2}}$$

Note that $\lambda(\boldsymbol{x})$ is a decreasing function of F. Moreover, $\{\lambda(\boldsymbol{x}) \leq c\} \Leftrightarrow \{F(1, n-1) \geq k\}$.

For a given α, we find k such that $\alpha = \mathsf{P}(F(1, n-1) \geq K)$. You mind want to refresh your memory here about the derivation of the F statistic. If

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1), \Rightarrow \qquad \frac{n(\overline{X} - \mu_0)^2}{\sigma^2} \sim \chi^2(1).$$

and also

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})^2/\sigma^2=\frac{(n-1)S^2}{(n-1)\sigma^2}=\frac{\chi^2(n-1)}{n-1};$$

thus the ratio is $F(m,n)=\frac{V/m}{W/n}$ provided $V \perp\!\!\!\perp W$, and that $W\sim\chi^2(m), W\sim\chi^2(n)$.

Theorem 3.13 (Generalized likelihood ratio test)

Suppose we want to test $H_0:\theta\in\Theta_0\subseteq\Theta$ versus $H_1:\Theta_1=\Theta\setminus\Theta_0$ then the generalized likelihood ratio test is

$$\lambda_n(\boldsymbol{x})=\frac{\sup_{\theta\in\Theta_0}\mathcal{L}(\theta;x_1,\ldots,x_n)}{\sup_{\theta\in\Theta}\mathcal{L}(\theta;x_1,\ldots,x_n)}$$

where the rejection region is given by $C=\{\boldsymbol{x}:\lambda_n(\boldsymbol{x})\leq c\}$ where c is chosen such that

$$\sup_{\theta\in\Theta_0}\mathsf{P}_\theta(C)=\alpha. \tag{3.47}$$

We find c such that (3.47) is fulfilled. We see that it reduces to Neyman-Pearson when the two hypothesis are simple. Often times, we cannot do this easily and have to rely on asymptotic[23]:

$$C=\{\boldsymbol{x}:-2\log\lambda_n(\boldsymbol{x})\geq k\}.$$

If $\boldsymbol{\theta}=(\boldsymbol{\eta},\boldsymbol{\xi})$ where $\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\xi}$ are respectively of dimensions $r,r',r-r'$. Then, we may want to test $H_0:\boldsymbol{\eta}=\eta_0$ against $H_1:\boldsymbol{\eta}\neq\eta_0$. In such case, we don't even need H_1 as it is $\Theta\setminus\Theta_0$. We can establish some great results under some regularity conditions[24]. Indeed,

$$-2\log\lambda_n\xrightarrow{\mathcal{D}}\chi^2_{r-r'}$$

Although we don't prove this result, here is nevertheless the essence of the

[23] Note that here the concept of large sample depends heavily on the dimension of parameter to estimate

[24] These regularity conditions resemble that of the Cramer-Rao lower bound.

proof. We can interchange logarithms and supremum since the log is a monotone function. Then, we get the supremum of the log-likelihood. Expanding this value and computing under H_0, we get two terms, one of which is like the derivative and will vanish. The other one is almost normal, and we can normalize properly to get convergence in distribution to the chi-square distribution.

Section 3.3. Confidence intervals

Often times, people are not happy with sharp hypothesis; it can often be too restrictive and we don't have necessarily most powerful tests for them. We would want to look at results of the form $|\boldsymbol{\eta} - \eta_0| < \varepsilon$, but as mentioned no result is as good as those we derived to do that.

$$C^* = \{\boldsymbol{x} : -2 \log \lambda_n(\boldsymbol{x}) \geq \chi^2_{r-r',\alpha}\}$$

where

$$\sup_{\theta \in \Theta_0} \mathsf{P}\left(\chi^2_{r-r'} \geq \chi^2_{r-r',\alpha}\right) = \alpha.$$

If we use this, we can show that $\lim_{n\to\infty} \mathsf{P}_\theta(C*) = 1$ for all $\theta \in \Theta_1$. If the null is incorrect, then the probability of rejecting when the null is false is 1; the power of the test increase to 1. We pursue with a proper

Definition 3.14 (Confidence set)

A family of subsets $S(x)$ of $\Theta \subseteq \mathbb{R}^k$ is said to constitute a family of confidence sets at confidence level $1 - \alpha$ if[25]

$$\mathsf{P}_\theta\{S(x) \supset \theta\} \geq 1 - \alpha \; \forall \; \theta \in \Theta. \tag{3.48}$$

