Statistical Theory

Yisrael Haber

October 26, 2021

Abstract

This is based on the lecture notes in the course "Statistical Theory" in Bar-Ilan university given by professor Simi Haber in spring semester of 2021. I do not take responsibility for errors in this document, I recommend to use this in addition to official sources and not instead of such sources.

Contents

1 First Lecture

1.1 Introduction To Probability

Definition 1.1. A probability space is a 3-tuple $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\Omega \neq \emptyset$ (The sample space), $\mathcal{F}$ is a $\sigma-$algebra and $\mathbb{P} : \mathcal{F} \to [0, 1]$ additive function.

Definition 1.2. A $\sigma-$algebra $\mathcal{F}$ is a collection of sets over a set Ω such that -

(1). $\mathcal{F} \subseteq P(\Omega)$.

(2). $\Omega, \emptyset \in \mathcal{F}$.

(3). $\mathcal{F}$ is closed to the complements i.e. if $A \in \mathcal{F}$ then $A^C \in \mathcal{F}$.

(4). $\mathcal{F}$ is closed under countable union i.e. if $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{F}$ then $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

The sets in $\mathcal{F}$ are called the measurable sets, the tuple $(\Omega, \mathcal{F})$ is called a measurable space.

Definition 1.3. A function $\mathbb{P} : \mathcal{F} \to [0, 1]$ over a measurable space $(\Omega, \mathcal{F})$ is a probability function if and only if the following happen:

(1). $\mathbb{P}(\emptyset) = 0, \mathbb{P}(\Omega) = 1$.

(2). For a collection of non-intersecting $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$ we have:

$$\mathbb{P}(\biguplus_{n\in\mathbb{N}} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

For countable sets Ω we usually take $\mathcal{F} := 2^{\Omega}$, the problem arises when we deal with non-countable sets where if we do take the power set to be the σ-algebra we get contradictions to the definition of the probability measure (for example take the Vitaly set).

Therefore when dealing, for example, with $\Omega := [0, 1]$ the normal way to define a probability space (with uniform probability) is to define for every interval $I \subseteq \Omega$ the probability to be it's length (in proportion to the original length which in this case is normalized to be 1). We can then take the $\sigma-$algebra to be the smallest one such that it contains every open interval in Ω, this is called the Borel $\sigma-$algebra ($\mathcal{B}$).

Definition 1.4. A function $X : \Omega \to \mathbb{R}$ such that for all $B \in \mathcal{B}$ we have $X^{-1}[B] \in \mathcal{F}$ is called measurable function, over a probability space it is called a random variable.

Definition 1.5. Given a random variable X over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ define a function $f_X : \mathbb{R} \to \mathbb{R}$ (the distribution) such that for all k we take $f_X(k) = \mathbb{P}(X = k)$.

Definition 1.6. Given a random variable X over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ define a function $F_X(r) = \mathbb{P}(X \leq r)$ (This is called the cumulative distribution function or the CDF).

Characteristics Of A CDF:

- F_X is monotone non-decreasing.
- $\lim_{x \to \infty} F_X(x) = 1$.
- $\lim_{x \to -\infty} F_X(x) = 0$.
- F_X is continuous from the right.

Theorem 1.1. *If a function F satisfies the previous characteristics then there is a random variable X such that F is it's CDF.*

Definition 1.7. For a random variable X with differentiable CDF F_X If F_X the density function f_X is the derivative of F_X.

Characteristics Of A Density Function:

- For all $x \in \mathbb{R}$ we have $f(x) \geq 0$.

- $\int_{\mathbb{R}} f(x)dx = 1$.

Theorem 1.2. *Every integrable function with the previous characteristics is a density function for some random variable.*

Definition 1.8. For $X, Y : \Omega \to \mathbb{R}$ random variables, we say that they are independent if and only if for all $A, B \in \mathcal{B}$ we have

$$\mathbb{P}(X \in A \wedge Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B)$$

Otherwise they are said to be dependent

1.2 Important Discrete Distributions:

(1): Uniform Distribution - Given a finite set A we pick an element $a \in A$ with equal probability $\frac{1}{|A|}$ for every element. We denote $U[a,b]$ to be the uniform distribution on $\{a, a+1, \ldots, b\}$. It's support is the set itself, and we have -

$$f_X(k) := \begin{cases} \frac{1}{b-a+1}, \ a \leq k \leq b \\ 0, \ \text{Otherwise} \end{cases}$$

(2): Bernoulli Distribution with parameter p ($p \in [0,1]$). We say that $X \sim \text{Ber}(p)$ if

$$\mathbb{P}(X = r) = \begin{cases} p, \ r = 1 \\ 1 - p, \ r = 0 \end{cases}$$

(3): Binomial Distribution over n experiments and parameter p. We say that $X \sim \text{Bin}(n, p)$ if and only if for all $k \in \{0, 1, \ldots, n\}$ we have

$$\mathbb{P}(X = k) = \tbinom{n}{k} p^k (1-p)^k$$

(4): The Geometric Distribution over $\mathbb{N}$ with parameter p. We say that $X \sim \text{Geo}(p)$ if and only if for all $k \in \mathbb{N}$ we have

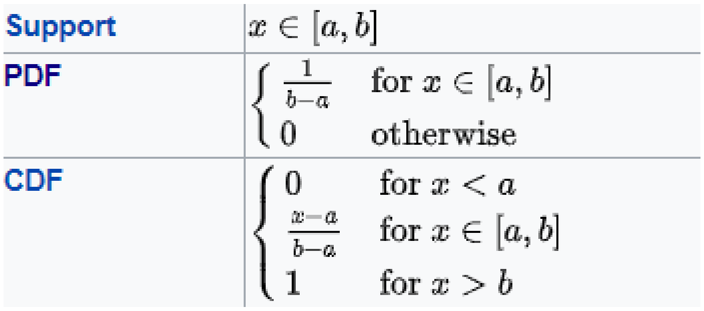
$$\mathbb{P}(X = k) = (1-p)^{k-1} p$$

(5): The Poisson Distribution over $\mathbb{N}$ with parameter λ. We say that $X \sim \text{Poi}(\lambda)$ if and only if for all $k \in \mathbb{N}$

$$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

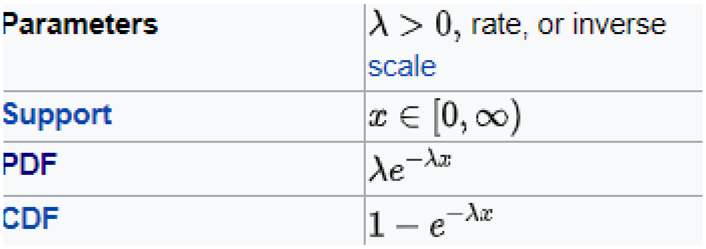
1.3 Important Uniform Distributions:

(1). Continuous Uniform Distribution over an interval $[a,b] \subset \mathbb{R}$. We say that $X \sim U[a,b]$ if it has the following characteristics



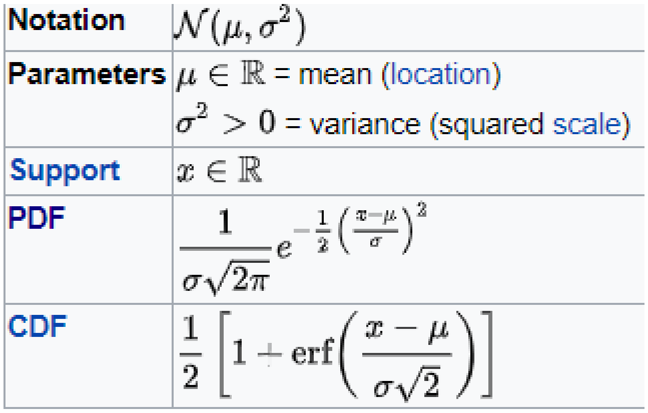
Support	$x \in [a,b]$
PDF	$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$
CDF	$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b] \\ 1 & \text{for } x > b \end{cases}$

(2). Exponential Distribution over $[0,\infty)$ with parameter $\lambda > 0$. We say that $X \sim \text{Exp}(\lambda)$ if and only if it has the following characteristics



Parameters	$\lambda > 0$, rate, or inverse scale
Support	$x \in [0,\infty)$
PDF	$\lambda e^{-\lambda x}$
CDF	$1 - e^{-\lambda x}$

(3). Normal Distribution over $\mathbb{R}$ with parameters μ, σ $(\sigma > 0)$. We say that $X \sim \mathcal{N}(\mu, \sigma^2)$ if and only if it has the following characteristics



Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 > 0$ = variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
CDF	$\frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$

2 Second Lecture

2.1 Conditional Probability

Definition 2.1. Given $B \in \mathcal{F}$ an event with $\mathbb{P}(B) > 0$ then we define the conditional probability measure -

$$\mathbb{P}_B(A) = \mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Formally this moves us to a know probability space - instead of $(\Omega, \mathcal{F}, \mathbb{P})$ we now have $(B, \mathcal{F} \mid_B, \mathbb{P}_B)$.

We can now use this to also consider density functions for a pair of random variables (X,Y) and have $f_Y(b) \cdot f_{X|Y=b} = f_{X,Y}(a,b)$.

Example 2.1. *Take $X \sim Exp(\lambda)$, meaning $f_X(t) = \lambda e^{-\lambda t}$. Consider the event $(X - a \leq b \mid X > a)$. This is -*

$$\mathbb{P}(X \leq a + b \mid X > a) = \frac{\mathbb{P}(X \leq a+b, X > a)}{\mathbb{P}(X > a)} = \frac{\int_a^{a+b} \lambda e^{-\lambda t} dt}{\int_a^{\infty} \lambda e^{-\lambda t} dt} =$$

$$= \frac{-e^{-\lambda t} \mid_a^{a+b}}{-e^{-\lambda t} \mid_a^{\infty}} = \cdots = 1 - e^{-\lambda b} = \mathbb{P}(X \leq b)$$

This is called the memorylessness characteristic of the exponential distribution.

Definition 2.2. For a discrete random variable X we define the expected value of X to be - $\mathbb{E}[X] = \sum_a a\mathbb{P}(X = a)$. For a continuous random variable the expected value to be - $\mathbb{E}[X] = \int_{\mathbb{R}} t f(t) dt$.

Example 2.2. *Take the binomial distribution $X \sim Bin(n,p)$. There are many ways to calculate the expected value, it can be done by hand but the more pleasant way of doing this is by using the linearity of the expected value - $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ for any 2 random variables X and Y, and by noticing that the binomial distribution is a sum of n bernoulli experiments $\{X_i\}_{i=1}^n$ each with expected value p and therefore -*

$$\underline{\overline{\mathbb{E}[X]}} = \mathbb{E}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \mathbb{E}[X_i] = \sum_{i=1}^n p = \underline{\overline{np}}$$

Claim 2.1. The law of returning expected values is that for any 2 random variables X,Y we have - $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]]$.

Proof. For this we need to notice that $Y \mid X$ is a function such that every $b \in \mathbb{R}$ maps to the random variable $Y \mid X = b$ and we can ask what it's expected value at each $b \in \mathbb{R}$. The function $\mathbb{E}[Y \mid X]$ does exactly that - for every $b \in \mathbb{R}$ it returns the value $\mathbb{E}[Y \mid X = b]$. Now the claim is that the expected value of $\mathbb{E}[Y \mid X]$ is just $\mathbb{E}[Y]$. ■

Example 2.3. *Take a school of kids - the sample space is the the set of all the kids in the school. Y measures the height of a child, and X says what class the child is in. What this tells us is that averaging per class and then over all of the classes is similar to averaging over all of the kids in the school at once.*

Example 2.4. *Flip a fair coin until you get "tails", and then toss a fair die the number of times we have tossed "heads" and sum over all of the rolls.*

We can then take $N \sim \text{Geo}(\frac{1}{2})$, and $X_i \sim U(\{1, 2, 3, 4, 5, 6\})$ and finally take $Y = \sum_{i=1}^{N} X_i$.

Using the previous claim to see that

$$\underline{\overline{\mathbb{E}[Y]}} = \mathbb{E}[\mathbb{E}[Y \mid N]] = \mathbb{E}[\mathbb{E}[X_1 \mid N] + \mathbb{E}[X_2 \mid N] + \cdots + \mathbb{E}[X_N \mid N]] =$$

$$= \mathbb{E}[\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_N]] = \mathbb{E}[N]\mathbb{E}[X_i] = \frac{7}{2} \cdot 2 = \underline{\overline{7}}$$

Definition 2.3. For a random variable X we can define it's variance in the following way -

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

We can see the following characteristics quite easily -

- $\text{Var}(X) \geq 0$.

- For all $a \in \mathbb{R}$ we have $\text{Var}(X + a) = \text{Var}(X)$, and $\text{Var}(a \cdot X) = a^2\text{Var}(X)$.

Claim 2.2. The law of total variance says that -

$$\mathbf{Var}(Y) = \mathbf{E}[\mathbf{Var}(Y \mid X)] + \mathbf{Var}(\mathbf{E}[Y \mid X]).$$

2.2 Sufficient Statistics Part 1

Denote $\bar{y} = (y_1, \dots, y_n)$ to be a sample (a realization) from a random vector $\bar{Y} = (Y_1, \dots, Y_n)$ with some distribution $f_Y(\bar{y})$. We will assume that Y comes from a family of distributions $\mathcal{G}$, which is characterized by a vector of parameters $\bar{\theta} \in \Theta$ where Θ is the parameter space.

