MATH 357 Honors Statistics

Yuyan Chen

January 2022

Lecture 1b

1 Chapter 1: Random Sampling

1.1 Basic Concepts

Definition 1.1. *The random variables (vectors) $X_1, \cdots, X_n$ are called a* ***random sample*** *if they are iid with some common distribution P. P is called the* ***population distribution*** *and n is called the* ***sample size***. ***Data*** *are the observations (or realizations) of $X_1, \cdots, X_n$, i.e.*

$$x_1, \cdots, x_n.$$

Note: We regard P as **unknown**; it is a proxy for our lack of knowledge of some phenomenon. Our goal is to infer (learn) P or some of its properties from the basis of the observed data $x_1, \cdots, x_n$.

Example 1.2.

Lecture 2a

Recall the definition of a random sample. This sampling model is also called sampling from an **infinite** population. Independence implies the distribution of X_2 is unaffected by having sampled $X_1 = x_1$.

Remark 1.3 (Finite population (N) with P(sampled) = 1/N)**.**

1. *Sample with replacement*
2. *Sample without replacement: $X_1, \cdots, X_n$ are identically distributed but NOT independent. However when N is much langer than n, the independence assumption may be a good enough approximation.*

1.2 Descriptive Statistics

Definition 1.4 (statistic)**.** *Let $X_1, \cdots, X_n$ be a random sample from P on $\mathbb{R}^d$. Let $T : \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}^h$ be a measurable mapping that does NOT depend on any unknown parameters. The random vector $T(X_1, \cdots, X_n)$ is called a* ***statistic****.*

Note that with Borel measure, all **continuous** functions are **measurable**.

Example 1.5.

$$(\frac{1}{n}\sum_{i=1}^{n} 1(X_i = 0) - p_0)^2$$

is not a statistic since p_0 is unknown.

Rule of thumb: You must be able to evaluate a statistic. The observed value must be a scalar, not a term or formula.

Definition 1.6. *Let $X_1, \cdots, X_n$ be a random sample from P on $\mathbb{R}$. Then*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is called the ***sample mean*** *(a measure of central tendency). Furthermore,*

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

is called the ***sample variance*** *(a measure of variability), and S is called the* ***sample standard deviation****. The observed values are denoted $\bar{x}, s^2, s$.*

Theorem 1.7. *For arbitrary $x_1, \cdots, x_n \in \mathbb{R}$,*

(a)

$$\min_{a\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

(b)

$$(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

Proof.

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - a)^2$$

□

Lemma 1.8. *Let $X_1, \cdots, X_n$ be a random sample from P on $\mathbb{R}$, $X \sim P$, g measurable so that $E\, g(X)$ and $var\, g(X)$ exist. Then*

$$E\left(\sum_{i=1}^{n} g(X_i)\right) = n \cdot E(g(X))$$
$$var\left(\sum_{i=1}^{n} g(X_i)\right) = n \cdot var(g(X)))$$

Note that

$$E(g(X)) = \int g(x)f(x)dx$$

Theorem 1.9. *Let $X_1, \cdots, X_n$ be a random sample from P on $\mathbb{R}$, $X \sim P$, $EX = \mu$ and $\sigma^2 = var\, X$ are finite. Then,* *(a)* $E\bar{X} = \mu$

(b) *var* $(\bar{X}) = \frac{\sigma^2}{n}$

(c) $E(S^2) = \sigma^2$.

Note: Theorem 1.9 holds for all P *such that* $EX = \mu$ *and* $\sigma^2 = var\ X$ *are finite.*

Example 1.10.

Definition 1.11 (order statistics)**.** *Let $X_1, \cdots, X_n$ be a random sample from P on $\mathbb{R}$. Placed in ascending order,*

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)},$$

the ordered random variables are called the ***order statistics****. $X_{(r)}$ is called the r^{th} order statistic.*

- $X_{(1)} \cdots$ *sample* ***minimum***
- $X_{(n)} \cdots$ *sample* ***maximum***
- $R = X_{(n)} - X_{(1)} \cdots$ *sample* ***range***
- $X_{med} \cdots$ *sample* ***median*** *(a measure of central tendency)*

$$X_{med} = \begin{cases} X_{\frac{n+1}{2}}, \text{ if } n \text{ is odd} \\ \\ \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}, \text{ if } n \text{ is even} \end{cases}$$

- *sample* $(100 \cdot p)^{th}$ ***percentile****, where $p \in (\frac{1}{2n}, 1 - \frac{1}{2n})$ is:*
  - $X_{(\{np\})}$ *if* $p \in (\frac{1}{2n}, \frac{1}{2})$
  - X_{med} *if* $p = \frac{1}{2}$
  - $X_{(\{n+1-n(1-p)\})}$ *if* $p \in (\frac{1}{2}, 1 - \frac{1}{2n})$

  where $b \in [0, \infty)$, $\{b\}$ is the integer so that

  $$j - \frac{1}{2} \leq b < j + \frac{1}{2}.$$

  The definition of the $(100 \cdot p)^{th}$ percentile is rigged so that if the $(100 \cdot p)^{th}$ percentile is $X_{(i)}$, the i^{th} smallest observation, the $(100 \cdot (1-p))^{th}$ percentile is the i^{th} largest observation, $X_{(n+1-i)}$.

- *the 25^{th} percentiled is called the **first quartile** (Q1)*
- *the 75^{th} percentiled is called the **third quartile** (Q3)*
- *their differntce $IQR = Q_3 - Q_1$ (a measure of variability) is called **interqurtile range**.*

Lemma 1.12 (Mean absolute error)**.** *For any $x_1, \cdots, x_n \in \mathbb{R}$, let X_{med} be the observed value of the sample median. Then for any $a \in \mathbb{R}$,*

$$\frac{1}{n}\sum_{i=1}^{n} |x_i - a| \geq \frac{1}{n}\sum_{i=1}^{n} |x_i - x_{med}|.$$

Example 1.13.

Graphical data visualization

(a) Boxplot

(b) Histogram (for continuous data)

Partition the range $[x_{(i),x_{(n)}}]$ into k (chosen) bins.

h_j is so that

$$\begin{aligned} h_j \cdot (b_{j+1} - b_j) &= \frac{1}{n}\sum_{i=1}^{n} 1(x_i \in [b_j, b_{j+1}]) \\ &\approx P(X \in [b_j, b_{j+1}]) \end{aligned}$$

The idea is that the histogram approximates the pdf of P.

(c) Bar chart/ bar plot (for discrete data) We observed k distinct value.

$$h_j = \frac{1}{n}\sum_{i=1}^{n} 1(x_i = b_j) \approx P(X = b_j)$$

Bar chart approximates the pmf of P.

1.3 Sampling distribution

Definition 1.14 (sampling distribution)**.** *Consider a statistic* $T(X_1, \cdots, X_n)$*. Its distribution is called the sampling distribution of* $T(X_1, \cdots, X_n)$*.*

Theorem 1.15. *Consider a random sample from* P *on* $\mathbb{R}$*,* $X \sim P$ *and assume that* X *has a MGF (moment generating function)* M_X *on the interval* I*. Then* $\bar{X}$ *has MGF*

$$M_{\bar{X}}(t) = (M_X(t/n))^n$$

Example 1.16.

- $X \sim \mathcal{N}(\mu, \sigma^2)$, $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$
- $X \sim Bin(m, p)$, $n \cdot \bar{X} \sim Bin(m \cdot n, p)$
- $X \sim Gamma(\alpha, \beta)$, $\bar{X} \sim Gamma(\alpha \cdot n, \beta/n)$.

Observation: the sampling distribution of $T(X_1, \cdots, X_n)$ **depends** on the population distribution P.

Theorem 1.17. *Let* $X_1, \cdots, X_n$ *be a random sample from* P *on* $\mathbb{R}$*. Then from any* $x \in \mathbb{R}$*,* $r \in \{1, \cdots, n\}$*,*

$$P(X_{(r)} \leq x) = F_{X_{(r)}}(x) = \sum_{k=r}^{n} \binom{n}{k} \{F(x)\}^k \{1 - F(x)\}^{n-k}$$

Proof. Fix $x \in \mathbb{R}$, $r \in \{1, \cdots, n\}$. Let

$$\begin{aligned} Y &= \#i : X_i \leq x \\ &= \sum_{i=1}^{n} 1(X_i \leq x), \text{ iid Bernoulli}(F(x)), \text{since } P(X_i \leq x) = F(X) \end{aligned}$$

Hence, $Y \sim Bin(n, F(x))$.

$$\begin{aligned} P(X_{(r)} \leq x) &= P(Y \geq r) \\ &= \sum_{k=r}^{n} \binom{n}{k} (F(x))^k (1 - F(x))^{n-k} \end{aligned}$$

$\square$

Note: if P has a pdf f, then $X_{(r)}$ has a pdf

$$f_{(X_{(r)})}(x) = \frac{n!}{(r-1)!(n-r)!} \{F(x)\}^{r-1} f(x) \{1 - F(x)\}^{n-r}.$$

Example 1.18. *Suppose $U_1, \cdots, U_n$ from $U(0,1)$. Then $U_{(r)}$ has a pdf*

$$f_{U_{(r)}}(u) = \frac{n!}{(r-1)!(n-r)!} u^{r-1} (1-u)^{n-r}.$$

Note that $\Gamma(n) = (n-1)!$ Hence, $U_{(r)} \sim Beta(r, n-r+1)$. In particular,

$$E(U_{(r)}) = \frac{r}{n+1}.$$

Note: for $\mathcal{U}(a,b)$, $f(x) = 1/(b-a)$ for $x \in [a,b]$, 0 otherwise.

1.4 Sampling from the Normal Population

Throughout this section, $X \sim \mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 are unknown.

Theorem 1.19. *Let $X_1, \cdots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Let $\bar{X}$ and S^2 be the sample mean and variance. Then,*

(a)

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

(b) $\bar{X}$ and S^2 are independent.