Example 3.8

Suppose we have $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \ldots, n$ and where μ is unknown,

[25] Note here that $S(x)$ is a random set. In the one-dimensional case, this is a set whose endpoints are random variables.

but σ^2 is known. Recall that

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Look now at the centred distribution

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

and let $\alpha = 0.05$. Then $\mathsf{P}(|Z_n| < 1.96) = 0.95$ and so

$$0.95 = \mathsf{P}\left(\left|\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}\right| < 1.96\right) = \mathsf{P}\left(\overline{X}_n - \frac{1.96\sigma}{\sqrt{n}} < \mu < \overline{X}_n + \frac{1.96\sigma}{\sqrt{n}}\right)$$

yields the random interval $(\overline{X}_n - 1.96\sigma/\sqrt{n}, \overline{X}_n + 1.96\sigma/\sqrt{n})$. If σ^2 is unknown, we can use the sample variance and thus the ratio will be distributed according to a Student-t. Note that we relied on the asymptotic of the distribution of ratio and in this case we could easily isolate the parameter of interest- this doesn't always happen.

Definition 3.15 (Pivotal quantity)

If the distribution of $T(\boldsymbol{x}, \boldsymbol{\theta})$ does not depend on any unknown parameter, we say $T(\boldsymbol{x}, \boldsymbol{\theta})$ is a pivotal quantity.

We will now distinguish between the large-sample case and the small sample case, as the two approaches are very different in practise.

Large sample confidence interval

If the X_i's have common mean μ and variance σ^2 and are independent, then

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{n\to\infty} \mathcal{N}(0,1)$$

using the Central Limit Theorem (the statistic is asymptotically pivotal). This holds as long as $S \xrightarrow{\mathsf{P}} \sigma$ and so

$$\left(\overline{X}_n - \frac{\mathfrak{z}_{\alpha/2} S}{\sqrt{n}}, \overline{X}_n + \frac{\mathfrak{z}_{\alpha/2} S}{\sqrt{n}}\right)$$

is an approximate $(1-\alpha)$ confidence interval for μ. We had earlier the exact distribution, we now have the asymptotic one and $\mathsf{P}(Z > \mathfrak{z}_{\alpha/2}) = \alpha/2$. One may ask what the sample size has to be and how large it needs in order to be able to draw valid inference for this. We have a result that links the uniform norm of the normalized parameter with the cumulative distribution function of the normal for given n.

Theorem 3.16 (Berry-Esseen)

We have for $X_i \overset{\text{iid}}{\sim} f$ and sufficient sample size, where $\mathsf{E}X_i^3 < \infty$ the following bound

$$\sup_t \left(|F_n(t) - \Phi(t)|\right) = \|F_n(t) - \Phi(t)\|_\infty \leq \frac{C\rho}{\sigma^3\sqrt{n}} \tag{3.49}$$

for fixed value of n and this allows us to get an idea of n, since this has convergence $\mathcal{O}(1/\sqrt{n})$. Since the skewness is not known, we can rely on the histogram. The best estimate for C is for the moment 0.4785, which is a big improvement over the original upper bound of 7.59 (that dates back to Esseen in 1942.) Esseen showed that the lower bound is given by $C_l \approx 0.4097$.

A second approach for this is to use the asymptotic distribution of $\hat{\theta}_{ML}$ and using the fact that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I^{-1}(\theta))$$

and from this, we can make a pivotal quantity again, with

$$\sqrt{nI(\theta)}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

such that $\mathsf{P}(\sqrt{nI(\theta)}|\hat{\theta}_n - \theta| < \mathfrak{z}_{\alpha/2}) \approx 1 - \alpha$ for large n. This is equivalent to saying

$$\mathsf{P}\left(\hat{\theta}_n - \frac{\mathfrak{z}_{\alpha/2}}{\sqrt{nI(\theta)}} < \theta < \hat{\theta}_n + \frac{\mathfrak{z}_{\alpha/2}}{\sqrt{nI(\theta)}}\right) \approx 1 - \alpha$$

and we often times replace $I(\theta)$ by a consistent estimator. For instance, if $I(\theta)$ is a continuous function of θ, then $I(\hat{\theta}_n)$ is a consistent estimate of $I(\theta)$. Indeed, recall that if $X_n \xrightarrow{\mathsf{p}} X$ and g is continuous, then $g(X_n) \xrightarrow{\mathsf{p}} g(X)$. Also, if we look back at the definition of Fisher's information,

$$I(\theta) = \mathsf{E}\left[\left\{\frac{\partial}{\partial\theta}\log f_\theta(x)\right\}^2\right].$$