Additionally we will also assume that $Y_1, \dots, Y_n$ are i.i.d. from some distribution $f_\theta(y)$ from a family $\mathcal{G}_\theta$ for $\theta \in \Theta$.

Example 2.5. *In a hospital we record the distribution of births - whether they are male (M) or female (F). Let's say this comes out to be -*

(F,M,F,F,M,M,F,M,F,F,F,F,M,M,F,F,F,M,M,F).

We can now define Y_i to be the indicator function that tells us that the i-th birth was a female. It is natural to assume that $Y_i \sim Ber(p)$ for some $p \in (0, 1)$ i.i.d. So our parameter space is $\Theta = (0, 1)$.

Example 2.6. *Someone wants to see how much codeine there is in a pill for headache relief, she purchases 5 pills. On the packaging it says that the amount of codeine in a pill is 200 mg. In the ones she checks at home she observes the following measurements -*

(200.3,195.0,192.4,201.2,190.1)

A model that fits these measurements could be the normal distribution - $Y_i = \mu + \epsilon_i$ where ϵ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$. She wants to know what μ, σ^2 are most likely to be.

Therefore the first task should be the likelihood measurements -

- From example 2.5, is the probability for a female child $\frac{12}{20}$?

- The average measurement that she measures in example 2.6 is 196.6 mg, and this is different from what the package says. Is the package "lying"?

2.3 Likelihood:

Given a sample $y = (y_1, \dots, y_n)$ that is a realization of the distribution $Y = (Y_1, \dots, Y_n)$ with distribution $f_\theta(y)$ for $\theta \in \Theta$.
Now assume that Y is a discrete random variable, what is the probability to sample y?

$$\mathbb{P}_\theta(Y_1 = y_1, \dots, Y_n = y_n)$$

Definition 2.4. The likelihood function is $L(\theta; y) := \mathbb{P}_\theta(Y = y)$. We consider θ_1 to be a more likely parameter than θ_2 if $L(\theta_1; y) > L(\theta_2; y)$.

If the Y_i are i.i.d. then we have the following result -

$$L(\theta; y) = \prod_{i=1}^{n} \mathbb{P}_\theta(Y_i = y_i)$$

For example for example 2.5 we have -

$$L(p; y) = p^{y_1}(1-p)^{1-y_1} \cdots p^{y_n}(1-p)^{1-y_n} = p^{\sum_i y_i}(1-p)^{n-\sum_i y_i}$$

We can now look at different values of p and compare to see what parameters are more or less likely.

Notice that we have defined this for discrete Y, now assume Y is continuous with density function $f_\theta(y)$. The likelihood function is now defined to be - $L(\theta, y) = f_\theta(y)$. Now we lose the exact probabilities but we are still able to compare for different θ's. As from before we have that for i.i.d. $y_1, \dots, y_n$ -

$$L(\theta; y) = \prod_{i=1}^{n} f_\theta(y_i)$$

For example 2.6 we can therefore have -

$$L(\mu, \sigma; y) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y_i - \mu}{\sigma})^2}$$

3 Third Lecture

3.1 Sufficient Statistic

Definition 3.1. A statistic is a measurable function on the sampling and it is denoted by $T(Y)$.

Example 3.1. *Given a sampling $(y_1, \dots, y_n)$, the following are statistics -*

1. $\sum_{i=1}^{n} y_i$
2. $\max\{y_i\}$
3. $Y_1 - Y_2^{Y_3}$

This of course precludes using functions with information that the sampling doesn't reveal to us - for example a function that already contains our constants θ.

Example 3.2. *A statistic $T(Y)$ is sufficient for an unknown parameter θ if the conditional distribution of the sampling given the statistic is independent of θ.*

Example 3.3. *We will return to example 2.5. We will show that $\sum_{i=1}^{n} Y_i$ is a sufficient statistic for parameter p (assuming every birth is bernoulli with parameter p). We want to check the following value -*

$$\mathbb{P}\left(Y \mid \sum_{i=1}^{20} \overrightarrow{y_i} = t\right) = \frac{\mathbb{P}(\overrightarrow{Y} = \overrightarrow{y},\ \sum_{i=1}^{20} \overrightarrow{y_i} = t)}{\mathbb{P}(\sum_{i=1}^{20} \overrightarrow{y_i} = t)} =$$

$$\begin{cases} \frac{1}{\binom{20}{t}}, & \text{if } \sum y_i = t \\ 0, & \text{Otherwise} \end{cases}$$

And this is independent of p and therefore this is a sufficient statistic by the previous definition.

Given a statistic it is now known how to check whether it is sufficient or not, but how can you generate a sufficient statistic?

Theorem 3.1 (Fisher-Neyman Factorization Theorem). *A statistic $T(\overrightarrow{Y})$ is sufficient for a parameter θ if and only if for all $\theta \in \Theta$ we have -*

$$L(\theta; \overrightarrow{y}) = g(T(\overrightarrow{y}), \theta) \cdot h(\overrightarrow{y})$$

Where the first part isn't directly dependent on $\overrightarrow{y}$ and the second part is independent of θ.

Proof. We will show for discrete random variables, but this can be easily be extended to continuous random variables.

$\overrightarrow{y}$ is a realization of $\overrightarrow{Y}$, $T(\overrightarrow{y}) = t$.

($\Longrightarrow$): Now assume that $T(\overrightarrow{Y})$ is sufficient. Notice that

$L(\theta; \overrightarrow{y}) = \mathbb{P}(\overrightarrow{Y} = \overrightarrow{y}) = \mathbb{P}(\overrightarrow{Y} = \overrightarrow{y},\ T(\overrightarrow{y}) = t) = \mathbb{P}(\overrightarrow{Y} = \overrightarrow{y} \mid T(\overrightarrow{y}) = t) \cdot \mathbb{P}(T(\overrightarrow{y}) = t)$

Now take $\mathbf{h}(\overrightarrow{y}) := \mathbb{P}(\overrightarrow{Y} = \overrightarrow{y} \mid T(\overrightarrow{y}) = t)$. From sufficiency h is independent of θ.

Now define $\mathbf{g}(T(\overrightarrow{y}), \theta) := \mathbb{P}(T(\overrightarrow{y}) = t)$ And we have showed what we wanted.

($\Longleftarrow$): Now assume that the factorization exists, we now need to show that the statistic is sufficient meaning it is independent of θ. Now assume

$$L(\theta; \overrightarrow{y}) = g(T(\overrightarrow{y}), \theta) \cdot h(\overrightarrow{y})$$

Notice that

$$\mathbb{P}(\overrightarrow{Y} = \overrightarrow{y} \mid T(\overrightarrow{y}) = t) = \begin{cases} \frac{\mathbb{P}(\overrightarrow{Y} = \overrightarrow{y})}{\sum_{\overrightarrow{y} \mid T(\overrightarrow{y} = t)} \mathbb{P}(\overrightarrow{Y} = \overrightarrow{y})}, & T(\overrightarrow{y} = t) \\ 0, & \text{Otherwise} \end{cases} =$$

$$= \begin{cases} \frac{h(\overrightarrow{y})}{\sum_{\overrightarrow{y} \mid T(\overrightarrow{y} = t) h(\overrightarrow{y})} \mathbb{P}(\overrightarrow{Y} = \overrightarrow{y})}, & T(\overrightarrow{y} = t) \\ 0, & \text{Otherwise} \end{cases}$$

This is independent of θ and therefore we are done. ■

Claim 3.1. Every one-to-one measurable function on a sufficient statistic is also a sufficient statistic for the same parameter.

Proof. Assume g is a one-to-one measurable function, and $T(\overrightarrow{y})$ is a sufficient statistic for θ. Now we want to show that

$$\mathbb{P}(\overrightarrow{Y} = y \mid f(T(\overrightarrow{y})) = t)$$

is independent of θ. g is one-to-one and therefore there is a unique s such that $g(s) = t$ which tells us that

$$\mathbb{P}(\overrightarrow{Y} = y \mid f(T(\overrightarrow{y})) = t) = \mathbb{P}(\overrightarrow{Y} = y \mid T(\overrightarrow{y}) = s)$$

Now T is a sufficient statistic and therefore it is independent of θ.

Alternatively this can also be shown using the factorization theorem. ■

Example 3.4. *Take a poisson distribution with constant λ, where $Y_i \sim Poi(\lambda)$ i.i.d. now we can see that the sum is also a sufficient statistic in this case to -*

$$L(\lambda; \overrightarrow{y}) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{y_i}}{y_i!} = e^{-n\lambda} \lambda^{\sum y_i} \prod_{i=1}^{n} \cdot \frac{1}{y_i!}$$

And form the factorization theorem we have that it is a sufficient statistic.

Example 3.5. *Going back to example 2.6 we can see*

$$L(\mu, \sigma; \overrightarrow{y}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

And similarly we will get that this is a sufficient statistic.

Now we can consider the following statistic -

$$T(\overrightarrow{y}) = \left(\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \frac{1}{n} \sum_{i=1}^{n} (\bar{y} - y_i)^2 \right)$$

Rewrite this to get

$$\sum_{i=1}^{n} (y_i - \mu)^2 = \sum_{i=1}^{n} (y_i - \bar{y} + \bar{y} - \mu)^2 = \sum_{i=1}^{n} \left[(y_i - \bar{y})^2 + 2(y_i - \bar{y})(\bar{y} - \mu) + (\bar{y} - \mu)^2 \right] =$$

$$\cdots = \sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2$$

Now it can be seen that

$$L(\mu, \sigma; \overrightarrow{y}) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{2\sigma^2}} \cdot e^{-\frac{n(\bar{y} - \mu)^2}{2\sigma^2}}$$

And from the factorization theorem this is a sufficient statistic.

3.2 Minimal Sufficient Statistic

We have seen that

- T is a sufficient statistic if and only if for every one-to-one function f, $f \circ T$ is a statistic.

- If T is a sufficient statistic then for some other function S we have that (T, S) is a sufficient statistic.

We are looking for a minimalist representation of the sufficient statistic.

Definition 3.2. A sufficient statistic $S(\overrightarrow{y})$ is called minimal if for any other sufficient statistic $T(\overrightarrow{y})$ there is a function f such that $S(\overrightarrow{y}) = f(T(\overrightarrow{y}))$

Theorem 3.2. *There exists a minimal sufficient statistic. (Uniqueness is not necessarily trivial)*

Proof. First define an equivalence relation. For $\overrightarrow{y_1}, \overrightarrow{y_2} \in \Omega$. Then

$$\overrightarrow{y_1} \sim \overrightarrow{y_2} \iff \frac{L(\theta; \overrightarrow{y_1})}{L(\theta; \overrightarrow{y_2})} \text{ is independent of } \theta$$

Define a statistic $T(\overrightarrow{y})$ as a one-to-one function from the set of equivalence classes. Now we will show that T is a minimal sufficient statistic.

$$\mathbb{P}(\overrightarrow{Y} = \overrightarrow{y} \mid T(\overrightarrow{y}) = t) = \begin{cases} \frac{\mathbb{P}(\overrightarrow{Y} = \overrightarrow{y})}{\sum_{\overrightarrow{y_2} \mid T(\overrightarrow{y_2}) = t} \mathbb{P}(\overrightarrow{Y} = \overrightarrow{y_2})}, & T(\overrightarrow{y}) = t \\ 0, & \text{Otherwise} \end{cases} =$$

$$= \begin{cases} \frac{L(\theta; \overrightarrow{y_1})}{\sum_{\overrightarrow{y_2} \mid T(\overrightarrow{y_2}) = t} L(\theta; \overrightarrow{y_2})}, & T(\overrightarrow{y}) = t \\ 0, & \text{Otherwise} \end{cases} =$$

$$= \begin{cases} \frac{1}{\sum_{\overrightarrow{y_2} \mid T(\overrightarrow{y_2}) = t} \frac{L(\theta; \overrightarrow{y_1})}{L(\theta; \overrightarrow{y_2})}}, & T(\overrightarrow{y}) = t \\ 0, & \text{Otherwise} \end{cases}$$

And this is independent of θ which tells us that T is a sufficient statistic, and what we have left is to prove is that it is minimal.

Assume that S is a different sufficient statistic, and define an equivalence relation -

$$\overrightarrow{y_1} \sim_s \overrightarrow{y_2} \iff S(\overrightarrow{y_1}) = S(\overrightarrow{y_2})$$

What we want to prove is that this equivalence relation has the following characteristic

$$\overrightarrow{y_1} \sim_s \overrightarrow{y_2} \implies \overrightarrow{y_1} \sim \overrightarrow{y_2}$$

and we will be done.