Proof. (b) Let X_i^* be the standardized variable such that

$$X_i^* = \frac{X_i - \mu}{\sigma}.$$

Then, $X_i^* \sim \mathcal{N}(0, 1)$. We have

$$\bar{X}^* = \frac{\bar{X} - \mu}{\sigma}$$
$$(S^*)^2 = \frac{S^2}{\sigma^2}.$$

Both are one-to-one function to $\bar{X}$ and S^2, respectively. Hence, WLOG, we can assume $\mu = 0$ and $\sigma^2 = 1$ and if $\bar{X}^* \perp (S^*)^2$, $\bar{X} \perp S^2$. Note that

$$S^2 = \frac{1}{n-1}\left(\underbrace{(-\sum_{i=2}^{n}(X_i - \bar{X})}_{=X_1 - \bar{X}})^2 + \sum_{i=2}^{n}(X_i - \bar{X})^2\right).$$

Lemma 1.20. *$X_2, \cdots, X_n$ iid $\mathcal{N}(0, 1)$. Then,*

$$\bar{X} \perp (X_2 - \bar{X}, \cdots, X_n - \bar{X}).$$

Proof. Define $T : \mathbb{R}^n \to \mathbb{R}^n$ as

$$(x_1, \cdots, x_n) \to (\bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x}).$$

Then, $T^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ is

$$(y_n, \cdots, y_n) \to (\underbrace{\quad y_1 - \sum_{i=2}^{n} y_i \quad}_{=n \cdot y_1 - \sum_{i=2}^{n}(y_i + y_1)}, y_2 + y_1, \cdots, y_n + y_1).$$

Jacobi matrix $|J| = n$.

$$\begin{aligned} f_{(Y_1,\cdots,Y_n)}(y_1, \cdots, y_n) &= f_{(X_1,\cdots,X_n)}(T^{-1}(y_1, \cdots, y_n)) \cdot |J| \\ &= ((\frac{1}{\sqrt{2\pi}})^n \exp(-\frac{1}{2}((y_1 - \sum_{i=2}^{n} y_i)^2 + \sum_{i=2}^{n}(y_i + y_1)^2))) \cdot n \\ &= \sqrt{n}(\frac{1}{\sqrt{2\pi}}) \exp(-\frac{1}{2}(ny_1^2)) \\ &\cdot \sqrt{n}(\frac{1}{\sqrt{2\pi}})^{n-1} \exp(-\frac{1}{2}((\sum_{i=2}^{n} y_i)^2 + \sum_{i=2}^{n} y_i^2)) \\ &= f_1(y_1) \cdot f_2(y_2, \cdots, y_n) \end{aligned}$$

Theorem 12.7 (from Jacod & Protter) Let $X = (X_1, \cdots, X_n)$ have joint density f. Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable and injective, with non-vanishing Jacobian. Then $Y = g(X)$ has density

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))|\det J_{g^{-1}}(y)|, & \text{if } y \text{ is in the range of } g \\ 0, & \text{otherwise.} \end{cases}$$

□

Since S^2 is a function of $(X_2 - \bar{X}, \cdots, X_n - \bar{X})$ which we now know is independent of $\bar{X}$. □

Definition 1.21 (Chi-squared distribution)**.** *The χ^2_ν distribution has a pdf given, for all $x > 0$,*

$$f(x;\nu) = \frac{1}{2^{\nu/2}\Gamma(\frac{\nu}{2})} \cdot x^{\nu/2-1} \cdot e^{-x/2}$$

and 0 otherwise. The χ^2_ν distribution is in fact the Gamma$(\frac{\nu}{2}, 2)$. The MGF of χ^2_ν is given, for all $t < \frac{1}{2}$, by $M_{\chi^2_\nu} = (1-2t)^{-\nu/2}$.

Lemma 1.22.

(a) When $X \sim \chi^2_\nu$, then $EX = \nu$ and var $X = 2\nu$

(b) $X_1 \sim \chi^2_{\nu_1}$, $X_2 \sim \chi^2_{\nu_2}$, and $X_2 \perp X_1$, then $X_1 + X_2 \sim \chi^2_{\nu_1+\nu_2}$

(c) $X \sim \mathcal{N}(0,1)$ then $X^2 \sim \chi^2_1$.

Theorem 1.23. *Supposet that $X_1, \cdots, X_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$. Then,*

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

Motivation for t distribution: Consider

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

where σ is unknown. Instead:

$$\sqrt{n}\frac{\bar{X}-\mu}{S} \equiv T.$$

Note that T is a statistic.

Definition 1.24 (Student t distribution). *The Student t distribution with ν degrees of freedom, t_ν, has pdf*

$$f(x;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\cdot\Gamma(\frac{\nu}{2})}(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}},\ x \in \mathbb{R}.$$

Lemma 1.25. *Let $X \sim t_\nu$. The the following holds:*

(a) $EX = 0$ if $\nu > 1$. If $\nu \leq 1$, EX does not exist. Note: t_1 is Cauchy(1).

(b) $varX = \frac{\nu}{\nu-2}$ if $\nu > 2$. If $\nu \leq 2$, then $varX$ does not exist.

(c)

$$X \stackrel{d}{=} \frac{Z}{\sqrt{V/\nu}}$$

where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2_\nu$, and $Z \perp V$.

Theorem 1.26. *Suppose that $X_1, \cdots, X_n$ is a random sample from $\mathcal{N}(\mu,\sigma^2)$. Then,*

$$T = \sqrt{n}\cdot\frac{\bar{X}-\mu}{S} \sim t_{n-1}$$

Proof. Lemma 1.25 (c). □

Definition 1.27. *The Fisher-Snedecor F_{ν_1,ν_2} with ν_1 and ν_2 dof is the distribution of*

$$\frac{V_1/\nu_1}{V_2/\nu_2}$$

where $V_1 \sim \chi^2_{\nu_1}$, $V_2 \sim \chi^2_{\nu_2}$, $V_1 \perp V_2$.

Theorem 1.28. *Let $X_1, \cdots, X_n$ be a random sample from $\mathcal{N}(\mu_1, \sigma_1^2)$. Let $Y_1, \cdots, Y_m$ be a random sample from $\mathcal{N}(\mu_2, \sigma_2^2)$. Suppose that $(X_1, \cdots, X_n)$ and $(Y_1, \cdots, Y_n)$ are independent; let S_X^2 and S_Y^2 be their respective sample variances, then*

$$\underbrace{\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2}}_{\text{not a statistic since } \sigma_1^2 \text{ and } \sigma_2^2 \text{ unknown}} \sim F_{n-1,m-1}.$$

<u>Remark:</u> Theorem 1.28 will serve as later to derive the so-called F test. Imagine we want to assess whether $\sigma_1^2 = \sigma_2^2$.

$$\underbrace{\frac{S_X^2}{S_Y^2}}_{\text{is a statistic}} \neq 1 \sim F_{n-1,m-1}.$$

2 Chapter 2: Theory of point estimation

2.1 Parametric model

Throughout this chapter, we will assume that $X_1, \cdots, X_n$ is a random sample from P and that

$$P \in \mathcal{P} = \{P_\theta, \theta \in \Theta\}.$$

- $\mathcal{P}$ is called a **parametric model** for P.
- θ is called a **parameter**.
- Θ is called a **parameter space** and we assume that $\Theta \in \mathbb{R}^k$.

We will denote the CDF of P_θ by F_θ and its pdf/pmf by $f(x;\theta)$, $x \in \mathbb{R}$.

Example 2.1. *For Newcomb's measurements, we may assume*

$$\mathcal{P} = \{\underbrace{\mathcal{N}(\mu, \sigma^2)}_{P_\theta}, \underbrace{(\mu, \sigma^2)}_{\theta} \in \underbrace{\mathbb{R} \times (0, \infty)}_{\Theta}\}$$

Note: A parametric model for P is an **assumption**. It is always an **approximation** to the reality which may or may NOT be true. Our goal is to estimate the unknown parameter θ from the observed data $x_1, \cdots, x_n$.

Definition 2.2. *A* ***point estimator*** *is any statistic* $W(X_1, \cdots, X_n)$ *which has been constructed with the aim to estimate* θ*. The observed value of* W*, i.e.* $W(x_1, \cdots, x_n)$ *is called the* ***estimate*** *of* θ*.*

Note: we do NOT require that the range of W is Θ.
Notation: estimators are often denoted $\hat{\theta}$, $\hat{\theta}(X_1, \cdots, X_n)$, $\tilde{\theta}$, and θ_n.

2.2 Methods of finding estimators

Recall: an estimator is a statistic $W(X_1, \cdots, X_n)$.

2.2.1 Method of moments

sample moment:

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j.$$

From Theorem 1.9, we know that if $EX^j < \infty$, $E(m_j) = EX^j$. If $E(X^j)^2 < \infty$, then from the weak law of large numbers,

$$m_j \xrightarrow{P} EX^j \text{ as } n \to \infty$$

Now suppose $\theta = (\theta_1, \cdots, \theta_k)$. The method of moments proceeds as follows:

1. Calculate k moments of P_θ (population moments), i.e:

$$EX^j = \mu_j(\theta),\ j = 1, \cdots, k.$$

2. Calculate the j^{th} sample moment

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j,\ j = 1, \cdots, k.$$

3. Equate

$$m_j = \mu_j(\theta),\ j = 1, \cdots, k.$$

If there is a unique solution, it is called a **method of moments estimator** of θ.

- “easy”

- usually consistent since

$$Y \xrightarrow{P} y \implies f(Y_n) \xrightarrow{P} f(Y)$$

- usually biased (e.g. Jensen inequality)

<u>Remark</u> You may need to choose moments other than the first k, depending on the distribution P_θ.

Example 2.3. *Suppose $X_1, \cdots, X_n$ is a random sample from the Normal distribution, i.e:*

$$P \in \{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)\}.$$

The method-of-moment estimator of (μ, σ^2) is

$$(\bar{X}, \underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}_{\frac{n-1}{n}S^2}).$$

Example 2.4. *Consider a random sample $X_1, \cdots, X_n$ from $Bin(N, p)$, i.e.*

$$P \in \{Bin(N, p), p \in (0, 1)\}$$

where N is known. The method of moment generator of p is

$$\hat{p} = \frac{1}{N}\bar{X}.$$

If N is unknown, the method-of-moment estimator of (p, N) is

$$\left(\frac{\bar{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}{\bar{X}}, \frac{(\bar{X})^2}{\bar{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\right).$$

<u>Note:</u> the method of moment estimators above may well be negative. The estimator of N may not be an integer.