If you have a bunch of samples, you can easily estimate this with

$$\hat{I}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial}{\partial\theta}\log f_\theta(x_i)\right\}^2\Bigg|_{\theta=\hat{\theta}}$$

In the assignment, plug $\hat{\theta}$ for θ and using Zehna's theorem, this will be asymptotically best; $\hat{\theta}_n \pm \dfrac{\mathfrak{z}_{\alpha/2}}{\sqrt{n\hat{I}(\theta)}}$ is an approximate $1-\alpha$ confidence interval for θ.

Example 3.9

Let $X_i \overset{\text{iid}}{\sim} \mathcal{B}(1, p)$ and recall that

$$\mathsf{E}\left[\left\{\frac{\partial}{\partial\theta}\log f_\theta(x)\right\}^2\right] = -\mathsf{E}\left\{\frac{\partial^2}{\partial\theta^2}\log f_\theta(x)\right\}$$

under regularity conditions. then

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the MLE of p and

$$\mathsf{P}_p(x) = \mathsf{P}(X = x) = p^x(1-p)^{1-x}, \qquad x = 0, 1.$$

The log-likelihood is $\log \mathsf{P}_p(x) = x \log p + (1-x)\log(1-p)$ and

$$\frac{\partial}{\partial p} \log \mathsf{P}_p(x) = \frac{x}{p} - \frac{1-x}{1-p}.$$

Using the above formula for the information , we compute

$$\frac{\partial^2}{\partial p^2} \log \mathsf{P}_p(x) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

and

$$\begin{aligned} -\mathsf{E}\left[\frac{\partial^2}{\partial p^2} \log \mathsf{P}_p(x)\right] &= \mathsf{E}\left[\frac{x}{p^2} + \frac{1-x}{(1-p)^2}\right] \\ &= \mathsf{E}\frac{x}{p} + \mathsf{E}\frac{1-x}{(1-p)^2} \\ &= \frac{p}{p^2} + \frac{1-p}{(1-p)^2} \\ &= \frac{1}{p(1-p)}. \end{aligned}$$

Therefore, a confidence interval is given by

$$\hat{\theta}_n \pm \frac{\mathfrak{z}_{\alpha/2}}{\sqrt{n} I(\theta)} = \hat{p}_n \pm \frac{\mathfrak{z}_{\alpha/2}}{\sqrt{n}(\hat{p}_n(1-\hat{p}_n))^{-1}} = \hat{p}_n \pm \frac{\mathfrak{z}_{\alpha/2}\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}}.$$

election confidence interval for surveys are of the type presented above.

All that we presented so far was drawn from cases with large samples. We may also wish to see how to derive pivotal quantities for the case of small sample sizes.

Small sample confidence interval

We will use the probability integral transform to get a pivotal quantity. If $X \sim F$, then $F(t) = \mathsf{P}(X \leq t)$ and taking $Y = F(X)$, then $Y \sim \mathcal{U}(0,1)$ if X is a continuous variable. Can take the uniform distribution, generate observations and approximatively invert $F(x)$, that is $X_i = F^{-1}(Y_i)$ given

F monotone.

Suppose that $X \sim \mathcal{E}(\lambda)$ and

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise .} \end{cases}$$

Then

$$F_\lambda(x) = \int_0^x \lambda e^{-\lambda t} dt = -e^{-\lambda t}\Big|_0^x = 1 - e^{-\lambda x}.$$

If we define $Y = 1 - e^{-\lambda X}$ and find the distribution

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}\mathsf{P}(Y \leq y) \\ &= \frac{d}{dy}\mathsf{P}(1 - e^{-\lambda X} \leq y) \\ &= \frac{d}{dy}\mathsf{P}(e^{-\lambda X} \geq 1 - y) \\ &= \frac{d}{dy}\mathsf{P}(-\lambda X \geq \log(1-y)) \\ &= \frac{d}{dy}\mathsf{P}(X \leq -\lambda^{-1}\log(1-y)) \\ &= \frac{d}{dy}\left(1 - e^{-\frac{\lambda \log(1-y)}{\lambda}}\right) = \frac{d}{dy} y = 1 \end{aligned}$$

and so $Y \sim \mathcal{U}(0,1)$. This result holds in more generality, and the proof for an arbitrary absolutely continuous distribution function is left as an exercise (simple application of the change of variable theorem).