So assume $\overrightarrow{y_1} \sim_s \overrightarrow{y_2}$. We want to show now that

$$\frac{L(\theta; \overrightarrow{y_1})}{L(\theta; \overrightarrow{y_2})} \text{ is independent of } \theta$$

And so

$$\frac{L(\theta; \overrightarrow{y_1})}{L(\theta; \overrightarrow{y_2})} = \frac{g(S(\overrightarrow{y_1}, \theta)) \cdot h(\overrightarrow{y_1})}{g(S(\overrightarrow{y_2}, \theta)) \cdot h(\overrightarrow{y_2})} = \frac{h(\overrightarrow{y_1})}{h(\overrightarrow{y_2})}$$

Which is independent of θ and therefore we are done. ■

4 Fourth Lecture

Example 4.1. *Take a Bernoulli distribution Be(p), and 2 sampling series - $\overrightarrow{y_1}, \overrightarrow{y_2}$. We want to see when they are equivalent (under the equivalency that appears in the previous proof). This means we want to know when the following value is independent of p -*

$$\frac{L(\theta, \overrightarrow{y_1})}{L(\theta, \overrightarrow{y_1})} = \frac{p^{(\sum_{i=1}^{n} y_{1,i})}(1-p)^{(n-\sum_{i=1}^{n} y_{1,i})}}{p^{(\sum_{i=1}^{n} y_{2,i})}(1-p)^{(n-\sum_{i=1}^{n} y_{2,i})}}$$

This is independent of p if and only if

$$\sum_{i=1}^{n} y_{1,i} = \sum_{i=1}^{n} y_{2,i}$$

Which means that the sum

$$T(\overrightarrow{y}) := \sum_{i=1}^{n} y_i$$

should be a minimal sufficient statistic in this case.

Example 4.2. *Consider a normal distribution, and look at the same fraction -*

$$\frac{L(\theta, \overrightarrow{y_1})}{L(\theta, \overrightarrow{y_1})} = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{\sum_{i=1}^{n}(y_{1,i}-\mu)^2}{\sigma^2}}}{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{\sum_{i=1}^{n}(y_{2,i}-\mu)^2}{\sigma^2}}} = \exp\left(-\frac{\sum_{i=1}^{n}(y_{1,i}-\mu)^2 - \sum_{i=1}^{n}(y_{2,i}-\mu)^2}{2\sigma^2}\right) =$$

$$= \exp\left(-\frac{\sum_{i=1}^{n}(y_{1,i}-\bar{y_1})^2 - \sum_{i=1}^{n}(y_{2,i}-\bar{y_2})^2 + n(\bar{y_1}-\mu)^2 - n(\bar{y_2}-\mu)^2}{2\sigma^2}\right)$$

And this happens when the exponent is 0 and this happens when

$$\sum_{i=1}^{n}(y_{1,i}-\bar{y_1})^2 = \sum_{i=1}^{n}(y_{1,i}-\bar{y_2})^2$$

And

$$\bar{y_1} = \bar{y_2}$$

Which means we can define

$$T(\overrightarrow{y}) = \left(\bar{y}, \sum_{i=1}^{n}(y_i - \bar{y})^2\right)$$

to be a minimal sufficient statistic.

4.1 Complete statistic:

Definition 4.1. A statistic T is said to be complete if and only if for all measurable functions g:

$$\forall\theta:\ \mathbb{E}\left[g(T(\overrightarrow{Y}))\right]=0\Longrightarrow\forall\theta:\ \mathbb{P}_\theta\left(g(T(\overrightarrow{Y}))=0\right)=1$$

Example 4.3. *$Y_1,\ldots,Y_n\sim Be(p)$, and take the known minimal statistic*

$$T(\overrightarrow{y})=\sum_{i=1}^{n}y_i$$

Take a measurable function g and assume that for all $p\in(0,1)$

$$\mathbb{E}\left[g(T(\overrightarrow{Y}))\right]=0$$

Since T distributes binomially we have that -

$$\mathbb{E}\left[g(T(\overrightarrow{Y}))\right]=\sum_{k=0}^{n}g(k)\mathbb{P}\left(T(\overrightarrow{Y})=k\right)=\sum_{k=0}g(k)\binom{n}{k}p^k(1-p)^{n-k}=$$

$$=(1-p)^n\sum_{i=1}^{n}g(k)\binom{n}{k}\left(\frac{p}{1-p}\right)^k$$

Assume that for all p this is 0, we can now ignore what come before the sum and notice that this is a polynomial in $\frac{p}{1-p}$ that has infinite amount of zeros. This is only possible if the polynomial itself has coefficients that are only zeros, $\binom{n}{k}$ is always non-negative values and so the function g has to be the zero function.

Example 4.4. *Assume $Y_1,Y_2\sim\mathcal{N}(\mu,1)$ i.i.d., and define $T=Y_1+Y_2$. Notice that $T\sim\mathcal{N}(2\mu,2)$. Given a measurable function g*

$$\mathbb{E}\left[g(T(\overrightarrow{Y}))\right]=\int_{\mathbb{R}}g(t)\frac{1}{2\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{t-2\mu}{2})^2}dt=\frac{1}{\sqrt{8\pi}}e^{-\frac{1}{2}\mu^2}\int_{\mathbb{R}}g(t)e^{-\frac{t^2}{8}}e^{\frac{1}{2}\mu t}dt$$

If we now take

$$h(t):=g(t)e^{-\frac{t^2}{8}}$$

We can see that if the above integral is 0 always we also have that the MGF of h will be zero everywhere which means that $h=0$ a.s. meaning $g=0$ a.s. which gives us what we want - T is a complete statistic.

Theorem 4.1. *If T is a sufficient complete statistic, it is also a minimal statistic (up to probability 0 sets).*

Proof. Let S be a sufficient minimal statistic. (existence was proven last lecture). We know there is a function g_1 such that $S = g_1(T)$.

Define

$$g_2(s) := \mathbb{E}[T \mid S]$$

g_2 is a statistic since S is a sufficient statistic. Now we know -

$$\mathbb{E}[T] = \mathbb{E}[\mathbb{E}[T \mid S]] = \mathbb{E}[g_2]$$

Now for all θ we have that

$$\mathbb{E}[T - g_2] = 0$$

S is a function of T and therefore we have that

$$\forall\theta : \ \mathbb{E}[T - g_2(g_1(T))] = 0$$

From completeness of Twe have

$$\forall\theta : \ \mathbb{P}[T = g_2(g_1(T))] = 0$$

S therefore is also a sufficient minimal statistic (since T is a function of it and every other statistic can be derived from this). It is also equal to T almost everywhere. ■

4.2 Estimation

Definition 4.2. An estimator is a statistic that is supposed to estimate θ. The value $T(\overrightarrow{y})$ is called an estimate.

5 Fifth Lecture: Estimation Strategies

5.1 Maximum Likelihood Estimation

Given the likelihood function it is natural to try and find the coefficients that maximize the likelihood function.

Example 5.1. *An MLE (Maximum Likelihood Estimation) is an estimator* $\hat{\theta}_{MLE}$ *such that*

$$\hat{\theta}_{MLE} = argmax_{\theta \in \Theta} L(\theta; \overrightarrow{y})$$

Often there is an assumption of the samples being i.i.d. and therefore the likelihood is a multiplication. This is then translated to a sum by applying a logarithm function to both sides and solving that minimization problem.

Example 5.2. *given a bernoulli distribution* $Be(p)$ *we know that -*

$$L(p; \overrightarrow{y}) = p^{\sum_i y_i}(1-p)^{n-\sum_i y_i}$$

And so applying the logarithm we get

$$l(p; \overrightarrow{y}) = \log p \left(\sum_i y_i \right) + \log(1-p) \left(n - \sum_i y_i \right)$$

Finding the minimum of this gives the following natural result

$$\hat{\theta}_{MLE} = \frac{\sum_i y_i}{n} = \bar{y}$$

Example 5.3. *Given a uniform distribution where the density is*

$$f_Y(y) := \begin{cases} \frac{1}{\theta}, & 0 \leq y \leq \theta \\ 0, & Otherwise \end{cases}$$

The likelihood function will be

$$L(\theta; y) \begin{cases} \frac{1}{\theta^n}, & \max\{y\} \leq \theta \\ 0, & Otherwise \end{cases}$$

And so the MLE estimator will be $\max\{y\}$

Example 5.4. *Assume* $Y_i \sim \mathcal{U}[\theta, \theta+1]$ *i.i.d. Meaning*

$$f_Y(y) := \begin{cases} 1, & \theta \leq y \leq \theta + 1 \\ 0, & \textit{Otherwise} \end{cases}$$

And so the likelihood function -

$$L(\theta; y) \begin{cases} 1, & \max\{y\} - 1 \leq \theta \leq \min\{y\} \\ 0, & \textit{Otherwise} \end{cases}$$

Meaning in this case the solution is not unique - any theta that satisfies that the above function is 1 will give us a solution.

Example 5.5. *Given* $t_1, \ldots, t_n \sim$ *Exp*(θ) *i.i.d. It is known that*

$$f_\theta(t) = \theta \cdot e^{-\theta t}$$

And so

$$L(\theta; \overrightarrow{y}) = \theta^n e^{-\theta \sum_i t_i}$$

Applying the logarithm we get

$$l(\theta; \overrightarrow{y}) = n \log(\theta) - \theta \sum_i t_i$$

Finding the minimum in the obvious way will give us that the MLE will $\frac{n}{\sum_i t_i} = \bar{t}^{-1}$

Definition 5.1. Given a function g on the parameter space we define the likelihood for $z = g(\theta)$ to be

$$L_{(z)}(z; \overrightarrow{y}) = \sup_{\{\theta : g(\theta) = z\}} L_{(\theta)}(\theta; \overrightarrow{y})$$

Claim 5.1. Given such a function g, and the MLE $\hat{\theta}$ then a MLE for z is

$$\hat{z}_{MLE} = g(\hat{\theta})$$

The proof for this is quite straightforward.

Therefore we know that the MLE is a function of every sufficient statistic. This can be seen through the fact that you can use Fischer-Neyman decomposition, and get rid of the part that is solely dependent on the sampling to get that the MLE is

$$\text{argmax}_\theta L(\theta; \overrightarrow{y}) = \text{argmax}_\theta g(T(\overrightarrow{y}), \theta)$$

5.2 Method Of Moment Estimator

Given a random variable you can calculate the moments

$$\mu_k(\theta) := \mathbb{E}\left[Y^k\right]$$

Now for a sampling we can define the moment of this sampling to be -

$$M_k := \frac{1}{n}\sum_{i=1}^{n} y_i^k$$

Now it is natural to demand that the most accurate θ will agree with the moment sampling. Meaning we will have a system of equations -

$$M_k := \mu_k(\theta)$$

To get $\hat{\theta}_{\mathrm{MME}}$

Example 5.6. *Assume the sampling is from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ i.i.d.*

This means the equations we get are

$$\hat{\mu}_{MME} = \frac{1}{n}\sum_i y_i$$

$$\hat{\mu}_{MME}^2 + \hat{\sigma}_{MME}^2 = \frac{1}{n}\sum_i y_i^2$$

The second equation can be shortened using the first one to get that $\hat{\sigma}_{MME}^2$ is just the variance of the sampling.

Example 5.7. *Assume $Y \sim \mathcal{U}[o, \theta]$. And so we have*

$$\mathbb{E}[Y] = \frac{\theta}{2}$$

And so MME will give us

$$\hat{\theta}_{MME} = \frac{1}{2n}\sum_i y_i = 2\bar{y}$$

5.3 Comparison between estimators

Definition 5.2. Given an estimator $\hat{\theta}$ we define the Mean-Square-Error

$$MSE(\hat{\theta}) = \mathbb{E}_\theta \left[(\hat{\theta} - \theta)^2 \right]$$

It can be seen (similar to how you calculate variance of a random variable) that this is equal to -

$$\text{bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta})$$

Where the bias is

$$\text{bias}(\hat{\theta}) := \theta - \mathbb{E}[\hat{\theta}]$$

Definition 5.3. An estimate $\hat{\theta}$ such that $\text{bias}(\hat{\theta}) = 0$ for all estimators θ is called un-biased

Example 5.8. *Assume $Y_i \sim f_\theta(\overrightarrow{y})$ i.i.d. We want to estimate the expected value, using MME we will just guess the average of the sampling. Using linearity of expected value, the average of the sampling is always unbiased. Therefore for the MSE we need only calculate the Variance -*

$$Var(\bar{Y}) = \frac{1}{n^2} \sum_i Var(Y_i) = \frac{\sigma^2}{n}$$

Meaning that overall

$$MSE(\bar{Y}) = \frac{\sigma^2}{n}$$

Another option for example would be to estimate the expected value by taking the first sample only. This will also be unbiased. And again we only need the variance. When we check the variance we will see that

$$Var(Y_1) = \sigma^2$$

And so

$$MSE(Y_1) = \sigma^2$$

6 Sixth Lecture:

Example 6.1. *We can create 3 different estimators for the bernoulli example that we have brought up many times throughout the course $(Y_1, \dots, Y_n \sim Be(p))$*

$$(1).\hat{p_1} := \bar{Y}$$

$$(2).\hat{p_2} := \frac{\sum_{i=1}^{n} Y_i + 1}{n+2}$$

$$(3).\hat{p_3} := Y_1$$

We know that $\hat{p_1}$ is unbiased, it also has variance $\frac{p(1-p)}{n}$.