Example 2.5. *Consider a random sample from* $U(-\theta, \theta)$*,*

$$P \in \{U(-\theta, \theta), \theta \in (0, \infty)\}.$$

We have

$$EX = \frac{-\theta + \theta}{2} = 0,$$

which is not useful. Use the second moment, we obtain

$$\hat{\theta} = \sqrt{\frac{1}{2n} \sum_{i=1}^{n} X_i^2}.$$

Consider $x_0 = 0$*,* $x_1 = 1 \sim U(\theta, \theta)$*. We find* θ *to be*

$$\hat{\theta} = \sqrt{\frac{1}{4}(0 + 1)} = \frac{1}{2}.$$

However, $0, 1 \notin (-\frac{1}{2}, \frac{1}{2})$*.*

2.2.2 Method of Maximum Likelihood

Assume $X_1, \cdots, X_n$ is a random sample from

$$P \in \{P_\theta, \theta \in \Theta\}.$$

Assume also that for each $\theta \in \Theta$, P_θ has a PMF/PDF.

Definition 2.6. *Given the observed data $x_1, \cdots, x_n$, the function of θ defined by*

$$L(\theta) = L(\theta; x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

*is called the **likelihood function**.*

Note that the likelihood function is a function of θ for a fixed set $x_1, \cdots, x_n$.

Example 2.7.

<u>Interpretation of the likelihood function</u>

- If P_θ is discrete, then the value of L at θ_0 is

$$\begin{aligned} L(\theta_0) &= P_{\theta_0}(X_1 = x_1, \cdots, X_n = x_n) \\ &= L(\theta_0; x_1, \cdots, x_n) \end{aligned}$$

  $L(\theta_0)$ is the probability of observing the data we observed if the parameter $\theta = \theta_0$. For example, in Example 2.7,

$$L(1) = 3.8 \times 10^{-5}$$

  is the probablity (or "likelihood") of observing 1,2,2,5 when $\lambda = 1$.

- When P_θ is continuous, this interpretation is still used, but in an approximation sense. Because $P(X_1 = x_1, \cdots, X_n = x_n) = 0$, we need to consider

$$\begin{aligned}
&P(X_1 \in (x_1 - \varepsilon, x_1 + \varepsilon), \cdots, X_n \in (x_n - \varepsilon, x_n + \varepsilon)) \\
=&\int_{x_1-\varepsilon}^{x_1+\varepsilon} \cdots \int_{x_n-\varepsilon}^{x_n+e} \prod_{i=1}^{n} f(t_i;\theta) dt_n \cdots dt_1 \\
\approx&\prod_{i=1}^{n} f(t_i;\theta) \cdot (2\varepsilon)^n \\
=&L(\theta; x_1, \cdots, x_n) \cdot \qquad \underbrace{(2\varepsilon)^n}_{\text{does not contain } \theta}
\end{aligned}$$

provided that $\varepsilon > 0$ is very small. So,

$$L(\theta; x_1, \cdots, x_n) \propto P(X_1 \in (x_1 - \varepsilon, x_1 + \varepsilon), \cdots, X_n \in (x_n - \varepsilon, x_n + \varepsilon))$$

Whether P_θ is continuous or discrete, we can say that if

$$L(\theta_1; x_1, \cdots, x_n) \geq L(\theta; x_1, \cdots_2, x_n),$$

it is more "likely" to have observed $x_1, \cdots, x_n$ when $\theta = \theta_1$ than $\theta = \theta_2$.

Definition 2.8. *For an observed sample $x_1, \cdots, x_n$, the **maximum likelihood (ML) estimate** of θ, denoted $\hat{\theta}(x_1, \cdots, x_n)$ is a value such that*

$$L(\hat{\theta}(\underset{\sim}{x}); x_1, \cdots, x_n) = \sup_{\theta \in \Theta} L(\theta; x_1, \cdots, x_n)$$

provided it exists. If the ML estimate exists for almost all samples $x_1, \cdots, x_n$ and if the mapping $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^h$

$$(x_1, \cdots, x_n) \to \hat{\theta}(x_1, \cdots, x_n)$$

is measurable, $\hat{\theta}(X_1, \cdots, X_n)$ is called the ML estimator of θ.

"Almost all samples" means that $\hat{\theta}(\underset{\sim}{x})$ exists for all $\underset{\sim}{x} \in A$ when

$$P_\theta((X_1, \cdots, X_n) \in A) = 1$$

for all $\theta \in \Theta$.

In Definition 2.8, note that the ML estimate is the value $\hat{\theta}(\underset{\sim}{x})$ in Θ at which the sup is attained.

The **log-likelihood function** is defined as

$$l(\theta; x) = \log L(\theta; \underset{\sim}{x}) = \sum_{i=1}^{n} \log f(x_i; \theta).$$

Typically, l is smooth and we can look for its maximum by calculating

$$\frac{\partial l}{\partial \theta_j}(\theta; x_1, \cdots, x_n) = 0, \ j = 1, \cdots, k$$

and inspect the solutions.

Example 2.9. *Consider a random sample from a Binomial population with KNOWN size N:*

$$P \in \{Bin(N, P), p \in [0, 1]\}.$$

The likelihood function is

$$L(p; x_1, \cdots, x_n) = \prod_{i=1}^{n} \binom{N}{x_i} p^{x_i} (1-p)^{N-x_i}.$$

The ML estimator is thus $\hat{p} = \frac{\bar{X}}{N}$ *(and the same as the method-of-moment estimator.)*

Careful: If we choose

$$\{Bin(N, p), p \in (0, 1)\}$$

then ML estimate does not exist when $\bar{x} = 0$ or $\bar{x} = N$. Since $P_p(\bar{X} = 0) \neq 0$, $P_p(\bar{X} = N) \neq 0$, the ML estimator does not exist in this case.

Example 2.10. *Consider a random sample from*

$$P \in \{\mathcal{N}(\mu, 1), \mu \in \mathbb{R}\}.$$

ML estimator of μ is $\hat{\mu} = \bar{X}$. Suppose now we know that $\mu \geq 0$. In this case, $\bar{x}$ is not the ML estimate when $\bar{x} < 0$. Note that

$$\frac{\partial l}{\partial \mu} = n \cdot (\bar{x} - \mu) < 0$$

if $\bar{x} < \mu$. Hence, l is decreasing on $[0, \infty)$. Hence, l is maximized at $\tilde{\mu}(\underset{\sim}{x}) = 0$. In this (constrained) estimation problem, the MLE is

$$\tilde{\mu} = \max(\bar{X}, 0).$$

Example 2.11. *Take a random sample from $P \in \{U(0, \theta), \theta \in (0, \infty)\}$. To calculate the MLE,*

$$\begin{aligned} L(\theta; \underset{\sim}{x}) &= \prod_{i=1}^{n} \frac{1}{\theta} \cdot 1(x_i \in [0, \theta]) \\ &= (\frac{1}{\theta})^n \cdot 1(\min_{1 \leq i \leq n} x_i \geq 0) \cdot 1(\max_{1 \leq i \leq n} x_i \leq \theta). \end{aligned}$$

The MLE is

$$\tilde{\theta}(\underset{\sim}{x}) = \max_{1 \leq i \leq n} x_i.$$

<u>Note:</u> if the density function has a compact support, use the **indicator function** to denote the support.

Theorem 2.12 (Invariance Principle of the MLE)**.** *Consider a statistical model $\{P_\theta, \theta \in \Theta\}$ and suppose that $g : \Theta \to \mathbb{R}^m$ is an arbitrary measurable function. Set $\Gamma = g(\Theta)$ to be the range of g and suppose we wish to estimate $\gamma = g(\theta)$. Then if $\tilde{\theta}(\underset{\sim}{x})$ is the MLE of θ,*

$$\hat{\gamma} = g(\tilde{\theta}(\underset{\sim}{x}))$$

is the MLE of γ in the following sense: for

$$L^*(\gamma; \underset{\sim}{x}) = \sup_{\theta \in \Theta : g(\theta) = \gamma} L(\theta; \underset{\sim}{x})$$

then

$$L^*(\hat{\gamma}; \underset{\sim}{x}) = \sup_{\gamma \in \Gamma} (\gamma; \underset{\sim}{x})$$

Proof. WTS: $L^*(\hat{\gamma}; \underset{\sim}{x}) = \sup_{\gamma \in \Gamma} L^*(\gamma; \underset{\sim}{x})$.

$$\begin{aligned} L^*(\hat{\gamma}; \underset{\sim}{x}) &= \sup_{\theta \in \Theta : g(\theta) = \hat{\gamma}} L(\theta; \underset{\sim}{x}) \\ &= L(\hat{\theta}; \underset{\sim}{x}) \\ &= \sup_{\theta \in \Theta} L(\theta; \underset{\sim}{x}) \\ &= \sup_{\gamma \in \Gamma} \sup_{\theta \in \Theta : g(\theta) = \gamma} L(\theta; \underset{\sim}{x}) \\ &= \sup_{\gamma \in \Gamma} L^*(\gamma; \underset{\sim}{x}) \end{aligned}$$

□

Example 2.13.

- $\{Bin(N, p), p \in [0, 1]\}$, *$N$ is known.*
- $\{\textit{Exponential}(\lambda),\ \lambda > 0\}$. *The MLE of λ is $\bar{X}$.*

Example 2.14.

- $\{\mathcal{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 > 0\}$. *The MLE of (μ, σ^2) is $(\bar{X}, \frac{n-1}{n} S^2)$.*

In the Bayesian approach, our uncertainty (lack of knowledge) of θ is expressed by a probability density $\pi(\theta)$, called the **prior**. Once we have collected the data, we will update the prior by incorporating the information from the data. This leads to the so-called **posterior density**. Bayesian estimation tends to perform better for small sample size.