We can use this to devise a pivotal quantity. Suppose $X_i \sim F$ and consider $T(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i=1}^n F_{\boldsymbol{\theta}}(X_i)$ and taking logarithms, we obtain

$$-\log T(\boldsymbol{x}) = \sum_{i=1}^n -\log F_\theta(X_i)$$

where $F_{\boldsymbol{\theta}}(X_i) = Y_i$. Note that $Y_i \sim \mathcal{U}(0,1)$.

Exercise 3.1

If $U \sim \mathcal{U}(0,1)$, show that $-\log U \sim (E)(1) = \Gamma(1,1)$. Then $-\log T(\boldsymbol{x}, \boldsymbol{\theta}) =$

$\sum_{i=1}^{n} -\log Y_i = \sum_{i=1}^{n} -Z_i$ and so $Z_i \overset{\text{iid}}{\sim} \mathcal{E}(1)$ for $i = 1, \ldots, n$ so that $\sum_{i=1}^{n} Z_i \sim \Gamma(n, 1)$. If we define $W_n = -\log T(\boldsymbol{x}, \boldsymbol{\theta}) \sim \Gamma(\alpha = n, \beta = 1)$, using the table, we can find a, b such that $\mathsf{P}(a < W_n < b) = 1 - \alpha$. However, we may not in this case be able to isolate $\boldsymbol{\theta}$ in the middle. Note however that we have

$$\mathsf{P}(e^{-b} < T(\boldsymbol{X}, \boldsymbol{\theta}) < e^{-a}) = 1 - \alpha.$$

Remark

Confidence interval for a given confidence level is not unique. That is why, we introduce other optimality notions, such as Uniformly Most Accurate (UMA) confidence intervals, shortest- length (average length) confidence intervals and some others. The former is closely tied to UMP tests. Indeed there is a one-to-one connection between them. As such the restrictions we were faced with when trying to find UMP tests are present here too. The latter, i.e. shortest-length (average length) confidence interval, are desirable and in interesting cases are available, such as normal distribution and hence for asymptotic confidence intervals. The confidence interval presented in class using asymptotic theory for MLE is shortest-length confidence interval.

Section 3.4. Goodness-of-fit tests

In parametric settings, we can get plausible models if we have large data. The question that arises however in practise is how we can know the distribution we stipulate is the correct one for the data. Many things need to be check: whether the random variables are truly independent or come from the same distribution, or even if the parameters of the distribution, say in the case of $\mathcal{N}(\mu, \sigma^2)$ are correctly specified. One thing we could do to is to compare our assumption with the data gathered through means of the empirical distribution, whivh is unbiased, consistent and equivalent to MLE in the non-parametric setting. One thing that is possible is to generate the quantiles and draw them against the assumption: concordance would come with the 45line. We want to measure thus the distance between $\widehat{F_n}$ and F, which leads to different test depending on your metric, such as Kolmogorov-Smirnoff test ($\|\widehat{F_n} - F\|_\infty$), the Cramer Von Mises test with l_2

norm: ($\|\widehat{F_n} - F\|_{l_2}$) which is natural (Euclidian norm), but hard to compute in practise and finally the Anderson-Darling which works in l_2, but with a ponderation function to accommodate for extreme-values. In any such case,