For $\hat{p_2}$ we can see that

$$\mathbb{E}[\hat{p_2}] := \mathbb{E}\left[\frac{\sum_{i=1}^{n} Y_i + 1}{n+2}\right] = \frac{np+1}{n+2}$$

Is this estimator biased? Notice that

$$Bias(\hat{p_2}) = \frac{np+1}{n+2} - p$$

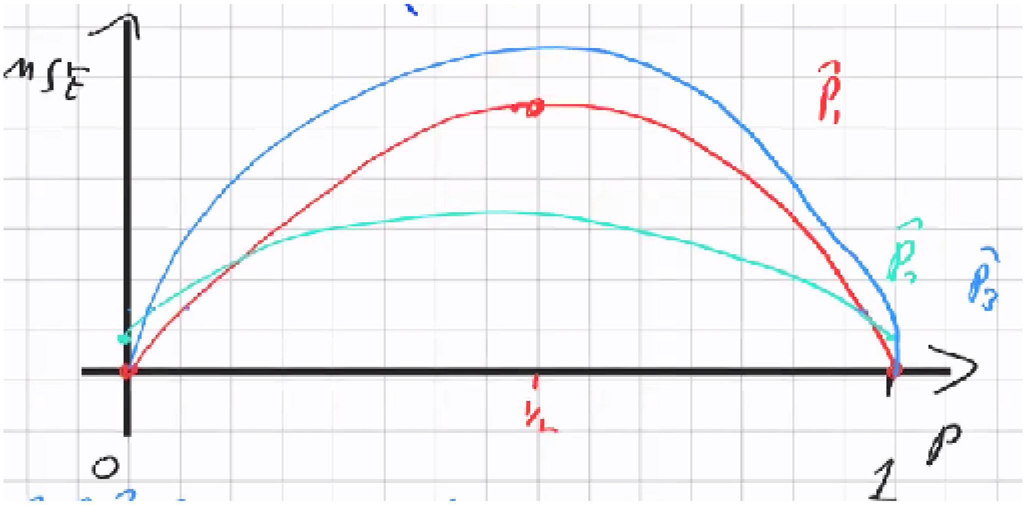
For $p = \frac{1}{2}$ this is 0 and otherwise it isn't and so this estimator is not unbiased. Now we need to check for it's variance.

$$Var(\hat{p_2}) = Var\left(\frac{\sum_{i=1}^{n} Y_i + 1}{n+2}\right) = \frac{np(1-p)}{(n+2)^2}$$

And so the MSE is

$$MSE(\hat{p_2}, p) = \frac{np(1-p)}{(n+2)^2} + (\frac{1-2p}{n+2})^2 = \frac{np - np^2 + 1 - 4p + 4p^2}{(n+2)^2}$$

Now we can graph out the MSE of all the estimators and get something like this



And so for different values of p we get different values of mean square error for each estimator and we might want to choose one over the either depending the values of p.

6.1 Unbiased Estimation

You can see from the previous example that there is a trade-off between a low bias and low variance. For example to get the minimal variance you can just choose a constant regardless of the examples, yet this will tend to maximize the bias and so this isn't a good choice. On the other end you can minimize the bias (in absolute value) by taking the first sampling but this will tend to make the variance very large and so this isn't necessarily a good choice. This means that to get the best results for MSE there has to be some choice that will probably have non-zero bias and non-zero variance but that alltogether minimizes MSE.

Sometimes people still choose to focus on unbiased estimators. We will assume the existence of the expected value from now.

Example 6.2. *Take $Y_1, \ldots, Y_n \sim f_\theta(y)$ i.i.d. random variables with expected value μ, and variance σ^2. Now we can ask whether or not the variance of the sample is an unbiased estimator for the variance. Where the variance of the sample is*

$$\hat{\sigma^2} := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

This means we need to calculate

$$\mathbb{E}\left[\hat{\sigma^2}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[Y_i^2\right]\right) + \frac{1}{n}\mathbb{E}\left[\frac{-2Y_i}{n}\sum_i Y_i\right] + \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\left(\frac{1}{n}\sum_i Y_i\right)\cdot\bar{Y}\right] =$$

$$= \frac{1}{n}\sum_i \mathbb{E}\left[Y_i^2\right] - \frac{2}{n}\mathbb{E}\left[\sum_i Y_i\bar{Y}\right] + \frac{1}{n}\mathbb{E}\left[\sum_i Y_i\bar{Y}\right] = \frac{1}{n}\sum_i \mathbb{E}\left[Y_i^2\right] - \frac{1}{n}\mathbb{E}\left[\sum_i Y_i\bar{Y}\right] =$$

$$= \frac{1}{n}\sum_i \mathbb{E}\left[Y_i^2\right] - \mathbb{E}\left[\bar{Y}^2\right]$$

Notice that

$$\mathbb{E}\left[Y_i^2\right] = \sigma^2 + \mu^2$$

And

$$\mathbb{E}\left[\bar{Y}^2\right] = \frac{\sigma^2}{n} + \mu^2$$

And so

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \sigma^2\left(1 - \frac{1}{n}\right) = \frac{n-1}{n}\sigma^2$$

We can now see that the sample variance isn't an unbiased estimator for the variance. Yet notice that the only thing that is preventing it from being unbiased *is a multiplicative factor. To get an unbiased estimator for the variance we can then take*

$$\hat{S^2} := \frac{n}{n-1}\hat{\sigma}^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{n-1}$$

Notice though that this means that the variance was multiplied by $(\frac{n}{n-1})^2$*, which again means we have a trade off between minimizing the variance and minimizing the bias.*

Problem 6.1: Assume that the expected value, μ, is known and show that $\frac{1}{n}\sum_i (Y_i - \mu)^2$ is an unbiased estimator for σ^2.

Now we can ask the following question - Is there always an unbiased estimator for our values? Notice that uniqueness doesn't happen since we have seen that the sample average and the first sample are both unbiased estimators for the expected value. Now notice that if we have 2 unbiased estimators $\hat{\theta_1}, \hat{\theta_2}$, then every convex sum $\alpha\hat{\theta_1} + (1-\alpha)\hat{\theta_2}$ is also an unbiased estimator. Now we will give an example where there isn't an unbiased estimator.

Example 6.3. *Take the bernoulli example -* $Y_1, \ldots, Y_n \sim Be(p)$*, and look for an estimator for* $\theta := \frac{p}{1-p}$*. Assume* $\hat{\theta}$ *is an estimator for* θ*. And so*

$$\mathbb{E}[\hat{\theta}] = \sum_{y_1,\ldots,y_n \in \{0,1\}^n} \hat{\theta}(\overrightarrow{y})p(\overrightarrow{y}) = \sum_{\overrightarrow{y}} \hat{\theta}(\overrightarrow{y})p^{\sum_i y_i} \cdot (1-p)^{n-\sum_i y_i}$$

This is a polynomial in p *and* θ *doesn't have a full finite taylor polynomial expansion and so it cannot be estimated by an unbiased estimator and so this shows that there isn't always an unbiased estimator for every value.*

6.2 Unbiased Estimator With Uniformly Minimal Variance

Definition 6.1. An estimator $\hat{\theta}$ is called an unbiased estimator with uniformly minimal variance (umvue) if it satisfies the following conditions

(1). $\hat{\theta}$ is an unbiased estimator.

(2). For any other unbiased estimator $\hat{\theta_1}$

$$\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\hat{\theta_1})$$

For all $\theta \in \Theta$.

Notice that there isn't always a umvue since there isn't always an unbiased estimator. If though there is a umvue, then it is unique.

Proof. For 2 umvue's $\hat{\theta}_1, \hat{\theta}_2$, define $\hat{\theta} := \frac{\theta 1 \hat{+} \theta 2}{2}$. We get that

$$\mathrm{Var}(\hat{\theta}) = \frac{\mathrm{Var}(\hat{\theta}_1) + \mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_2) + \mathrm{Var}(\hat{\theta}_1)}{4} \leq \frac{\mathrm{Var}(\hat{\theta}_1) + 2\sqrt{\mathrm{Var}(\hat{\theta}_1)\mathrm{Var}(\hat{\theta}_2)} + \mathrm{Var}(\hat{\theta}_2)}{4} \leq$$

$$\leq \left(\frac{\sqrt{\mathrm{Var}(\hat{\theta}_1)} + \sqrt{\mathrm{Var}(\hat{\theta}_2)}}{2} \right)^2 \leq \left(\max\{\sqrt{\mathrm{Var}(\hat{\theta}_2)}, \sqrt{\mathrm{Var}(\hat{\theta}_1)}\} \right)^2$$

There is equality if and only if

$$\sqrt{\mathrm{Var}(\hat{\theta}_2)} = \sqrt{\mathrm{Var}(\hat{\theta}_1)}$$

And

$$\mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_2) = 2\sqrt{\mathrm{Var}(\hat{\theta}_1)\mathrm{Var}(\hat{\theta}_2)}$$

From the second equation we know that they are a linear function of each other, yet they have the same variance and the same expected value which means that

$$\hat{\theta}_1 = \hat{\theta}_2$$

This proves uniqueness. (Almost surely)

Assuming there is a umvue, is this the best estimator? it turns out that the answer is no. ■

7 Seventh Lecture: Identifying a UMVUE

Theorem 7.1 (Cramer-Rao Bound)**.** *Let there be $Y_1, \dots, Y_n \sim f_\theta(\overrightarrow{y})$ where Θ is one-dimensional. Assume also*

(1). The support of the distribution is independent of θ (meaning where the density or PDF is non-zero is independent of θ).

(2). For every statistic $T = T(\overrightarrow{Y})$ with finite expected value we have the following equalities

$$\frac{d}{d\theta}\mathbb{E}[T] = \frac{d}{d\theta}\int_\Omega T(\overrightarrow{y})f_\theta(\overrightarrow{y})d\overrightarrow{y} = \int_\Omega T(\overrightarrow{y})\frac{d}{d\theta}f_\theta(\overrightarrow{y})d\overrightarrow{y}$$

(The interesting equality is the 2nd since it is the one that provides new information, where you can change the order between integration and derivation).

Under these assumptions we have that for any unbiased estimator with finite variance $\hat{\theta}$ for θ we have that

$$Var(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

Where

$$I(\theta) := \mathbb{E}\left[\left(\frac{d}{d\theta}\log f_\theta(\overrightarrow{y})\right)^2\right]$$

More generally we have that if T is an unbiased estimator for $g(\theta)$ then we have

$$Var(T) \geq \frac{g'(\theta)^2}{I(\theta)}$$

($I(\theta)$ is called Fischer information)

Claim 7.1. Under these assumptions we have

$$\mathbb{E}\left[\frac{d}{d\theta}\log(f_\theta(\overrightarrow{y}))\right] = 0$$

Proof.

$$\mathbb{E}\left[\frac{d}{d\theta}\log(f_\theta(\overrightarrow{y}))\right] = \int \frac{d}{d\theta}\log(f_\theta(\overrightarrow{y}))f_\theta(\overrightarrow{y})d\overrightarrow{y} = \int \frac{\frac{d}{d\theta}f_\theta(\overrightarrow{y})}{f_\theta(\overrightarrow{y})}f_\theta(\overrightarrow{y})d\overrightarrow{y} =$$

$$= \int \frac{d}{d\theta}f_\theta(\overrightarrow{y})d\overrightarrow{y} = \frac{d}{d\theta}\int f_\theta(\overrightarrow{y})d\overrightarrow{y} = \frac{d}{d\theta}1 = 0$$

This means that the Fischer information is just the variance of the derivative log-likelihood function ■

Proof Of Cramer-Rao Bound. Let there be an unbiased estimator T for $g(\theta)$. We know that

$$g(\theta) = \mathbb{E}[T] = \int T(\overrightarrow{y})f_\theta(\overrightarrow{y})d\overrightarrow{y}$$

Meaning

$$g'(\theta) = \frac{d}{d\theta}\int T(\overrightarrow{y})f_\theta(\overrightarrow{y})d\overrightarrow{y} = \int T(\overrightarrow{y})\frac{d}{d\theta}f_\theta(\overrightarrow{y})d\overrightarrow{y} = \mathbb{E}\left[T\cdot\frac{d}{d\theta}\log(f_\theta(\overrightarrow{y})\right] =$$

$$= \mathrm{Cov}(T(\overrightarrow{y}), \frac{d}{d\theta}\log f_\theta(\overrightarrow{y})) + \mathbb{E}[T]\cdot\mathbb{E}[\frac{D}{d\theta}\log f_\theta(\overrightarrow{y})] = \mathrm{Cov}(T(\overrightarrow{y}), \frac{d}{d\theta}\log f_\theta(\overrightarrow{y}))$$

We know that for general random variables X,Y that

$$\mathrm{Cov}(X,Y) \leq \sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}$$

Meaning

$$g'(\theta) \leq \sqrt{\mathrm{Var}(T(\overrightarrow{y}))\cdot\mathrm{Var}(\log f_\theta(\overrightarrow{y}))} = \sqrt{\mathrm{Var}(T(\overrightarrow{y}))\cdot I(\theta)}$$

And so we can see that

$$\mathrm{Var}(T) \geq \frac{g'(\theta)^2}{I(\theta)}$$

■

Claim 7.2. If $Y_1, \ldots, Y_n \sim f_\theta(\overrightarrow{y})$ are a series of i.i.d. random variables with the assumptions above. Define

$$I^* := \mathbb{E}\left[\left(\frac{d}{d\theta}\log f_\theta(\overrightarrow{y})\right)^2\right]$$

We have that

$$I(\theta) = nI^*(\theta)$$

Example 7.1. *Assume $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ where we know σ^2 (meaning the parameter space is only for the mean). We are looking for an unbiased estimator for μ. Let's check the regularity conditions that appear in the Cramer-Rao bound -*

(1). The support is $\mathbb{R}$ for all θ, which implies independence.