Assume for simplicity that θ is univariate and let π be the pmf/pdf of the prior distribution (i.e. a distribution on Θ of your choice). Suppose the density (pmf/pdf) of $(X_1, \cdots, X_n)$ given θ

$$\prod_{i=1}^{n} f(x_i; \theta).$$

The posterior density is the conditional density of θ given the observed data (i.e. conditionally on $X_1 = x_1$, $\cdots$, $X_n = x_n$). The posterior density is given by

$$\pi(\theta | x_1, \cdots, x_n) = \frac{\prod_{i=1}^{n} f(x_i; \theta)}{m(x_1, \cdots, x_n)} \cdot \pi(\theta)$$

where

$$m(x_1, \cdots, x_n) = \int_{\Theta} \prod_{i=1}^{n} f(x_i; \theta) \pi(\theta) d\theta$$

is the marginal density of $X_1, \cdots, X_n$ (unconditional). A Bayesian estimate of θ could be the mean of the posterior distribution with density (pmf/pdf) $\pi(\theta | x_1, \cdots, x_n)$.

Example 2.15. *$X_1, \cdots, X_n$ a Bernoulli random sample, $X_i \sim$ Bernoulli(p). $\Theta(0,1)$. The prior density is* ***chosen*** *to be Beta(α, β). The Bayesian estimate p_B as the expected value of the posterior:*

$$p_B = \frac{n\bar{x} + \alpha}{n + \alpha + \beta} = \frac{n}{n + \alpha + \beta} \cdot \underbrace{\bar{x}}_{\text{sample mean}} + \frac{\alpha + \beta}{n + \alpha + \beta} \cdot \underbrace{\frac{\alpha}{\alpha + \beta}}_{\text{expectation of the prior}}$$

Trick to avoid integration:

$$\pi(\theta|x_1,\cdots,x_n) = \underbrace{c(x_1,\cdots,x_n)}_{\text{normalizing constant}} \cdot \underbrace{\prod_{i=1}^{n} f(x_i;\theta)}_{\text{likelihood}} \cdot \underbrace{\pi(\theta)}_{\text{prior}}$$

$$\propto \text{likelihood} \times \text{prior}$$

Example 2.16. *$X_1,\cdots,X_n$ a random sample from Exponential(λ). The parameter space is $(0,\infty)$.*

- *Likelihood is $\lambda^n e^{-n\bar{x}\lambda}$*
- *Prior: Gamma(α, β)*

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{\lambda\beta},\ \lambda > 0$$

- *Posterior: Gamma($n+\alpha$, $n\bar{x}+\beta$)*
- *Bayesian estimator of λ:*

$$\hat{\lambda_B} = \frac{n+\alpha}{n\bar{x}+\beta} \underset{n\to\infty}{\to} \frac{1}{\bar{x}}$$

2.3 Method of evaluating estimators

Definition 2.17. *Consider a statistical model*

$$P = \{P_\theta, \theta \in \Theta\}$$

and $\gamma : \Theta \to \mathbb{R}^m$. *Let* $T(X_1, \cdots, X_n)$ *be an estimator of* $\gamma(\theta)$. *Then:*

(a) T *is called* ***unbiased*** *if* $\forall \theta \in \Theta$,

$$E_\theta T(X_1, \cdots, X_n) = \gamma(\theta).$$

The difference $E_\theta T(X_1, \cdots, X_n) - \gamma(\theta)$ *is called the* ***bias*** *of* T, *and denoted* $bias_\theta(T)$.

(b) If for all $\theta \in \Theta$,

$$\lim_{n\to\infty} E_\theta T(X_1, \cdots, X_n) = \gamma(\theta),$$

then T *is called* ***asymtotically unbiased***.

(c) (Weak consistency) T *is called* ***consistent*** *if for all* $\theta \in \Theta$

$$T(X_1, \cdots, X_n) \xrightarrow{P_\theta} \gamma(\theta)$$

as $n \to \infty$.

(d) The ***mean square error*** *of* T *is*

$$MSE_\theta = E_\theta\{T(X_1, \cdots, X_n) - \gamma(\theta)\}^2.$$

Note: the expectation, variance, etc. of T is taken w.r.t. P_θ and hence **depends** on θ. For all $\theta \in \Theta$:

$$\begin{aligned} MSE_\theta T &= E_\theta(T - \gamma(\theta))^2 \\ &= E_\theta(T - E_\theta T + E_\theta T - \gamma(\theta))^2 \\ &= E_\theta(T - E_\theta T)^2 + (E_\theta T - \gamma(\theta))^2 + 2(E_\theta T - \gamma(\theta)) \cdot E_\theta(T - E_\theta T) \\ &= var_\theta T + (bias_\theta T)^2 \end{aligned}$$

Example 2.18. *Consider a random sample $X_1, \cdots, X_n$ from $\mathcal{N}(\mu, \sigma^2)$. We know from Theorem 1.9 that $E\bar{X} = \mu$, $ES^2 = \sigma^2$.*

$$\begin{aligned} MSE(\bar{X}) &= var\bar{X} = \frac{\sigma^2}{n} \\ MSE(S^2) &= varS^2 = \frac{2\sigma^2}{n-1}. \end{aligned}$$

The MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{n-1}{n} S^2.$$

and

$$bias(\hat{\sigma}^2) = -\frac{1}{n}\sigma^2.$$

Hence, $\hat{\sigma}^2$ is asymptotically unbiased.

$$\begin{aligned} MSE(\hat{\sigma}^2) &= var(\hat{\sigma}^2) + (bias(\hat{\sigma}^2))^2 \\ &= \underbrace{\frac{2\sigma^4}{n-1}}_{MSE(S^2)} \cdot \underbrace{\frac{2n^2 - 3n + 1}{2n^2}}_{\leq 1} \\ &\leq MSE(S^2) \end{aligned}$$

Trade-off between the bias and the variance

- *Increasing the $(bias)^2$ led to a **decrease** of the variance and an overall decrease of the MSE.*

- *The MSE is just a criterion, meaning that we should not discard S^2 based on the MSE alone.*

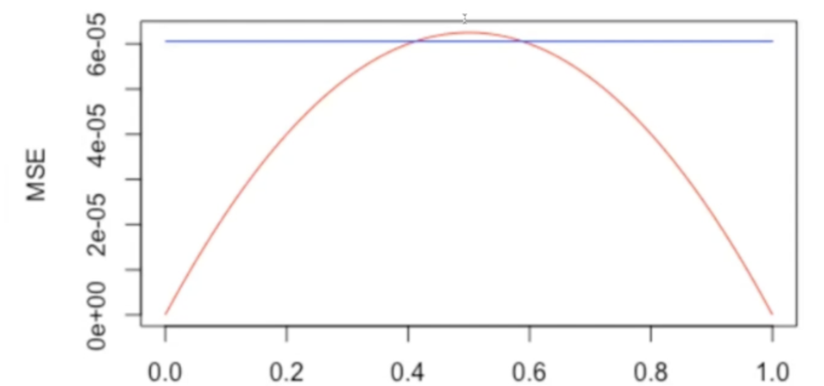
Example 2.19. *The Bayesian estimator of p is*

$$\hat{p}_B = \frac{n\bar{X} + \alpha}{n + \alpha + \beta}.$$

Clearly, $\hat{p}_B$ is biased.

$$MSE\hat{p}_B = \frac{\alpha^2 + p(n - 2\alpha^2 - 2\alpha\beta) + p^2(-n + \alpha^2 + \beta^2 + 2\alpha\beta)}{(n + \alpha + \beta)^2}.$$

We can decide to choose α and β so that the $MSE_{\hat{p}_B}$ does not depend on p. We get $\alpha = \beta = \frac{\sqrt{n}}{2}$.



When $p = 1/2$, the Bayesian estimator (the blue line) has the biggest advantage over the MLE (the red line), since the expectation of the prior, $Beta(\alpha, \beta)$, is

$$\frac{\alpha}{\alpha + \beta} = \frac{1}{2}.$$

Theorem (2.20)**.** *Suppose that T is asymptotically unbiased estimator of $\gamma(\theta)$ and $var_\theta T \to 0$ as $n \to \infty$ for all $\theta \in \Theta$. Then T is a consistent estimator of $\gamma(\theta)$.*

Proof. Fix an arbitrary $\varepsilon > 0$, and $\theta \in \Theta$. By Markov inequality,

$$\begin{aligned} P_\theta(|T - \gamma(\theta)| > \varepsilon) &\leq \frac{E_\theta(T(X_1, \cdots, X_n) - \gamma(\theta))^2}{\varepsilon^2} \\ &= \frac{MSE_\theta(T)}{\varepsilon^2} \\ &= \frac{var_\theta T + (bias_\theta T)^2}{\varepsilon^2} \xrightarrow{n\to\infty} 0. \end{aligned}$$

□

<u>Remark:</u>

we see from the proof that if T is an estimator of $\gamma(\theta)$ and $MSE_\theta T \to 0$ as $n \to \infty$, then T is consistent for $\gamma(\theta)$.

2.4 Best Unbiased Estimators

- Comparisons based on MSE may not yield a clean winner among estimators
- There is no "best MSE" estimator. Consider

  $$\{Bernoulli(p), p \in (0,1)\}.$$

  Let

  $$p_{\text{silly}} = 0.5.$$

  This is silly because the estimator does not use the data at all, but

  $$\begin{aligned} MSE_p(\hat{p}_{silly}) &= (0.5 - p)^2 \\ &= 0 \text{ when } p = 0.5. \end{aligned}$$

  Now, we can devise such silly estimator for any $p_0 \in (0,1)$:

  $$\hat{p}_{silly;p_0} = p_0 \to MSE_{p_0}(\hat{p}_{silly;p_0}) = 0.$$

- MSE that uniformly minimize MSE of all possible estimators would have to be 0 for any $p \in (0,1)$.