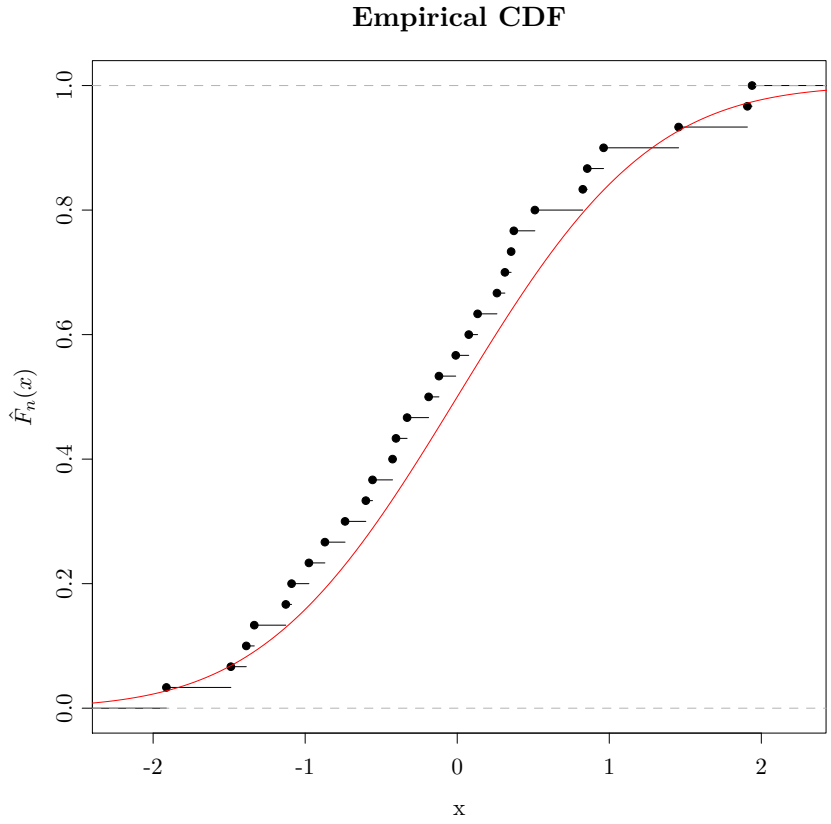


Figure 11: Empirical distribution for a sample of size 20 against the CDF for $\mathcal{N}(0,1)$

we still need to find the distribution, and as soon as sampling becomes more complex and we lose independence, this is a very hard problem.

Chi-square tests

The idea for the test is simple: chop the real line into different bins, take the observations and put into bins, and make the sets $A_1, \ldots, A_n$ and then com-

pare $\mathsf{P}(A_k) - \hat{\mathsf{P}}(A_k)$: compare the two vectors of n-dimension (and not functions). By doing so, we take $(\mathsf{P}(A_1), \dots, \mathsf{P}(A_m))$ with $(\hat{\mathsf{P}}(A_1), \dots, \hat{\mathsf{P}}(A_m))$: by splitting our data into a finite number of group, we reduces the dimension from ∞ to m and created a multinomial distribution in the process.

Theorem 3.17

Let $\boldsymbol{X} = (X, \dots X_{k-1})$ be a multinomial random variable with parameters n, $p_1, \dots, p_{k-1}$. Then, the random variable

$$U_k = \sum_{i=1}^{k-1} \frac{(x_i - np_i)^2}{np_i} \xrightarrow{\mathcal{D}} \chi^2(k-1)$$

as $n \to \infty$ and where k is the number of bins, X_i is the number of observations in the i^{th} bin and np_i is the expected number of observations in bin i under the null hypothesis.

Recall the PDF of the multinomial is given by

$$\mathsf{P}(\boldsymbol{X} = \boldsymbol{x}) = \binom{n}{x_1, \dots x_n} \prod_{i=1}^{n} p_i^{x_i}$$

and where $\sum_{i=1}^{n} x_i = n, \sum_{i=1}^{n} p_i = 1$ and $\boldsymbol{x} = (x_1, \dots, x_n)$, $0 \le x_i \le n$ and the margins are binomial if the random variables are independent. We write $\boldsymbol{X} = (X_1, \dots, X_n) \sim \mathcal{M}(n, p_1, \dots, p_k)$.

Remark

There are only $k-1$ independent random variables as one is completely specified by the others. By normalizing, we can get

$$\left\{ \frac{X_i - np_i}{\sqrt{np_i(1-p_i)}} \right\}^2 \xrightarrow{\mathcal{D}} \chi^2(k-1)$$

but the denominator does not match the one we have; we don't have $1 - p_i$ and the formula given would require $\mathsf{Var}X = \mathsf{E}X = p_i$, so that X be Poisson. For this, we need $np_n \approx \lambda$ so that $n \sim \frac{1}{p_n}$, that is n is reciprocal to p_n and $p_n = \mathcal{O}\left(\frac{1}{n}\right)$. What happens is that $np_n(1-p_n) \approx np_n - np_n^2 \to np_n$ and this formula tells us the number of bins needed.