(2). We need to show that we can exchange derivation by μ and integration according to y. This is true since the dependence of $f_\mu(\theta)$ on y and μ is negative super-exponential and so it is a fact that we have all the related convergences which also covers this condition.

Now we want to calculate the Cramer-Rao bound. The density is

$$f_\mu(y) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}$$

And so

$$\log(f_\mu(\theta)) = -\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2$$

Meaning

$$\frac{d}{d\mu}\log(f_\mu(\theta)) = \frac{y-\mu}{2\sigma^2} = \frac{y-\mu}{\sigma^2}$$

And so

$$I^*(\mu) = \mathbb{E}\left[\frac{(y-\mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^4}\mathbb{E}[(y-\mu)^2] = \frac{1}{\sigma^2}$$

And so alltogether

$$I(\mu) = \frac{n}{\sigma^2}$$

This tells us that for every unbiased estimator T for μ we have

$$MSE(T) \geq \frac{\sigma^2}{n}$$

We saw that the sample average is an unbiased estimator with this above MSE and so it is a UMVUE.

claim If in addition to the regularity assumptions from the bound we also assume that $\log f_\theta(\overrightarrow{y})$ is twice-differentiable then

$$I(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2}(\log f_\theta(\overrightarrow{y}))\right]$$

Proof.

$$\frac{d^2}{d\theta^2}(\log f_\theta(\overrightarrow{y})) = \frac{d}{d\theta}\left[\frac{\frac{d}{d\theta}f_\theta(\overrightarrow{y})}{f_\theta(\overrightarrow{y})}\right] = \frac{(\frac{d^2}{d\theta^2}f_\theta(\overrightarrow{y}))\cdot f_\theta(\overrightarrow{y}) - (\frac{d}{d\theta}f_\theta(\overrightarrow{y})^2}{f_\theta^2(\overrightarrow{y})}$$

We now want to calculate the expected value of all of this. If we calculate the first part we have

$$\mathbb{E}\left[\frac{\frac{d^2}{d\theta^2}f_\theta(y)}{f_\theta(\overrightarrow{y})}\right] = \int \frac{\frac{d^2}{d\theta^2}f_\theta(y)}{f_\theta(\overrightarrow{y})}f_\theta(\overrightarrow{y})d\overrightarrow{y} = \int \frac{d^2}{d\theta^2}f_\theta(\overrightarrow{y})d\overrightarrow{y} = \frac{d^2}{d\theta^2}1 = 0$$

Meaning that all we are left with is

$$\mathbb{E}\left[-\frac{(\frac{d^2}{d\theta^2}f_\theta(\overrightarrow{y}))^2}{f_\theta^2(\overrightarrow{y})}\right] = I(\theta)$$

And this ends the proof. ■

Example 7.2. *We can now look for an unbiased estimator for the variance of a normal distribution. Meaning we now consider the parameter space to be the variance while the mean is a constant.*

Remember

$$\log(f_\sigma(y)) = -\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2$$

And so

$$\frac{d}{d\sigma}\log f_\sigma(y) = -\frac{1}{\sigma} + \frac{(y-\mu)^2}{\sigma^3}$$

Differentiating again gives us

$$\frac{d^2}{d\sigma^2}\log f_\sigma(y) = \frac{1}{\sigma^2} - 3\frac{(y-\mu)^2}{\sigma^4}$$

Meaning

$$\mathbb{E}\left[\frac{d^2}{d\sigma^2}\log f_\sigma(y)\right] = \frac{1}{\sigma^2} - 3\frac{1}{\sigma^2} = \frac{-2}{\sigma^2}$$

And so

$$I(\theta) = n \cdot -\frac{-2}{\sigma^2} = \frac{2n}{\sigma^2}$$

We want to estimate σ^2 and so define $g(\sigma) = \sigma^2$ and see that for every unbiased estimator (using the bound from the beginning of the lecture) we have that

$$MSE(T) \geq \frac{(2\sigma)^2}{\frac{2n}{\sigma^2}} = \frac{2\sigma^4}{n}$$

Remember that the unbiased estimator we have found for the variance is

$$\hat{s^2} = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2$$

We will want to compare $MSE(\hat{s^2})$ to the bound we have found in the next lecture.

8 Eighth Lecture

Last lecture we saw that with a normal distribution, any unbiased estimate of the variance has

$$\mathrm{MSE}(T) \geq \frac{2\sigma^4}{n}$$

We now want to see how the unbiased estimator for the variance we have found in the past, $\hat{s^2}$, compares with this. Notice that since Y_i is a normal distribution, and the average is also a normal distribution (since the samples are independent) then $Y_i - \bar{Y}$ is also a normal distribution. And so

$$\sum_i \left(Y_i - \bar{Y}\right)^2 \sim \sigma^2 \chi^2_{n-1}$$

It is known that when $Z \sim \chi^2_k$ the variance is 2k. And so

$$\mathrm{Var}(\hat{s^2}) = \left(\frac{\sigma^2}{n-1}\right)^2 \cdot 2(n-1) = \frac{2\sigma^4}{n-1}$$

And so this doesn't achieve the lower bound that the Cramer Rao bound achieves so we can't immediately know whether this is or isn't a UMVUE for the variance (It is, this has to be proved in a different or more specific way).

8.1 The Family Of Exponential Distributions

Definition 8.1. Assuming the coefficient space Θ is an open interval.

We say that a distribution $f_\theta(\overrightarrow{y})$ is in the family of exponential distribution if and only if all of the following happen -

(1). The support of the $f_\theta(\overrightarrow{y})$ is independent of the choice of θ.

(2). The density can be written in the following way

$$f_\theta(\overrightarrow{y}) := \begin{cases} \exp(c(\theta)T(\overrightarrow{y}) + d(\theta) + s(\overrightarrow{y})), & \overrightarrow{y} \in \text{Supp} \\ 0, & \text{Otherwise} \end{cases}$$

$c(\theta)$ in the above definition is called the natural parameter of the distribution.

Notice that this isn't a definition that is hard to extend to discrete distributions.

Example 8.1. *We can take the exponential distribution -*

$$f_\theta(\overrightarrow{y}) := \theta e^{-\theta t}, \text{ for } \theta \in [0, \infty)$$

Notice that the support is all non-negative numbers and is therefore independent of θ. And it is easy to verify that this distribution can be written in the above way and so it is part of the family of exponential distributions.

Example 8.2. *Taking $n \in \mathbb{N}$ constant we can consider the distribution Bin?(n, p). Its support is $\{0, 1, \dots, n\}$ which is independent of p. What we need to verify is that it can be written in the fashion that is described in the definition*

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} = (1-p)^n \binom{n}{k} \left(\frac{p}{1-p}\right)^k =$$

$$= \exp\left(k \log\left(\frac{p}{1-p}\right) + \log\left(\binom{n}{k}\right) + n \log(1-p)\right)$$

Therefore this is in the exponential distribution. Notice that this is only when n is constant, if we allow n to not be constant the problem we get is that the support is not independent of the parameters and so it doesn't fall into this exponential family distribution.

Example 8.3. *Taking a distribution that follows a power law -*

$$\mathbb{P}(X = k) = \frac{1}{\zeta(\gamma)} k^{-\gamma}, \text{ for } k \in \mathbb{N}$$

This is defined over the coefficient space $\Gamma = (1, \infty)$ (otherwise the distribution itself isn't well-defined). And it's support is $\mathbb{N}$ which is of course is independent of γ. Now notice that

$$\mathbb{P}(X = k) = \exp\left(-\gamma \log k + \log \zeta(\gamma)\right)$$

Most of the regular discrete distributions are also in this family. For example - bernoulli, binomial (with n being constant), poisson, geometric and so on

Additionally most of the regular continuous distributions belong to this family - normal (where at least one of the parameters are known), exponential and so on.

We can now consider a case where the space of coefficients are multi-dimensional.

Definition 8.2. Assume that Θ is an open subset of $\mathbb{R}^p$. We will say that a distribution $f_{\overrightarrow{\theta}}(\overrightarrow{y})$ is in the family of exponential distributions if all of the following occur

(1). The support is independent of $\overrightarrow{\theta}$.

(2). The distribution can be written in the following way

$$f_{\overrightarrow{\theta}}(\overrightarrow{y}) = \exp\left(\sum_{j=1}^{k} c_j(\overrightarrow{\theta})T_j(\overrightarrow{y}) + d(\overrightarrow{\theta}) + S(\overrightarrow{y})\right)$$

Here too we call $c_1(\overrightarrow{\theta}), \dots, c_k(\overrightarrow{y})$ the natural parameter of the distribution.

Example 8.4. *We can now take $\mathcal{N}(\mu, \sigma^2)$ the normal distribution. The parameter space is $\Theta = \mathbb{R} \times (0, \infty)$. The support is all $\mathbb{R}$, which is independent of the coefficients. Now we want to check if we can express the density function in the way the definition needs*

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) = \exp\left(-\frac{y^2 - 2y\mu + \mu^2}{2\sigma^2} - \log\sigma - \frac{1}{2}\log(2\pi)\right)$$

This can be decomposed pretty easily to get what we want and so this is in the family of exponential distributions.

Exercise 8.1. *Show that a multi-variate normal distribution with $(\overrightarrow{\mu}, \Sigma)$ is also in the family of exponential distributions. What is k?*

Exercise 8.2. *Given $Y \sim \mathcal{N}(\mu, (a\mu)^2)$ for some constant $\alpha > 0$, In what type of family of exponential distributions is this distribution? If in the multi-dimensional case what is k?*

GMM, gaussian mixture model which is a widely-used distribution in many areas, is not in the family of exponential distributions.

Theorem 8.1. *Assume that we know that $f_\theta(\overrightarrow{y})$ is in the one-dimensional family of exponential distributions, and that the natural parameter $c(\theta)$ is continuously differentiable and the derivative doesn't hit zero. Taking the representation from the original definition we have that $T(\overrightarrow{y})$ is a UMVUE for $\mathbb{E}_\theta[T(\overrightarrow{y})]$ that attains the Cramer-Rao bound.*

On the other hand if $f_\theta(\overrightarrow{y})$ is a distribution dependent on θ whose support is independent of θ. Assume there is an unbiased estimator $T(\overrightarrow{Y})$ for $g(\theta)$. If $T(\overrightarrow{Y})$ attains the Cramer-Rao bound for all θ then $f_\theta(\overrightarrow{y})$ is in the one dimensional family of exponential distributions.

We will continue this in the next lecture

9 Ninth Lecture

9.1 Multidimensional Cramer-Rao Bound

When we first talked about this bound we had the case of the one-dimensional bound. Now we will consider the multi-dimensional case. For this assume that the parameter space is p-dimensional. Now define the Fischer information matrix to be a p-dimensional matrix -

$$\forall\ j,k : (I(\theta))_{j,k} = \mathbb{E}_\theta \left[\left(\frac{d}{d\theta_j} \log f_{\overrightarrow{\theta}}(\overrightarrow{y}) \right) \cdot \left(\frac{d}{d\theta_j} \log f_{\overrightarrow{\theta}}(\overrightarrow{y}) \right) \right]$$

Theorem 9.1 (Cramer-Rao Bound)**.** *Assume the sampling $\overrightarrow{Y}$ has joint density function $f_{\overrightarrow{\theta}}(\overrightarrow{y})$. Assume Θ is p-dimensional, and that the following conditions all occur -*

(1). The support of the joint density function is independent of the choice of θ.