Definition 2.20. *An estimator T^* is called a uniform minimum variance unbiased estimator (UMVUE) of $\gamma(\theta)$ if:*

1. *T^* is unbiased: $E_\theta T^* = \gamma(\theta)$*
2. *T^* is "best" in terms of the variance: if T is an arbitrary unbiased estimator of $\gamma(\theta)$,*

   $$\forall \theta \in \Theta,\ \underbrace{var_\theta T^*}_{MSE_\theta T^*} \leq \underbrace{var_\theta T}_{MSE_\theta T}.$$

Example 2.21. *$X_1, \cdots, X_n$ a random sample from Poisson(λ), $\lambda \in (0, \infty)$. We derived earlier an estimator of λ:*

$$\hat{\lambda} = \bar{X}.$$

Theorem 2.22 (Cramer-Rao Inequality)**.** *Suppose that $X_1, \cdots, X_n$ is a random sample from P_θ, $\theta \in \Theta \subset \mathbb{R}$. Let $T(X_1, \cdots, X_n)$ be an unbiased estimator of $\gamma(\theta)$, i.e.*

$$\forall \theta \in \Theta, \; E_\theta T = \gamma(\theta).$$

Let $X \sim P_\theta$. Assume that the conditions (1), (2), (3) below holds:

(1) For all $\theta \in \Theta$, P_θ had a pdf/ pmf $f(x;\theta)$ and

$$\frac{\partial f}{\partial \theta}$$

exists for all $\theta \in \Theta$ and all $x \in N_\theta$.

(2) $\forall \theta \in \Theta$,

$$E_\theta \left(\frac{\partial log f}{\partial \theta}(X;\theta) \right) = 0$$

and

$$E_\theta \left((\frac{\partial log f}{\partial \theta}(X;\theta))^2 \right) = I(\theta) \in (0, \infty)$$

for all $\theta \in \Theta$. Here, $I(\theta)$ is called the <u>Fisher Information</u>.

(3) $var_\theta T(X_1, \cdots, X_n) < \infty$ for all $\theta \in \Theta$ and

$$\sum_{i=1}^{n} E_\theta \left\{ T(X_1, \cdots, X_n) \cdot \frac{\partial log f}{\partial \theta}(X_i;\theta) \right\} = \gamma'(\theta)$$

for all $\theta \in \Theta$.

Then

$$var_\theta T(X_1, \cdots, X_n) \geq \frac{(\gamma'(\theta))^2}{n \cdot I(\theta)}.$$

Proof. Cauchy-Schwarz inequality:

$$(cov(Z, W))^2 \leq varZ \cdot varW.$$

□

Remarks

- Note that if $X \sim P_\theta$,

  $$P_\theta(X \in \{x : f(x;\theta) > 0\}) = 1.$$

  So we can assume wlog that $f(x;\theta) > 0$ for all $x \in N_\theta$ and $\theta \in \Theta$. Then

  $$\frac{\partial log f}{\partial \theta} = \frac{\frac{\partial f}{\partial \theta}}{f}$$

  exists for all $\theta \in \Theta$ and $x \in N_\theta$.

- Assumptions (2) and (3) really mean that we can interchange differentiation and either integration or summation as the case may be.

- Check if it is an exponential family

Example 2.23. *$X_1, \cdots, X_n$ us Bernoulli(p), $p \in (0,1)$. $\bar{X}$ is UMVUE for p.*

Recall that Cauchy-Schwarz inequality,

$$cov(X,Y) \leq \sqrt{varXvarY}.$$

Equality holds if and only if $\exists a, b \in \mathbb{R}$ so that

$$Y = aX + b \text{ a.s.}$$

Denoting $T = T(X_1, \cdots, X_n)$, an unbiased estimator of $\gamma(\theta)$ with finite variance and

$$W = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} log f(X_i;\theta)$$

then we have

Corollary 2.24. *Under the condition of the CR theorem (Thm 2.22), T attains the CR lower boudn if and only if*

$$a(\theta)\cdot(T-\gamma(\theta)) = W \;\; P_\theta - a.s.$$

Example 2.23 (cont'd) $X_1, \cdots, X_n$, a random sample from Bernoulli(p), $p \in (0,1)$.

$$\begin{aligned} W &= \sum_{i=1}^{n} \frac{\partial}{\partial p} log f(X_i;p) \\ &= \sum_{i=1}^{n} (\frac{X_i}{p} + \frac{(1-X_i)}{1-p}) \\ &= \frac{n\bar{X} - np}{p(1-p)}. \end{aligned}$$

Suppose we wish to estimate the ODDs

$$\gamma(\theta) = \frac{p}{1-p}$$

In order for T to attain the CR lower bound

$$\frac{p}{n(1-p)^3},$$

we have to have that $T = a(n)\bar{X} + b(n)$, but $ET = a(n) \cdot p + b(n) \neq \frac{p}{1-p}$ for all $p \in (0,1)$. Hence, the CR lower bound for estimating the odds cannot be attained.

Definition 2.25 (One-parameter exponential family)**.** *A family of PDFs/ PMFs is called a one-paramter exponential family in $c(\theta)$ and $T(x)$, if, for all $\theta \in \Theta \subset \mathbb{R}$,*

$$f(x;\theta) = 1_A(x) \exp\{c(\theta)T(x) + d(\theta) + S(x)\}$$

for some set $A \subset \mathbb{R}$ which does not depend on θ and is a Borel set,, $c : \Theta \to \mathbb{R}$, and $S, T : \mathbb{R} \to \mathbb{R}$ Borel-measurable, and T is not a.s. constant on A.

Example 2.26. *Bernoulli(p):*

$$f(x;p) =_p p^x(1-p)^{1-x}, x \in \{0,1\}.$$

$$A = \{0,1\}.$$

On A,

$$\begin{aligned} f(x;p) &= \exp\{x \cdot \log p + (1-x) \cdot \log(1-p)\} \\ &= \exp\{\underbrace{x}_{T(x)} \cdot \underbrace{\log\frac{p}{1-p}}_{c(p)} + \underbrace{\log(1-p)}_{d(p)}\}. \end{aligned}$$

Remark

One can prove that for $\Theta = (a,b)$, $-\infty \leq a < b \leq \infty$, $c : \Theta \to \mathbb{R}$ is continuously differentiable with $c'(\theta) > 0$ for all $\theta \in \Theta$, then the assumptions of the CR Theorem 2.22 are fulfilled. Since

$$\frac{\partial}{\partial\theta} log f(x;\theta) = c'(\theta)T(x) + d'(\theta)$$

than

$$Z = \frac{1}{n}\sum_{i=1}^{n} T(X_i)$$

is an UMVUE of $\gamma(\theta) = ET(X)$ (assuming $ET^2(X) < \infty$) by Theorem 2.22.

Example 2.27 (Uniform $(0, \theta)$)**.** *A unbiased estimator of θ is*

$$T = \frac{n+1}{n} X(n).$$

$$varT = \frac{\theta^2}{n(n+2)} << \frac{\theta^2}{n}, \ \textit{CR lower bound.}$$

. Hence, we need a deeper theory to find UMVUE.

3 Chapter 3: Sufficiency and Completeness

3.1 Suffiency

Can we summarize the data without losing information about θ?

Notation: the support of $(X_1, \cdots, X_n)$, the so called sample space, is denoted by χ.

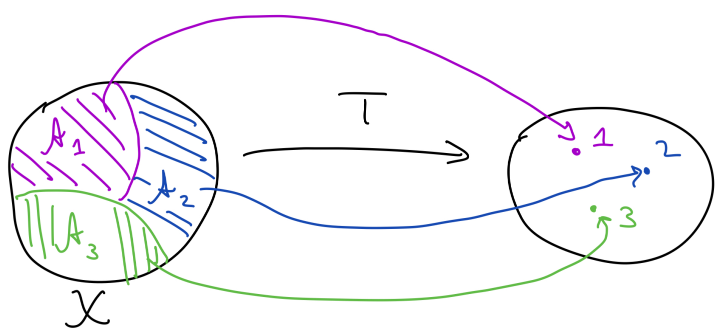
Basic observation Any statistic T induces a partition of χ. Indeed, let

$$\tau = \{t : t = T(\underset{\sim}{x}) \text{ for some } \underset{\sim}{x} \in \mathcal{X}\}.$$

The sets

$$\mathcal{A}_t = T^{-1}\{t\} = \{\underset{\sim}{x} \in \mathcal{X} : T(\underset{\sim}{x}) = t\}$$

form a partition of the sample space.



The statistic T summarizes the data (i.e. reduces information). $T = t$ really means that $(X_1, \cdots, X_n) \in \mathcal{A}_t$.

T contains all relevant information about θ if the exact value of $\underset{\sim}{x} \in \mathcal{A}_t$ contains no additional information about θ.

Definition 3.1 (Sufficient statistic)**.** *A statistic $T(X_1, \cdots, X_n)$ is a sufficient statistic for θ if the conditional distribution of $(X_1, \cdots, X_n)$ given $T(X_1, \cdots, X_n) = t$ does not depend of θ.*

Example 3.2.

- $(X_1, \cdots, X_n)$ *is sufficient for θ: the conditional distribution of $(X_1, \cdots, X_n)$ given $(X_1, \cdots, X_n) = \underset{\sim}{x}$ is degenerate.*
- $X_1, \cdots, X_n$ *be a random sample from Bernoulli(p), $p \in (0, 1)$.*

$$T(X_1, \cdots, X_n) = \sum_{i=1}^{n} X_i.$$

Here, $\chi = \{0, 1\}^n$, $T = \{0, 1, \cdots, n\}$,

$$\mathcal{A}_t = \{(x_1, \cdots, x_n) \in \{0, 1\}^n : \sum_{i=1}^{n} x_i = t\}.$$

For all $(x_1, \cdots, x_n) \in \mathcal{X}$, $t \in \tau$,

$$\begin{aligned} &P_\theta\left((X_1, \cdots, X_n) = (x_1, \cdots, x_n) | T(X_1, \cdots, X_n) = t\right) \\ =&\begin{cases} 0 & \text{if } \underset{\sim}{x} \notin \mathcal{A}_t \\ \frac{1}{\binom{n}{t}} & \text{if } \underset{\sim}{x} \in \mathcal{A}_t \end{cases} \end{aligned}$$

does not depend on p, so $T = \sum_{i=1}^{n}$ is sufficient for p.