We now present some important results

Theorem 3.18

Let $X_1, \ldots, X_n$ be a random sample on X. Suppose $H_0 : X \sim F_0$, where F_0 is completely known. Consider a collection of disjoint Borel sets $A_1, \ldots, A_k$ that form a partition of the real line. Let $\mathsf{P}\{X \in A_i\} = p_i$ for $i = 1, \ldots, k$ and assume that $p_i > 0$ for each i. Let Y_j denote the number of $X_i's$ in A_j for $j = 1, \ldots, k, i = 1, \ldots, n$. Then, the joint distribution of $(Y_1, \ldots, Y_k)$ is multinomial with parameters $n, p_1, \ldots, p_k$.

Theorem 3.19

Let $H_0 : X \sim F_0$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$ is unknown. Let $(X_1, \ldots, X_n)$ be a random sample on X and suppose that the MLE's of $\theta_1, \ldots, \theta_r$ exists and are respectively $\hat{\theta}_1, \ldots \hat{\theta}_r$. Let $A_1, \ldots, A_k$ be a collection of disjoint Borel sets that cover the real line (a partition) and let

$$\hat{p}_i = \mathsf{P}_{\hat{\boldsymbol{\theta}}}(X \in A_i) > 0, \quad i =, \ldots, k$$

where $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_r)$ and $\mathsf{P}_{\hat{\boldsymbol{\theta}}}$ is the probability distribution associated with $F_{\boldsymbol{\theta}}$. Let $Y_1, \ldots, Y_k$ be the random variables defined as follows, Y_i: number of $X_i \in A_j$, $i = 1, \ldots, n, j = 1, \ldots, k$. Then, the random variable

$$V_k = \sum_{j=1}^{k} \left\{ \frac{(y_j - n\hat{p}_j)}{n\hat{p}_j} \right\}^2 \xrightarrow{\mathcal{D}} \chi^2(k - r - 1).$$

Be warned that there are two ways here of finding MLE. We have

$$\mathcal{L}(\boldsymbol{\theta}, x_1, \ldots, x_n) \propto \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} p_j(\boldsymbol{\theta})$$

and thus $\mathsf{P}_j(\boldsymbol{\theta}) = \mathsf{P}_{\boldsymbol{\theta}}(x \in A_j)$. If $F_{\boldsymbol{\theta}}$ is the CDF of $\mathcal{N}(\theta, 1)$, then we take $\mathsf{P}_{\boldsymbol{\theta}}(x \in A_j) = \int_{A_j} (\sqrt{2\pi})^{-1} \exp\left(-\frac{1}{2}(x - \theta)^2\right) dx$, find a numerical approximation that we put back in the multinomial, before finding the maximizer of the multinomial likelihood. This is the right approach: maximizing with the Normal and transferring our result to the multinomial would give a wrong result.

We move along with an example, which illustrates the test for means. We could think for example of testing whether hearth attacks happen more on Monday.

Example 3.10 (Fair die)

A die is rolled 120 times with the following result. and we want to test

Value	1	2	3	4	5	6
Frequency of x_j	20	30	20	25	15	10

whether the die is fair, that is $H_0 : p_i = \frac{1}{6}, i = 1, \ldots, 6$. The test statistic is

$$u = \frac{1}{120\left(\frac{1}{6}\right)} \sum_{j=1}^{6} \left[x_j - 120\left(\frac{1}{6}\right)\right]^2 = \frac{2 \cdot 10^2 + 2 \cdot 5^2}{20} = 12.5.$$

Testing at 5% level, we reject H_0 since $12.5 > 11.07$. The intuition here is that small values are good for H_0.

References

[1] Rohatgi, V.K., A.K.M.E. Saleh, *An introduction to Probability and Statistics*, 2nd edition, Wiley, 2001, 716p.

[2] Hogg, Robert, Allen Craig, *Introduction to Mathematical Statistics*, 5nd edition, Prentice Hall, 1995, 564p.

[3] Shao, Jun, *Mathematical Statistics*, 2nd edition, Springer, 2003, 591p.

[4] Masoud Asgharian-Dastenaei, *MATH 357: Honours Statistics*, Notes taken from January to April 2012 at McGill University.

We are grateful to Pr. Asgharian for the content of these notes and the great lectures he gave. We also need to credit the following for pictures and images used through this set of notes:

1. Figure 4: Christophe Jorssen for the Chi-square distribution;

All other illustrations were made with TikZ, GnuPlot or R.

Special thanks to Sébastien Jean for reviewing the notes.

License