(2). The following equalities occur always -

$$\forall\ 1 \le j \le p : \frac{d}{d\theta_j} \int T(\overrightarrow{y}) f_\theta(\overrightarrow{y}) d\overrightarrow{y} = \int T(\overrightarrow{y}) \frac{d}{d\theta_j} f_\theta(\overrightarrow{y}) d\overrightarrow{y}$$

Let T be an unbiased estimator for a scalar function $g : \mathbb{R}^p \to \mathbb{R}$, with finite variance. Then we have

$$Var(T) \ge \left(\frac{dg(\overrightarrow{\theta})}{d\overrightarrow{\theta}} \right)^T \cdot (I(\theta))^{-1} \cdot \left(\frac{dg(\overrightarrow{\theta})}{d\overrightarrow{\theta}} \right)$$

$$\left(Where\ \frac{dg(\overrightarrow{\theta})}{d\overrightarrow{\theta}} := \left(\frac{dg(\overrightarrow{\theta})}{d\theta_1}, \ldots, \frac{dg(\overrightarrow{\theta})}{d\theta_p} \right) \right)$$

Proof. We will skip this proof, this is a generalization of the one-dimensional proof. ■

Note that when the samplings are i.i.d. we can again define $I^*(\theta)$ and get that

$$I(\theta) = n \cdot I^*(\theta)$$

Additionally if there is a second derivative and integration and second differentiation can switch we have that

$$I(\theta) = \mathbb{E} \left[\frac{d^2}{d\overrightarrow{\theta}(d\overrightarrow{\theta})^T} \log f_{\overrightarrow{\theta}}(\overrightarrow{y}) \right]$$

Example 9.1. *Consider a i.i.d. sampling of normal distributions $\mathcal{N}(\mu, \sigma^2)$. As we have previously calculated*

$$I_{(\mu,\mu)} = \frac{n}{\sigma^2},\ I_{(\sigma,\sigma)} = \frac{2n}{\sigma^2}$$

Now we want to calculate $I_{(\mu,\sigma)}$, what we want therefore is -

$$\frac{d^2}{d\mu d\sigma}\log f(y)=\frac{d^2}{d\mu d\sigma}\left(-\log\sqrt{2\pi}-\log\sigma-\frac{(y-\mu)^2}{2\sigma^2}\right)=\cdots=-\frac{2(y-\mu)}{\sigma^3}$$

Meaning

$$I^*_{(\mu,\theta)}=\mathbb{E}\left[-\frac{2(y-\mu)}{\sigma^3}\right]=0$$

Alltogether we have

$$I(\overrightarrow{\theta})=\begin{bmatrix}\frac{n}{\sigma^2} & 0\\ 0 & \frac{2n}{\sigma^2}\end{bmatrix}$$

9.2 Approaching UMVUE

Theorem 9.2 (Blackwell-Rao Theorem)**.** *Let T be an unbiased estimator for θ, and W be a sufficient statistic for θ. Now define*

$$T_1:=\mathbb{E}[T\mid W]$$

We have the following happen:

(1). T_1 is an unbiased estimator for θ.

(2). $Var(T_1)\leq Var(T_2)$.

Note that the sufficiency of W is necessary for T_1 to even be a statistic - since W is a sufficient statistic we have that $T\mid W$ is independent of the parameters and so T_1 is also independent of the parameters which tells us that it is a statistic

Proof. We can show that T_1 is in fact unbiased and that is because -

$$\mathbb{E}[T_1]=\mathbb{E}[\mathbb{E}[T\mid W]]=\mathbb{E}[T]=\theta$$

Now we want to compare the variances -

$$\mathrm{Var}(T)=\mathrm{Var}\left(\mathbb{E}[T\mid W]\right)+\mathbb{E}\left[\mathrm{Var}(T\mid W)\right]=\mathrm{Var}(T_1)+\mathbb{E}[\mathrm{Var}(T\mid W)]\geq\mathrm{Var}(T_1)$$

And so we are done. ■

Notice that this means that if $\mathrm{Var}(T\mid W)$ is non-zero at least sometimes that means that the inequality turns to a strict inequality. Also you can see that it might be wise to pick W to be a minimal sufficient statistic. Also note that the process stops after one iteration because of how conditional expected value works.

Claim 9.1. $T_1=T$ if and only if T is a function of W. (with probability 1)

This is a result using the proof of the previous theorem

Example 9.2. *Take $Y_i \sim Ber(p)$ i.i.d., we know that $\sum_i y_i$ is a sufficient statistic, and take $T = Y_1$. Now let's apply the theorem -*

$$T_1\mathbb{E}\left[Y_1 \mid \sum_i Y_i\right] = 1\cdot\mathbb{P}\left(Y_1 = 1 \mid \sum_i Y_i\right) + 0\cdot\mathbb{P}\left(Y_i = 0 \mid \sum_i Y_i\right) = \frac{\sum_i Y_i}{n} = \bar{Y}$$

Which is the UMVUE for this distribution.

What are the weaknesses of this process?

(1). It can be hard to find an initial unbiased estimate.

(2). It is sometimes very hard to evaluate $\mathbb{E}[T \mid W]$

Example 9.3. *When it comes to earthquakes, the number of earthquakes in a large enough interval time usually models as a Poisson distribution. Therefore if we have i.i.d. samples $Y_i \sim Poi(\lambda)$, we will be interested in estimating -*

$$\mathbb{P}(Y \geq 1) = 1 - e^{-\lambda}$$

Which is the probability that an earthquake happens in that interval of time. We have seen previously that for poisson distributions $\sum_i Y_i$ is a sufficient statistic. Now define an T to be an indicator function in the following way -

$$T := \begin{cases} 1, & Y_1 \geq 1 \\ 0, & Otherwise \end{cases}$$

Obviously T is an unbiased estimator for the event $\{Y \geq 1\}$.

Now we can apply he process from the theorem. And so

$$T_1 := \mathbb{E}[T \mid W] = 1\cdot\mathbb{P}\left(Y_1 \geq 1 \mid \sum_i Y_i = k\right) + 0\cdot\dots = 1 - \mathbb{P}\left(Y_i = 0 \mid \sum_i Y_i = k\right) =$$

$$= 1 - \frac{\mathbb{P}(Y_i = 0, \sum_i Y_i = k)}{\mathbb{P}(\sum_i Y_i = k)} = 1 - \frac{\mathbb{P}(Y_1 = 0)\mathbb{P}(\sum_{i=2}^n Y_i = k)}{\mathbb{P}(\sum_i Y_i = k)} = 1 - \frac{e^{-\lambda} \cdot e^{-(n-1)\lambda} \cdot \frac{((n-1)\lambda)^k}{k!}}{e^{-n\lambda} \cdot \frac{(n\lambda)^k}{k!}} =$$

$$= 1 - \left(\frac{n-1}{n}\right)^{\sum_i Y_i}$$

This is obviously not a trivial choice for an estimator. If we would have used MLE we would have gotten that the estimator would be

$$1 - e^{-\bar{Y}} = 1 - e^{-\frac{\sum_i Y_i}{n}}$$

Notice that these are very very close, where what we found is a UMVUE but the second one isn't necessarily even unbiased. And so it remains to see which one is "better".

9.3 Lehmann-Scheffe Theorem

Claim 9.2. If W is a complete statistic then there is at most 1 (with probability 1) unbiased estimators for $g(\theta)$ that is a function of W

Proof. By definition of completeness, just expand and get what you want. ■

Theorem 9.3 (Lehmann - Scheffe)**.** *If W is a sufficient and complete statistic for θ, and T is an unbiased estimator then*

$$T_1 = \mathbb{E}[T \mid W]$$

Is a UMVUE for θ.

Proof. From Blackwell-Rao theorem we have that T_1 is an unbiased estimator for θ. Now we need to show that its variance is minimal. Let S be an unbiased estimator for θ. We know that

$$\mathrm{Var}(\mathbb{E}[S \mid W]) \leq \mathrm{Var}(S)$$

Obviously $\mathbb{E}[S \mid W]$ is a function of W and so this has to be T_1 (with probability 1) which tells us that T_1 has a lower variance. ■

Example 9.4. *Consider the estimator $\hat{s^2}$ as an estimator in a normal distribution. We have seen that*

$$W := \left(\bar{Y}, \sum_i (Y_i - Y)^2\right)$$

is a complete and sufficient statistic for μ, σ^2. (we have seen sufficient, completeness can be seen by hand similar to things in the past). Obviously $\hat{s^2}$ is a function of W and so it is a UMVUE.

10 Tenth Lecture

10.1 Interval Estimation

Definition 10.1. Given n samples $\overrightarrow{Y}_1, \dots, \overrightarrow{Y}_n \sim f_\theta(\overrightarrow{Y})$ and $\alpha > 0$ and 2 statistics L, U such that

$$\mathbb{P}(L \leq \theta \leq U) = 1 - \alpha$$

We then say that $[L, U]$ is a confidence interval with significance α for θ.

Hisotrically The values for α that are usually chosen from $\{0.05, 0.01, 0.001\}$, even though theoretically it can be any value in $(0, 1)$.

Example 10.1. *Given $Y \sim [0, \theta]$, for all $t \leq \alpha$ the interval -*

$$\left[\frac{Y}{1-\alpha+t}, \frac{Y}{t}\right]$$

is a confidence interval with significance α for θ. This can be seen by the fact that

$$\mathbb{P}\left(\theta \geq \frac{Y}{1-\alpha+t} \wedge \theta \leq \frac{Y}{t}\right) = \mathbb{P}\left(t\theta \leq Y \leq \theta - \alpha\theta + t\theta\right) = \dots = 1 - \alpha$$

Definition 10.2. Often it is demanded that we have symmetry in the following sense -

$$\mathbb{P}(\theta < L) = \mathbb{P}(\theta > U)$$

And so returning to this previous example we have that -

$$\mathbb{P}(Y < t\theta) = \mathbb{P}(Y > \theta - \alpha\theta + t\theta)$$

Meaning

$$t = \alpha - t$$

Alltogether

$$t = \frac{1}{2}\alpha$$

This means we want to take the confidence interval to be

$$\left[\frac{2Y}{2-\alpha}, \frac{2Y}{\alpha}\right]$$

Definition 10.3. Under a Baysien world view we assume that there is a probability space on the constant space. And so given $\alpha > 0$ we look for L, U statistics such that

$$\mathbb{P}(L \leq \theta \leq U) = 1 - \alpha$$

And then we say that $[L, U]$ is a credible interval with confidence α for θ.

Definition 10.4. Given samples $Y_1, \ldots, Y_n$ samples we look for L, U such that

$$\mathbb{P}(L \leq Y_{n+1} \leq U) = 1 - \alpha$$

This is called a prediction interval

10.2 Finding a confidence interval

Definition 10.5. A function of the samples and the parameter

$$\psi\left(\overrightarrow{Y}, \theta\right)$$

such that it's distribution is independent of the parameter is called a pivot.

Given a pivot we look for an interval A_α such that

$$\mathbb{P}\left(\psi\left(\overrightarrow{Y}, \theta\right) \in A_\alpha\right) = 1 - \alpha$$

Given a sampling then $\overrightarrow{y}$ we will invert (not trivial) the pivot and rescue θ to get an interval

$$C_\alpha := \{\theta \mid \psi(\overrightarrow{y}, \theta) \in A_\alpha\}$$

Example 10.2. *Given a sampling $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. with known σ^2. And so we know that*

$$\bar{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

And so we can define a pivot

$$\psi\left(\overrightarrow{Y}, \mu\right) = \bar{Y} - \mu$$

Now choose the intervals in the following way (using symmetry of normal distributions)

$$\mathbb{P}\left(|\bar{Y} - \mu| \leq z_{\frac{1}{2}\alpha} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Where

$$z_\alpha$$

is the inversion of the CDF of the normal distribution. Then we get that the confidence interval is

$$C_\alpha = \left[\bar{y} - z_{\frac{1}{2}\alpha} \cdot \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\frac{1}{2}\alpha} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Claim 10.1. Given $X \sim F_X(x)$, where F is the CDF. We have that

$$\mathbb{P}(F_X(X) \leq \alpha) = \alpha$$

For an i.i.d. sample $Y_1, \ldots, Y_n$ the following is a pivot -

$$\prod_{i=1}^{n} F_\theta(y_i)$$

It is custom to take the negative log of this as a pivot. Meaning

$$-\sum \log F_\theta(y_i)$$

This gives us

$$-\sum \log F_\theta(y_i) \sim \frac{1}{2}\chi^2_{2n}$$

Notice that often it is very hard to rescue θ from the construction we have.

Definition 10.6. We say that f_X belongs to the scale-location family of distributions if f is controlled by 2 parameters μ and $\sigma > 0$ such that

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma} f_{0,1}\left(\frac{X-\mu}{\sigma}\right)$$

Where $f_{0,1}$ is called the standard distribution of the family.

Example 10.3. *Normal distribution, the "standardization" can be seen to be a pivot.*

Claim 10.2. Let $T(\overrightarrow{Y})$ be a sufficient statistic for θ whose distribution is f_T which is uni-modal, then there exists a confidence interval with significance α for θ with minimal expected length of interval based on the information from $T(\overrightarrow{y})$. This is

$$\{x \mid f_T(x) > c_\alpha\}$$

Where c_α is determined in the natural way

Proof. We will show that this is minimal. Denote (a_1, b_1) to be the confidence interval from the claim, and assume (a_2, b_2) is a different confidence interval with the same significance. We will consider the following intervals -

(1). $(a_1, b_1) \cap (a_2, b_2)$.

(2). $(a_1', b_1') = (a_1, b_1) \setminus (a_2, b_2)$.

(3). $(a_2', b_2') = (a_2, b_2) \setminus (a_1, b_1)$.

Notice that (we can also do this with (a_2', b_2'))

$$\int_{[a_1', b_1']} f_T(x)dx = 1 - \alpha - \int_{(a_1,b_1)\cap(a_2,b_2)} f_T(x)dx$$

Notice that

$$\int_{(a_1',b_1')} f_T(x)dx > \int_{(a_1',b_1')} c_\alpha dx = c_\alpha(b_1' - a_1')$$

Similarly

$$\int_{(a_2',b_2')} f_T(x)dx \leq \int_{(a_2',b_2')} c_\alpha dx = c_\alpha(b_2' - a_2')$$

Using this we can interpolate

$$b_1' - a_1' < \frac{1}{C_\alpha} \int_{(a_1',b_1')} f_T(x)dx = \frac{1}{c_\alpha} \int_{(a_2',b_2')} f_T(x)dx \leq \frac{1}{c_\alpha} \cdot c_\alpha(b_2' - a_2') = b_2' - a_2'$$

And so we have shown minimality. ■

11 Eleventh Lecture: Hypothesis Testing

11.1 Terminology

We start with a sampling of $\overrightarrow{Y} \sim f_\theta(\overrightarrow{y})$. We then get 2 **hypotheses -** H_0 (this is called the null hypothesis), and H_1 (alternative hypothesis). In theory these are interchangeable but, as the names suggest, traditionally H_0 is the "current thinking" or the one that is easier to check for and H_1 is the hypothesis that we want to check and compare. It is then widely said we "reject" or "accept" only about H_1.