Theorem 3.3 (Neyman-Fisher Factorization)**.** *Let $f(x_1, \cdots, x_n; \theta)$ denote the joint pdf/pmf of $(X_1, \cdots, X_n)$. A statistic T is sufficenit for θ if and only if for all $\theta \in \Theta$, there exists measurable function g_θ, h so that*

$$f(x_1, \cdots, x_n; \theta) = g_\theta(T(x_1, \cdots, x_n)) \cdot h(x_1, \cdots, x_n).$$

Proof. □

Example 3.4. *$X_1, \cdots, X_n$ is a random sample from $N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$.*

$$f(x_1, \cdots, x_n; \mu, \sigma^2) = (\frac{1}{2\pi})^{n/2} (\frac{1}{\sigma^2})^{n/2} \exp(-\frac{\sum_{i=1}^n (x_i - \mu)^2)}{2\sigma^2}.$$

Clearly, $(X_1, \cdots, X_n)$ is sufficient for (μ, σ^2). But

$$\begin{aligned} &\sum_{i=1}^n (x_i - \mu)^2 \\ =&\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \\ =&(n-1)s^2 + n(\bar{x} - \mu)^2 \end{aligned}$$

$$f(x_1, \cdots, x_n; \mu, \sigma^2) = \underbrace{(\frac{1}{2\pi})^{n/2}}_{h(\underset{\sim}{x})} \cdot \underbrace{(\frac{1}{\sigma^2})^{n/2} \exp(-\frac{(n-1)s^2 + n(\bar{x} - \mu)^2}{2\sigma^2})}_{g_{\mu,\sigma^2}(\bar{x}, s^2)}$$

Using Thm 3.3 (Neyman-Fisher factorization), we conclude that $(\bar{X}, S^2)$ is sufficient for (μ, σ^2). Assume now that σ^2 is known. Here, $(\bar{X}, S^2)$ is sufficient for μ. But, we can also write

$$f(x_1, \cdots, x_n; \mu, \sigma^2) = \underbrace{(\frac{1}{2\pi})^{n/2} (\frac{1}{\sigma^2})^{n/2} \exp(-\frac{(n-1)s^2}{2\sigma^2})}_{h(\underset{\sim}{x})} \cdot \underbrace{\exp(-\frac{n(\bar{x} - \mu)^2}{2\sigma^2})}_{g_\mu(\bar{x})}$$

Hence, $\bar{X}$ is sufficient for μ.

Remark: Sufficient statistic is generally not unique. Some statistics achieve greater data reduction than others. Also, the dimension of paramters nad the dimension of statistics are unrelated.

Example 3.5. *Consider a random sample form $U(\theta, \theta+1)$, $\theta \in \mathbb{R}$.*

$$
\begin{aligned}
& f(x_1, \cdots, x_n; \theta) \\
= & \begin{cases} 1, & \textit{if } \theta < x_i < \theta + 1 \\ 0, & \textit{otherwise} \end{cases} \\
= & \underbrace{1(\min_{1 \le i \le n} > \theta) \cdot 1(\max_{1 \le i \le n} < \theta + 1)}_{g\theta(\min_{1 \le i \le n} x_i;\ \max_{1 \le i \le n} x_i)}
\end{aligned}
$$

Using the Neyman-Fisher factorization, we have that

$$(\min_{1 \le i \le n} X_i, \max_{1 \le i \le n} X_i)$$

is sufficient for θ.

Example 3.6. *Consider a random sample from $U(0, \theta)$*

Consider a random sample from $U(0, \theta)$, $\theta > 0$.

$$
\begin{aligned}
& f(x_1, \cdots, x_n; \theta) \\
= & \begin{cases} (\frac{1}{\theta})^n, & \text{if } 0 < x_i < \theta \\ 0, & \text{otherwise} \end{cases} \\
= & \underbrace{(\frac{1}{\theta})^n \cdot 1(\max_{1 \le i \le n} x_i < \theta)}_{g_\theta(\max_{1 \le i \le n} x_i)} \cdot \underbrace{1(\min_{1 \le i \le n} x_i > 0)}_{h(x_1, \cdots, x_n)}
\end{aligned}
$$

By the Neyman-Fisher factorization, $\max_{1 \le i \le n} X_i$ is sufficient for θ.

3.2 The Rao-Blackwell Theorem

Recall X, Y random variables

$$E(X) = E(E(X|Y))$$

and $E(X|Y)$ is a measurable function of Y.

$$var(X) = E(var(X|Y)) + var(E(X|Y)).$$

Theorem 3.7 (Rao-Blackwell Theorem)**.** *Let W be an unbiased estimator of $\gamma(\theta)$ with finite varaince, and T be a sufficient statistic for θ. Let*

$$W^* = E(W|T).$$

Then

(a) W^* is an unbiased estimator of $\gamma(\theta)$.

(b) For all $\theta \in \Theta$:

$$var_\theta W^* \leq var_\theta W.$$

Example 3.8.

Remark

- *Process of conditioning on a sufficient statistic is called "Rao-Blackwellization".*
- *Theorem 3.7 implies that an UMVUE (if it exists) needs to be based on a sufficient statistic.*

Corollary 3.9. *Let W be an estimator of $\gamma(\theta)$ with finite variance, but not necessarily unbiased. Let T be a sufficient statistic for θ. Then for*

$$W^* = E(W|T),$$

$$MSE_\theta(W^*) \leq MSE_\theta(W) \quad \forall \theta \in \Theta.$$

3.3 Completeness

Suppose that T is a statistic and g is a measurable function such that

$$\forall\theta \in \Theta,\ E_\theta g(T) =$$

we have that

$$\forall\theta \in \Theta, \quad E_\theta g(T) = 0.$$

Assume, for simplicity $\Theta \in \mathbb{R}$ and we wish to estimate θ. Suppose W is an unbiased estimator of θ. Suppose that $g(T)$ is not degenerate (i.e. is a constant a.s.). Then for any $a \in \mathbb{R}$,

$$W_a = W + g(T) \cdot a$$

then W_a is also an estimator of θ :

$$\begin{aligned} E_\theta(W_a) &= E_\theta(W) + a \cdot E_\theta(g(T)) \\ &= \theta + a \cdot 0 = \theta. \end{aligned}$$

Assume further that W and $g(T)$ have a finite variance. Suppose that $cov_{\theta_0}(W, g(T)) \neq 0$ for some $\theta_0 \in \Theta$. Then, WLOG assume $cov_{\theta_0}(W, g(T)) <$ 0:

$$\begin{aligned} var_{\theta_0} = {} & var_{\theta_0}(W) + a^2 \cdot var_{\theta_0}(g(T)) \\ & + 2a \cdot cov_{\theta_0}(W, g(T)) \end{aligned}$$

Then,

$$\begin{aligned} var_{\theta_0} - var_{\theta_0}(W) = {} & a^2 \cdot var_{\theta_0}(g(T)) \\ & + 2a \cdot cov_{\theta_0}(W, g(T)). \end{aligned}$$

The RHS is negative if $a > 0$ and

$$a \cdot var_{\theta_0} g(T) < -2 \cdot cov_{\theta_0}(W, g(T))$$
$$a < \underbrace{\frac{-2 \cdot cov_{\theta_0}(W, g(T))}{var_{\theta_0}(g(T))}}_{=a^*>0}$$

Hence, for $a \in (0, a^*)$,

$$var_{\theta_0} W_a < var_{\theta_0} W.$$

Note that if T is complete, no such a^* exists.

Definition 3.10 (Completeness)**.** *A statistic T is called complete, if the family $\{P_\theta^T,\ \theta \in \Theta\}$ is complete, meaning that if for any measurale $g : T \to \mathbb{R}$ such that*

$$\forall \theta \in \Theta, \mathbb{E}(g(t)) = 0,$$

we have

$$\forall \theta \in \Theta,\ P_\theta(g(T) = 0) = 1.$$

Remark: T is complete if $\forall \theta \in \Theta,\ E_\theta(g(T)) = 0$ implies that $g(T) = 0$ $[P]$ *a.e.* Then, clearly, $cov_\theta(W, g(T)) = 0$ for all $\theta \in \Theta$, for any unbiased estimate W.

Example 3.11. *Completeness tells us something about the size of*

$$\{P_\theta^T,\ \theta \in \Theta\}.$$

Consider $X_1, \cdots, X_n$ a random sample from Bernoulli(p), $p \in \Theta \subset (0, 1)$. Take $T = \sum_{i=1}^n X_i$. Then $T \sim Binomial(n, p)$. Hence

$$E_p(g(T)) = \sum_{k=0}^{n} g(h) \binom{n}{k} p^k (1-p)^{n-k}.$$

So $E_p(g(T)) = 0$ for all $p \in \Theta$ means that

$$0 = \sum_{k=0}^{n} \underbrace{g(k)\binom{n}{k}}_{a_k} \cdot (1-p)^n \cdot \underbrace{(\frac{p}{1-p})}_{r}^k$$

$$(*) \quad 0 = \sum_{k=0}^{n} a_k r^k, \ p \in \Theta$$

For T to be complete, we need to conclude that $g(h) = 0$ for all $k = \{0, \cdots, n\}$, i.e. $a_k = 0$ for al $k \in \{0, \cdots, n\}$.

- *If $\Theta = (0,1)$, then $r = \frac{p}{1-p} \in (0, \infty)$. Hence, (*) means that the polynomial vanishes for all $r \in (0, \infty)$, and that indeed implies that $a_k = 0$ for all $k \in \{0, \cdots, n\}$, so T is complete.*

- *If Θ is finite and $|\Theta| \leq n$, it may well happen that $a_k \neq 0$ for some k. For example, if $\Theta = \{1/2\}$, then (*) becomes (say n = 1):*

  $$0 = g(0) + g(1)$$

  which does not imply

  $$g(0) = g(1) = 0.$$

  Hence, T is NOT complete.

Example 3.12. *Consider a random sample $X_1, \cdots, X_n$ from $U(0, \theta)$, $\theta > 0$.*

$$T = \max_{i \leq i \leq n} X_i.$$

Then,

$$P_\theta(T \leq t) = \prod_{i=1}^{n} P_\theta(X_i \leq t) = \begin{cases} (t/\theta)^n, \ t \in (0, \theta) \\ 0, \ t \leq 0 \\ 1, \ t \geq \theta \end{cases}$$

So T has a pdf:

$$f_\theta^T(t) = \frac{n}{\theta^n} \cdot t^{n-1}, \ t \in (0, \theta).$$

Suppose that g is measurable and such that $E_\theta g(T) = 0$ for all $\theta > 0$. Suppose that g is Riemann-integrable.

$$E_\theta g(T) = 0 \iff 0 = \int_0^\theta g(t) \cdot \frac{n}{\theta^n} \cdot t^{n-1} dt$$

Fix $\theta \in \Theta$ arbitrary. Then $E_\theta g(T) = 0$ implies

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta} \int_0^\theta g(t) \frac{n}{\theta^n} t^{n-1} dt \\
&= (\frac{\partial}{\partial\theta} \theta^{-n}) \cdot \underbrace{\theta^n \int_0^\theta g(t) \frac{n}{\theta^n} t^{n-1} dt}_{=0 \ \textit{because} \ E_\theta g(T)=0} \\
&+ \theta^{-n} \cdot \frac{\partial}{\partial\theta} \int_0^\theta g(t) n \cdot t^{n-1} dt \\
&= \theta^{-n}[g(\theta) n \cdot \theta^{n-1}] \\
&= \frac{g(\theta) \cdot n}{\theta} \ \textit{by Leibnitz rule}
\end{aligned}$$

Hence, $g(\theta) = 0$ implies $g(t) = 0$ for $t > 0$ for any $\theta > 0$. Then, $P_\theta(g(T) = 0) = 1$ for all $\theta > 0$. Hence, T is complete.

Theorem 3.13 (Lehmann-Scheffe)**.** *$X_1, \cdots, X_n$ a random sample from P_θ, $\theta \in \Theta$. Suppose that T is a sufficient and complete statistic. Let $\gamma(\theta)$ be a real-valued parameter, and let W be an unbiased estimator of $\gamma(\theta)$ with finite variance. Then*

$$W^* = E(W|T)$$

is UMVUE for $\gamma(\theta)$.

Remark:

- We see from the proof that the UMVUE is a.s. unique.

- If T is complete and sufficient and $W = h(T)$ is unbiased, then W is UMVUE.

Example 3.14.

- *$T = \max_{i \le i \le n} X_i$ is complete.*

- *T is sufficient*

- *$\frac{n+1}{n} T$ is an unbiased estimator of θ.*

Hence, by Lehmann-Scheffe theorem, $\frac{n+1}{n} \max_{1 \le i \le n}$ is UMVUE.

Theorem 3.15. *Suppose $X_1, \cdots, X_n$ are iid from a distribution in a J-parameter exponential family, that is, the PDF/PMF has the form*

$$f(x; \theta) = 1(x \in A) \exp\{\sum_{i=1}^{J} c_j(\theta) T_j(x) + d(\theta) + S(x)\}$$

where $J \ge 1$, $A \subset \mathbb{R}$ is a Borel set independent of θ, $c_1, \cdots, c_j$, $d : \Theta \to \mathbb{R}$; $T_1, \cdots, T_J$, $S : \mathbb{R} \to \mathbb{R}$ measurable and $T_1, \cdots, T_J$ are not a.s. constant. Then

$$T = \left(\sum_{i=1}^{n} T_1(X_i), \cdots, \sum_{i=1}^{n} T_J(X_i,) \right)$$

is sufficient for θ. If

$$\{(c_1(\theta), \cdots, c_J(\theta) : \theta \in \Theta)\}$$

contains an open subset in $\mathbb{R}^J$, T is complete.

Example 3.16.

- *Bernoulli:*

$$\begin{aligned} f(x; p) &= p^x (1-p)^{1-x} 1(x \in \{0, 1\}) \\ &= 1(x \in \{0, 1\}) \exp\{x \cdot \log \frac{p}{1-p} + \log(1-p)\} \end{aligned}$$

where $J = 1$, $S(x) = 0$. *By Theorem 3.15,* $\sum_{i=1}^n X_i$ *is sufficient for* p. *The set*

$$\{\log \frac{p}{1-p},\ p \in (0,1)\} = (-\infty, \infty).$$

Hence, $\sum_{i=1}^n X_i$ *is complete.*

- *Uniform:* $f(x;\theta) = \frac{1}{\theta}1(x \in (0,\theta))$ *is not an exponential form since* $A = (0,\infty)$ *depends on* θ.

4 Chapter 4: Hypothesis Tests

4.1 Basic terminology of hypothesis testing

Definition 4.1 (Hypothesis). *A hypothesis is a statement about a population parameter. Given a parametric model for the population distribution, viz*

$$\{P_\theta,\ \theta \in \Theta\}$$

we have

- *the null hypothesis ("the null")*

$$H_0 : \theta \in \Theta_0$$

  where $\Theta_0 \subset \Theta$ is some fixed subset of the parameter space.

- *the alternative hypothesis (the "alternative")*

$$H_1 : \theta \notin \Theta_0$$

  When $|\Theta_0| = 1$, H_0 is called simple; otherwise, it is called composite, and analogously for H_1.

Definition 4.2 (Hypothesis test). *A hypothesis test is a decision rule that specfies for which sample values H_0 is rejected and for which it is not. Formally, a hypothesis test is a measurable map*

$$\psi : \chi \to [0, 1].$$

The observed value $\psi(x_1, \cdots, x_n)$ is the probablity of rejecting H_0 when

$$(X_1, \cdots, X_n) = (x_1, \cdots, x_n).$$

- $$R = \{(x_1, \cdots, x_n) \in \mathcal{X} : \psi(x_1, \cdots, x_n) = 1\}$$

  is called the rejection region.

- $$A = \{(x_1, \cdots, x_n) \in \mathcal{X} : \psi(x_1, \cdots, x_n) = 0\}$$

  is called the acceptance region.

- $$U = \{(x_1, \cdots, x_n) \in \mathcal{X} : \psi(x_1, \cdots, x_n) \in 0, 1()\}$$

  is called the randomization region.

If $U \neq \emptyset$, ψ is called a randomized test.

Example 4.3. *Coffee bean: good - 0, spoiled - 1*
$X_1, \cdots, X_n$ sample of coffee beans

- *test statistic:*

$$T = \sum_{i=1}^{n} X_i = \text{ “number of spoiled beans”}$$

- *pick $c \in \{0, \cdots, n+1\}$*

- $$\psi(X_1, \cdots, X_n) = \begin{cases} 1, T \geq c \\ 0, T < c \end{cases} = 1(T \geq c)$$

Any test can have 4 possible outcomes:

TRUTH \ DECISION	Accept H_0	Reject H_0
H_0 is true	✓	Type I error "false positive"
H_0 is false	Type II error "false negative"	✓

- Medical test :
  - H_0: healthy
  - H_1: infected
- Trial :
  - H_0: innocent
  - H_1: guilty
- Exam :
  - H_0: student deserves to pass
  - H_1: student does not deserve to pass

		Exam	
		pass	fail
TRUTH	pass	✓	Type I failing a good student
	fail	Type II passing a poor student	✓

Extreme exams :

- super easy :

  ⇒ everyone passes

  ⇒ type I error does not occur

  ⇒ type II error blows up.

- super tough
  - every fails
  - type 2 error does not occur
  - type 1 error blows up
- Department chair: make sure that at most 5% (or α%) of good students fails $\implies$ control the Type 1 error $\implies$ LEVEL
- While controlling type 1 error, we can try to minimize the type 2 error, or maximize the power of the test (to detect the alternative, i.e. fail poor students)

Definition 4.4 (Power function). *The power function of a hypothesis test ψ is*

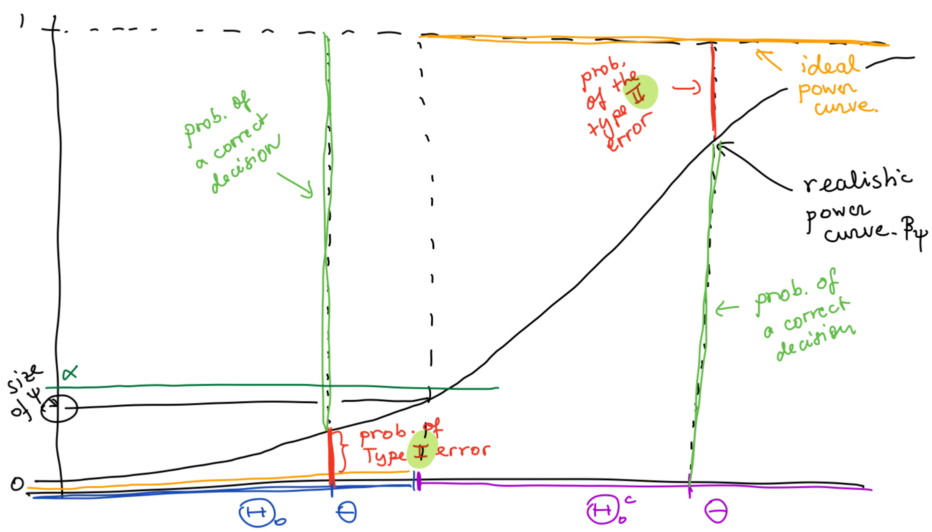
$$\begin{aligned} B_\psi : &\Theta \to [0,1] \\ &\theta \to E_\theta(\psi(X_1, \cdots, X_n)) \end{aligned}$$

If ψ is not randomized, $B_\psi(\theta)$ is the probablity of rejecting H_0. For a given $\alpha \in [0,1]$, ψ is called a level-α test if

$$\forall \theta \in \Theta_0 : B_\psi(\theta) \le \alpha.$$

The size of ψ is $\sup_{\theta \in \Theta_0} B_\psi(\theta)$.

" The power curve says it all "



A level-α test controls type 1 error, but not necessarily the type 2 error.

- Rejecting H_0 is a "safe" decision
- Accpting H_0 is NOT a "safe" decision. That's why we say "the data do not provide sufficient evidence to reject H_0" or "do not reject H_0".
- If possible, the scientific hypothesis we wish to prove should be the alternative. Sometimes, it is not possible. For example, we want to know if the snowfall is from a normal distribution.