What these hypotheses are in a sense are just the sets of coefficients that we find "usable" in our model. Formally

$$H_0 = \Theta_0 \subset \Theta,\ H_1 = \Theta_1 \subset \Theta_1 \text{ where } \Theta_1 \cap \Theta_2 = \emptyset$$

Hypotheses with one coefficient possible are called simple, and otherwise composite.

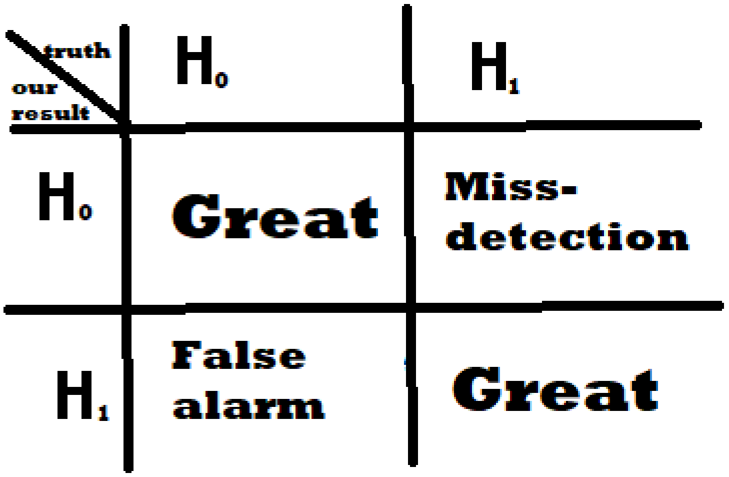
Definition 11.1. A test statistic is a statistic where it gets "low" values for the null hypothesis, and "high" for the alternative hypothesis.

Example 11.1. *If we take the birth sampling example the null hypothesis will naturally be that with equal probability females and males. And so it is natural to choose the following test statistic -*

$$T(\overrightarrow{Y}) = |\bar{Y} - 0.5|$$

In this example it is then natural to choose $C > 0$ such that if $T(\overrightarrow{Y}) > C$ we accept H_1 and otherwise we reject the alternative hypothesis. This C is called the critical value. If there isn't one emergent critical value for what we want it is accepted that you augment the statistic until there is a singular critical value.

Now there is additional terminology



truth / our result	H_0	H_1
H_0	Great	Miss-detection
H_1	False alarm	Great

Where miss-detection is an error of the second type, and false alarm is an error of the first type. The reason there is a distinction is again for real life reasons. We will assume for the lecture that both hypotheses are simple. We want to develop a strategy to choose which hypothesis is better. The probability to be wrong is the probability of having a type-1 error and the probability of having a type-2 error. These are

$$\mathbb{P}_{\theta_0}(T \geq C) \text{ for a type-1 error, } \mathbb{P}_{\theta_1}(T < C) \text{ for a type-2 error}$$

Now we denote

$$\Omega_0 := \{\overrightarrow{y} \mid T(\overrightarrow{y}) < C\},\, \Omega_1 := \{\overrightarrow{y} \mid T(\overrightarrow{y}) \geq C\}$$

Definition 11.2. We define the significance of a test statistic to be the probability that we get an error of type 1 -

$$\alpha := \mathbb{P}_{\theta_0}(\text{"Accept "}H_1) = \mathbb{P}_{\theta_0}(\overrightarrow{Y} \in \Omega_1)$$

And we define β to be

$$\beta := \mathbb{P}_{\theta_1}(\text{"Reject "}H_1) = \mathbb{P}_{\theta_1}(\overrightarrow{Y} \in \Omega_0)$$

The power of the test is

$$\pi = 1 - \beta$$

Example 11.2. *Assume there is a claim that a car gets 15 kilometers per liter, and the competition claims that their car gets 12 kilometers per liter. Let's assume a model where the sampling is $Y_i \sim \mathcal{N}(\mu, 4)$ where*

$$H_0 : \mu = 12, \;\; H_1 : \mu = 15$$

Of course it makes sense to choose $\bar{Y}$ as the test statistic. If for example we have 5 samplings we have that

$$\bar{Y} \sim \mathcal{N}(\mu, \frac{4}{5})$$

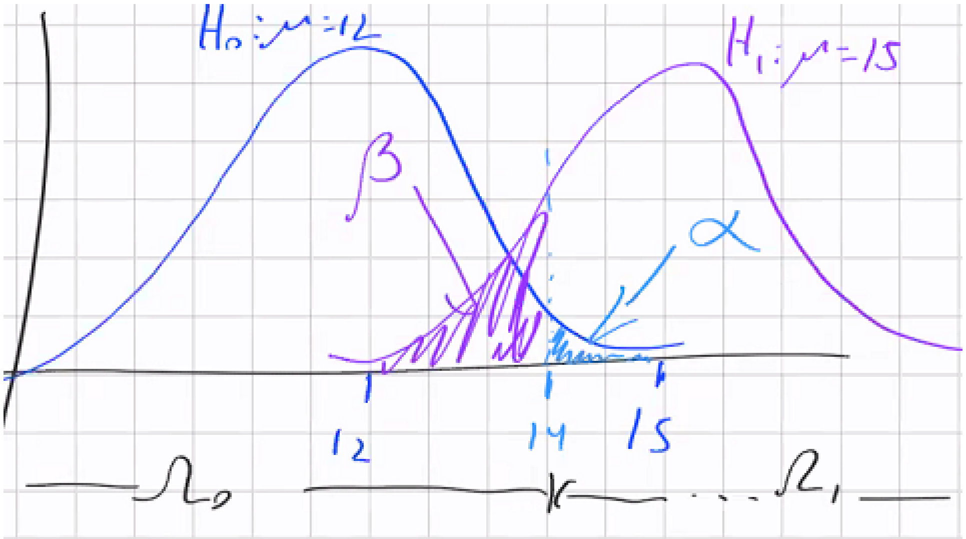
We now want to calculate α. If we for example take $C = 14$

$$\alpha = \mathbb{P}_{\mu=12}(Y \geq 14) = 1 - \Phi\left(\frac{14-12}{\sqrt{0.8}}\right) = 0.0127$$

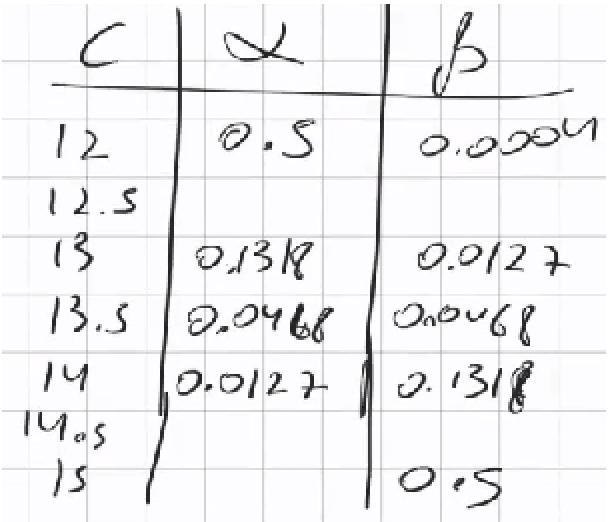
And we have that

$$\beta = \mathbb{P}_{\mu=15}(Y < C) = \Phi\left(\frac{14-15}{\sqrt{0.8}}\right) = 0.1318$$

This is all exemplified by the following image



Now we can wonder what critical value is good. We can consider a few values for the critical value and see how it affects α and β. We can get the following table



c	α	β
12	0.5	0.0004
12.5		
13	0.1318	0.0127
13.5	0.0468	0.0468
14	0.0127	0.1318
14.5		
15		0.5

Regularly α is chosen arbitrarily at 0.05.

11.1.1 p-value

Given a sampling $\overrightarrow{y}$ we can calculate

$$t_{\text{obs}} = T(\overrightarrow{y})$$

Definition 11.3. The p-value of an experiment is

$$\text{p-value} := \mathbb{P}_{H_0}\left(T(\overrightarrow{Y}) \geq t_{\text{obs}}\right)$$

And a lot of times it is demanded that the p-value be less than the previous threshold.

11.2 Tests With Maximal Power

For a given level of α it is natural to take a statistic that maximizes the power of the test, over all of the tests with a probability of error of type-1 $\leq \alpha$.

Definition 11.4. Given a sampling $\overrightarrow{Y} \sim f_\theta(\overrightarrow{y})$, and 2 simple hypotheses

$$H_0 : \theta = \theta_0,\ H_1 : \theta = \theta_1$$

The likelihood ratio is

$$\lambda(\overrightarrow{y}) := \frac{L(\theta_1; \overrightarrow{y})}{L(\theta_0; \overrightarrow{y})}$$

The likelihood ratio test (LRT), with significance α is if given a critical value of C we have that

$$\mathbb{P}_{\theta_0}(\lambda(\overrightarrow{y}) \geq C) = \alpha$$

Lemma 11.1 (Neyman-Pearson)**.** *Under these conditions LRT is a test with maximal power*

Proof. Will be shown next lecture. ■

12 Twelfth Lecture

12.1 LRT For Simple Hypotheses

Lemma 12.1 (Neyman - Pearson)**.** *Given a sampling* $\overrightarrow{Y} \sim f_\theta(\overrightarrow{y})$ *and 2 simple hypotheses -*

$$H_0 : \theta = \theta_0, \ \text{and } H_1 : \theta = \theta_1$$

The likelihood ratio test with the rejection area

$$\Omega_1 := \{\overrightarrow{y} \mid \lambda(\overrightarrow{y}) := \frac{f_{\theta_1}(\overrightarrow{y})}{f_{\theta_0}(\overrightarrow{y})} \geq C\}$$

Is a maximal power test at significance level $\alpha = \mathbb{P}_\theta(\lambda(\overrightarrow{Y}) \geq C)$

Proof. We will prove for continuous random variables, but it is extendable to discrete ones. It now makes sense to calculate the power of the LRT.

$$\pi := \mathbb{P}_{\theta_1}(\lambda(\overrightarrow{Y}) \geq C) = \mathbb{P}(\overrightarrow{Y} \in \Omega_1) = \int_{\Omega_1} f_{\theta_1}(\overrightarrow{y}) d\overrightarrow{y}$$

Now assume we have another test with significance $\alpha' \leq \alpha$ and with power π' and rejection region Ω_1'. Again we have

$$\pi' := \mathbb{P}_{\theta_1}(\lambda(\overrightarrow{Y}) \geq C) = \mathbb{P}(\overrightarrow{Y} \in \Omega_1') = \int_{\Omega_1'} f_{\theta_1}(\overrightarrow{y}) d\overrightarrow{y}$$

Also

$$\alpha' := \mathbb{P}_{\theta_0}(\overrightarrow{Y} \in \Omega_1') = \int_{\Omega_1'} f_{\theta_0}(\overrightarrow{y}) d\overrightarrow{y}$$

We now want to show that $\pi \geq \pi'$:

$$\pi - \pi' = \int_{\Omega_1} f_{\theta_1}(\overrightarrow{y}) d\overrightarrow{y} - \int_{\Omega_1'} f_{\theta_1}(\overrightarrow{y}) d\overrightarrow{y} =$$

$$= \int_{\Omega_0' \cap \Omega_1} f_{\theta_1}(\overrightarrow{y}) d\overrightarrow{y} + \int_{\Omega_1' \cap \Omega_1} f_{\theta_1}(\overrightarrow{y}) d\overrightarrow{y} - \int_{\Omega_1' \cap \Omega_1} f_{\theta_1}(\overrightarrow{y}) d\overrightarrow{y} - \int_{\Omega_1' \cap \Omega_0} f_{\theta_1}(\overrightarrow{y}) d\overrightarrow{y} =$$

$$= \int_{\Omega_1 \cap \Omega_0'} f_{\theta_1}(\overrightarrow{y}) d\overrightarrow{y} - \int_{\Omega_1' \cap \Omega_0} f_{\theta_1}(\overrightarrow{y}) d\overrightarrow{y}$$

Using the definition of Ω_1 we can now know that this tells us that

$$\pi - \pi' \geq \int_{\Omega_1 \cap \Omega_0} C f_{\theta_0}(\overrightarrow{y}) d\overrightarrow{y} - \int_{\Omega_1' \cap \Omega_0} C f_{\theta_0}(\overrightarrow{y}) d\overrightarrow{y} = C \cdot \left(\int_{\Omega_1} f_{\theta_0}(\overrightarrow{y}) d\overrightarrow{y} - \int_{\Omega_1'} f_{\theta_0}(\overrightarrow{y}) d\overrightarrow{y} \right) =$$

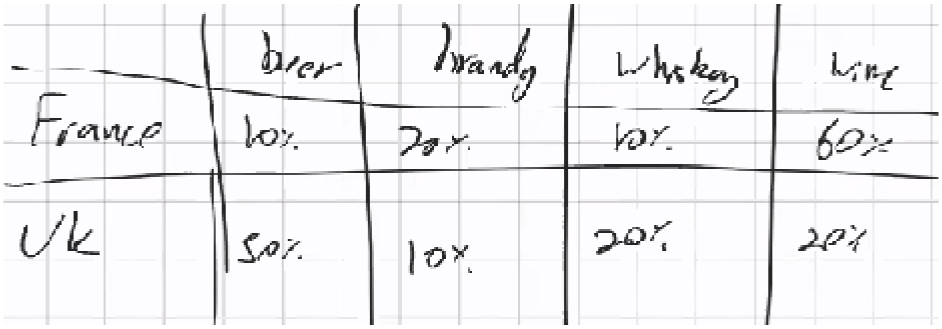
$$= C \cdot (\alpha - \alpha') \geq 0$$

Meaning we have that the power of LRT test is always greater or equal to the power of any other test, and we are done. ■

Example 12.1. *You are in a bar europe and you meet another person who you know is either french or british. You know that certain nationalities drink certain drinks in different probabilities. You can look at what drink they have, and you want to create a test that will allow you to know the nationality of this person. Assuming you take significance $\alpha = 0.25$ and your hypotheses are*

$$H_0 : \text{ "They are french", and } H_0 : \text{ "They are british"}$$

The table of drink per nation is the following



	beer	brandy	Whiskey	Wine
France	10%	20%	10%	60%
UK	50%	10%	20%	20%

Notice that we see what type of drinks the person drinks and we want to decide what nationality they are. Meaning when we see a certain drink we should already have a deterministic guess for nationality. In order to have the significance that was stated above the possible rejection regions are

$$\{Beer\}, \{Brandy\}, \{Whiskey\}\{Brandy,\ Beer\}$$

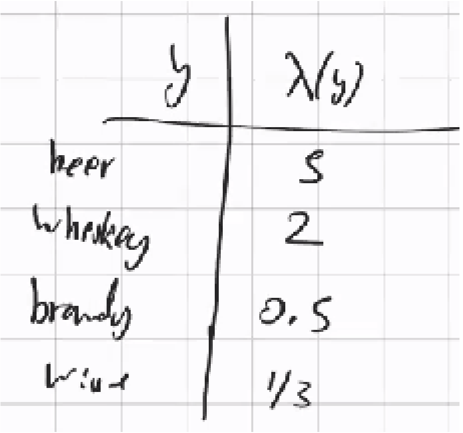
The powers are

$$\pi_{\{Brandy\}} = 0.1, \text{ and } \pi_{\{Brandy,\ Beer\}} = 0.7$$

This gives us a test for nationality based on what the persons drinks. Now we can consider the LRT. This is

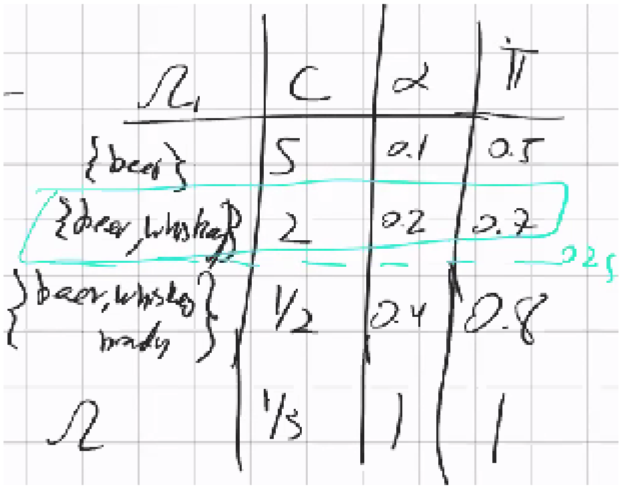
$$\lambda(\overrightarrow{y}) = \frac{\mathbb{P}_{UK}(\overrightarrow{y})}{\mathbb{P}_{France}(\overrightarrow{y})}$$

This gives us the following table of values of λ



y	$\lambda(y)$
beer	5
whiskey	2
brandy	0.5
wine	1/3

We can now examine how the powers change in the following table



Ω_1	C	α	π
{beer}	5	0.1	0.5
{beer, whiskey}	2	0.2	0.7
{beer, whiskey, brandy}	1/2	0.4	0.8
Ω	1/3	1	1

0.25

As we can see we have a discrete distribution and we can't find the specific significance we want. In continuous distributions though we can find the significance we want and so given a significance value α we need to rescue our C to get the test that we want for LRT. Now we can consider "composite" hypotheses (where there can be various different values of θ we can get).

12.2 Composite Hypotheses

Until now we have dealt with singletons for hypotheses, but in the real world usually the hypotheses are more than just a singleton and so we will need to give new definitions.

Definition 12.1. Given composite hypotheses we define the power function

$$\pi(\theta) := \mathbb{P}_\theta(\text{"reject } H_0\text{"}) = \mathbb{P}_\theta(\overrightarrow{Y} \in \Omega_1)$$

Definition 12.2. The significance of a test on composite hypotheses is

$$\alpha := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\overrightarrow{Y} \in \Omega_1)$$

Definition 12.3. If $T(\overrightarrow{Y}) = t_{\text{obs}}$ then the p-value as

$$p_{\text{value}} := \sup_{\theta \in \Theta_0} \mathbb{P}_{\theta_0}(T(\overrightarrow{Y}) \geq t_{\text{obs}})$$

13 Thirteenth Lecture

13.1 Composite Hypotheses

Assume that

$$\Theta = \Theta_0 \cup \Theta_1 \cup Theta_{\text{other}}$$

Is the set of coefficients where our hypotheses are -

$$H_0 : theta \in \Theta_0,\ H_1 : \theta \in \Theta_1$$

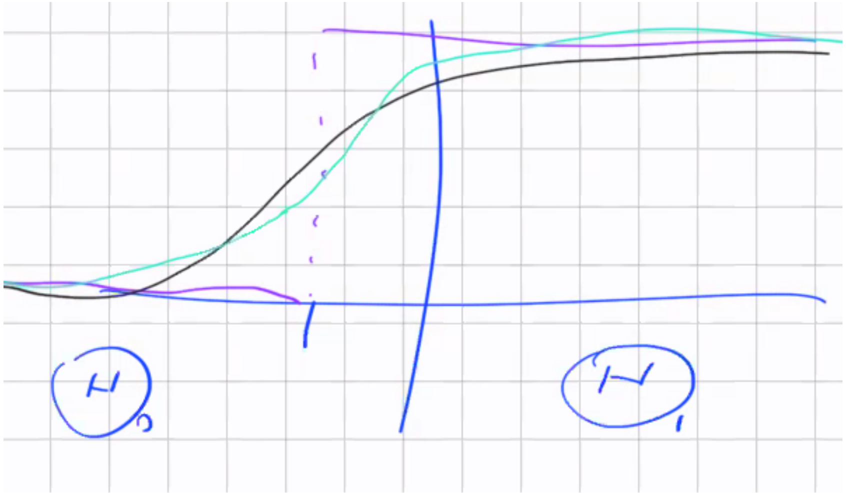
Now for composite hypotheses we have that

$$\alpha := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\overrightarrow{y} \in \Omega_1)$$

And the power function

$$\pi(\theta) : \mathbb{P}_\theta(\overrightarrow{y} \in \Omega_1)$$

What we want is the following graph of π:



Definition 13.1. A uniformly most powerful test with significance α is a test that satisfies

- $$\sup_{\theta \in \Theta_0} \pi(\theta) = \alpha$$
- Assume that $\pi_1(\theta)$ is a power function for a different test with significance α, we have that for all θ
$$\pi(\theta) \geq \pi_1(\theta)$$

Definition 13.2. For a family of distributions F_θ where Θ is one-dimensional. This family has monotone likelihood ratios if there is a statistic $T(\overrightarrow{Y})$ such that for all $\theta_1 < \theta_2$ we have that

$$\frac{f_{\theta_0}(\overrightarrow{y})}{f_{\theta_1}(\overrightarrow{y})}$$

is a monotone non-decreasing function of $T(\overrightarrow{Y})$.

Claim 13.1. Assume $\overrightarrow{Y} \sim f_\theta(\overrightarrow{y})$ where $f_\theta \in F_\theta$ from a family of monotone likelihood ratios with statistic $T(\overrightarrow{Y})$ Then there is a uniformly most powerful test with significance α for the hypothesis

$$H_1 : \theta > \theta_0$$

where the test is

$$T(\overrightarrow{Y}) \geq C$$

And

$$\alpha = \mathbb{P}_{\theta_0}(T(\overrightarrow{Y}) \geq C)$$

Proof. Assume first that $\theta_1 > \theta_0$. We will create a test with maximal power for the hypothesis

$$H_0 : \theta = \theta_0,\ H_1 : \theta = \theta_1$$

Using Neyman-Pearson the maximal power test is the LRT. Meaning

$$\lambda(\overrightarrow{y}) = \frac{f_{\theta_1}(\overrightarrow{y})}{f_{\theta_0}(\overrightarrow{y})} > C$$

Where C is determined such that

$$\mathbb{P}_{\theta_0}(\lambda(\overrightarrow{y} > C)) = \alpha$$

This is independent of θ_1 and so this is the maximal power test for $\theta_1 \in \Theta_1$. Now assume that $\Theta_0 := (-\infty, \theta_0]$. It suffices to show that

$$\text{For every } \theta \in \Theta_0 wehavethat \mathbb{P}_\theta(\overrightarrow{y} \in \Omega_1) \leq \alpha$$

Take $\theta_2 < \theta_0$ and build the following test that distinguishes between the following hypotheses:

$$H_0 : \theta = \theta_2,\ H_1 : \theta = \theta_0$$

From Neyman-Pearson we know that the LRT is the maximal power test

$$\lambda(\overrightarrow{y}) \geq C'$$

where

$$\mathbb{P}_{\theta_2}(T(\overrightarrow{y}) \geq C) = \alpha'$$

and

$$\pi = \mathbb{P}_{\theta_0}(T(\overrightarrow{Y}) \geq C) = \alpha$$

We want to show that $\alpha \geq \alpha'$. We will build another test - flip a biased coin with probability α' for heads and declare H_1, otherwise maintain null hypothesis . It can be easily seen that the power and significance α' and so $\alpha' \leq \alpha$ and we are done. (Notice that this means appending a sampling of a bernoulli variable to the sampling) ■

13.2 Generalized Likelihood Ratio Test

Given composite hypotheses we can redefine the likelihood ratio

$$\lambda = \frac{\sup_{\theta \in \Theta} f_\theta(\overrightarrow{y})}{\sup_{\theta \in \Theta_0} f_\theta(\overrightarrow{y})}$$

This is called the general likelihood ratio.

Definition 13.3. A GLRT is a test for composite hypothesis where the test is

$$\lambda(\overrightarrow{y}) \geq C$$

Notice that both the denominator and the numerator are MLE's for different coefficient spaces. Usually in order to find a good way to express this is using some statistic $T(\overrightarrow{Y})$ to help us inverse the λ function.

Example 13.1. *Sample $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$, with unknown σ^2. Now consider 2 hypotheses*

$$H_0 : \mu = \mu_0,\ H_1 : \mu \neq \mu_0$$

. MLE we have already seen for the normal distribution. This is

$$\left(\bar{Y}, \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{Y}\right)^2\right)$$

MLE estimate for H_0 is

$$\hat{\mu_0} = \mu_0, \hat{\sigma_0^2} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mu\right)^2$$

The likelihood in general is

$$L(\hat{\mu}, \hat{\sigma^2}; \overrightarrow{y}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(\frac{y_i - \bar{y}}{\hat{\sigma}})^2}$$

Looking at the ratio between this and

$$L(\hat{\mu_0}, \hat{\sigma_0^2}; y)$$

We have

$$\left(1 + \frac{1}{n-1}\left(\frac{\bar{y} - \mu_0}{s/\sqrt{n}}\right)^2\right)^{\frac{1}{2}n}$$

Where

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \mu_0)^2$$

Now take statistic

$$T(\overrightarrow{Y}) = \frac{\bar{y} - \mu_0}{S/\sqrt{n}}$$

Therefore the test will look like

$$|T(\overrightarrow{Y})| \geq C \textit{ where } \mathbb{P}_{\mu_0}(|T(\overrightarrow{Y}) \geq C|) = \alpha$$

For composite hypotheses we enter a supremum in the following sense

$$\text{p-value} := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left(T(\overrightarrow{Y}) > T(\overrightarrow{y}) \right)$$

13.3 Connection Between hypothesis testing to the significance interval

A significance interval with significance $1-\alpha$ for θ is the interval $(L(\overrightarrow{y}), U(\overrightarrow{y}))$ Such that

$$\mathbb{P}_\theta(L \leq \theta \leq U) \geq 1-\alpha$$

If we define $H_0 : \theta = \theta_0$. We will define the test where we find the significance interval. If $\theta_0 \in (L, U)$ then H_0 otherwise H_1. This is a test with significance α. (This is a pretty simple observation).

And we are done