Example 4.1 (cont'd)

$H_0 : \theta \leq \frac{1}{100}$ $\quad$ $H_1 : \theta > \frac{1}{100}$

$$T = \sum_{i=1}^{n} X_i \sim Binomial(n, \theta).$$

$$B_\psi(\theta) = P_\theta(T \geq c) = \sum_{k=c}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

- if $c = 0$, $B_\psi(\theta) = 1$ for all $\theta \in (0, 1)$.
- if $c = n + 1$, $B_\psi(\theta) = 0$ for all $\theta \in (0, 1)$
- if $c \in \{1, \cdots, n\}$: B_ψ is strictly increasing in θ. $\implies$ The size of ψ is $B_\psi(\frac{1}{100})$.
- To choose c:
  - Control type 1 error:

    $$B_\psi(\frac{1}{100}) \leq \alpha = 0.05$$

    The larger c, the smaller the size.
  - Maximize the power: maximize B_ψ for $\theta > 1/100$. The smaller c, the larger the power.
  - Note: typically, increasing the sample size leads to a better power.

4.2 Likelihood Ratio Test

General strategy how to construct tests. Typically, we construct a test statistic

$$W(X_1, \cdots, X_n)$$

and identify values in the sample space χ for which W has an unlikely value if H_0 holds. This set of values in χ will form a rejection region R. The (non-randomized) test will be

$$\psi(X_1, \cdots, X_n) = 1((X_1, \cdots, X_n) \in R).$$

For test problems about the parameter θ,

$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \notin \Theta_0$$

a large class of tests can be obtained as follows:

Definition 4.5 (Likelihood ratio test)**.** *The likelihood ratio statistic for testing*

$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \notin \Theta_0$$

is $\lambda(X_1, \cdots, X_n)$ *given, at any* $(x_1, \cdots, x_n)$ *by,*

$$\lambda = \frac{\sup_{\theta \in \Theta_0 L(\theta; x_1, \cdots, x_n)}}{\sup_{\theta \in \Theta L(\theta; x_1, \cdots, x_n)}}.$$

A likelihood ratio test(LRT) has the rejection region

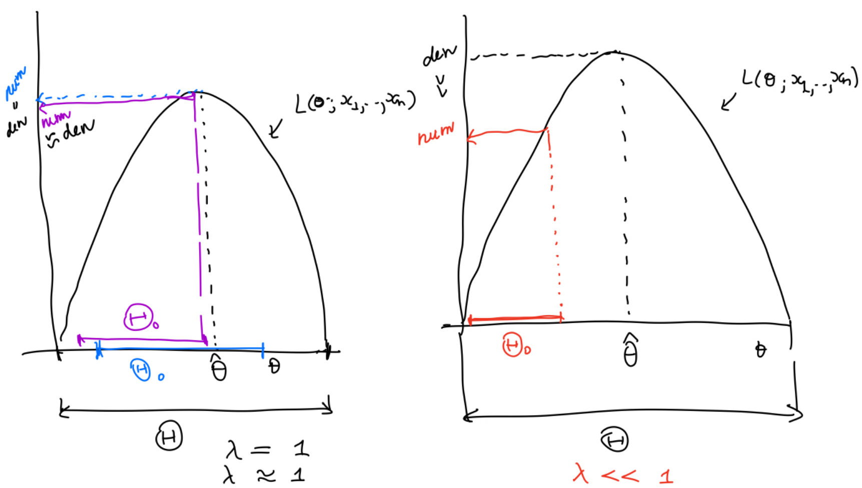
$$R = \{(x_1, \cdots, x_n) : \lambda(x_1, \cdots, x_n) \leq c\}$$

for some suitable chosen critical value c*, chosen as a function of* α *(the level of the test).*

Illustration:

1) H_0 holds

2) H_1 holds



How do we calculate the LR statistic λ?

- If $\hat{\theta}$ is MLE of θ and $\hat{\theta_0}$ is $\hat{\theta_0} = argmax_{\theta \in \Theta_0} L(\theta; X_1, \cdots, X_n)$, then

$$\lambda = \frac{L(\hat{\theta_0}; x_1, \cdots, x_n)}{L(\hat{\theta}; x_1, \cdots, x_n)}$$

Example 4.6. *We wish to test $H_0 : p \leq p_0$ vs $H_1 : p > p_0$ based on a random sample $X_1, \cdots, X_n$ from Bernoulli(p) (viz. Example 4.1). To construct a LRT, recall*

$$L(p; x_1, \cdots, x_n) = p^{n \cdot \bar{x}}(1-p)^{n(1-\bar{x})}, \; p \in [0,1]$$

we already know (Ex. 2.9) that the MLE of p is $\bar{X}$.

$$\hat{p_0} = arg \max_{0 \leq p \leq p_0} L(p; x_1, \cdots, x_n) = \min(p_0, \bar{x}).$$

4.3 p-value

Definition 4.7. *Let $W(X_1, \cdots, X_n)$ be a test statistic such that small (large) value of W give evidence against H_0 (are unlikely under H_0). For each*

$$(x_1, \cdots, x_n) \in \mathcal{X},$$

let

$$p(x_1, \cdots, x_n) = \sup_{\theta \in \Theta_0} P_\theta(W(X_1, \cdots, X_n) \leq (\geq) \underbrace{W(x_1, \cdots, x_n)}_{\textit{observed value of } W}),$$

"probablity of observing a value of W that is even more unlikely under H_0 than the one actually observed"
The random variable $p(X_1, \cdots, X_n)$ is called the p-value.

Definition 4.7 Let $W(X_1, \dots, X_n)$ be a test statistic such that small (large) values of W give evidence against H_0 (are unlikely under H_0)

For each $(x_1, \dots, x_n) \in \mathcal{X}$, let

$$* \quad p(x_1, \dots, x_n) = \sup_{\theta \in \Theta_0} P_\theta\left(W(X_1, \dots, X_n) \overset{\geq}{\leq} \underbrace{W(x_1, \dots, x_n)}_{\text{observed value of } W} \right)$$

"probability of observing a value of W that is even more unlikely under H_0 than the one actually observed"

The random variable $p(X_1, \dots, X_n)$ is called the p-value ∎

Note: the p-value is NOT the probability that H_0 holds!

Example 4.8 (p-value of a LRT)**.**

$$p(x_1, \cdots, x_n) = \sup_{\theta \in \Theta_0} (\lambda(X_1, \cdots, X_n) \leq \lambda(x_1, \cdots, x_n)).$$

Example 4.9 (Bernoulli)**.**

Theorem 4.10. *In the context of Definition 4.7, the test that rejects H_0 if $p(X_1, \cdots, X_n) \leq \alpha$ is a level-α test for all $\alpha \in [0, 1]$.*

Lemma 4.11. *For any random variable Y with distribution function G, $P(G(Y) \leq u) \leq u$ for all $u \in [0, 1]$.*

Proof. wlog:

$$p(x_1, \cdots, x_n) = \sup_{\theta \in \Theta_0} P_\theta(W \leq w(x_1, \cdots, x_n)).$$

For all $\theta \in \Theta$, let

$$\begin{aligned} p_\theta(x_1, \cdots, x_n) &= P_\theta(W(X_1, \cdots, X_n) \leq w(x_1, \cdots, x_n)) \\ &= F_\theta^W(W(x_1, \cdots, x_n)) \end{aligned}$$

From Lemma 4.11

$$\begin{aligned} &P_\theta(p_\theta(X_1, \cdots, X_n) \leq \alpha) \\ =&P_\theta(F_\theta^W(W(X_1, \cdots, X_n)) \leq \alpha) \leq \alpha \end{aligned}$$

Hence, for all $\theta^* \in \Theta_0$

$$P_{\theta^*}(p(X_1, \cdots, X_n) \leq \alpha) \leq P_{\theta^*}(p_{\theta^*}(X_1, \cdots, X_n) \leq \alpha) \leq \alpha$$

since

$$p(X_1, \cdots, X_n) = \sup_{\theta \in \Theta_0} p_\theta(X_1, \cdots, X_n) \geq p_{\theta^*}(X_1, \cdots, X_n)$$

Note: if you report the p-value

- the reader can choose α
- the smaller the p-value, the stronger the evidence against H_0.

$\square$

4.4 Small Sample Tests for Normal Samples

Throughout this lecture: $X_1, \cdots, X_n$ is a random sample from $N(\mu, \sigma^2)$.

Example 4.12 (z-test)**.** *Assume that $\sigma^2 \equiv \sigma_0^2$ is KNOWN and we wish to test*

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

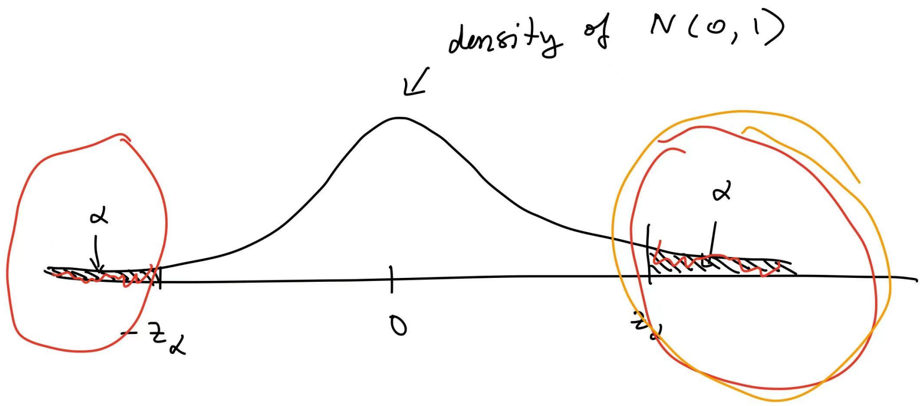
The Z statistic is

$$\sqrt{n}\frac{\bar{X} - \mu_0}{\sigma_0} \sim N(0, 1).$$

Definition 4.13 ((1-α)$\cdot$ 100% quantile of $N(0, 1)$)**.** *The* $(1-\alpha)100\%$ *quantile of $N(0, 1)$ is a value z_α such that*

$$1 - \Phi(z_\alpha) = \alpha = \Phi(-z_\alpha)$$

where Φ is the CDF of $N(0, 1)$.



- Two-sided z test: the level-α LRT for testing

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0$$

is

$$\psi(X_1, \cdots, X_n) = 1(\frac{\sqrt{n}}{\sigma_0}|\bar{X} - \mu_0| \geq z_{\alpha/2}).$$

p-value:

$$2(1 - \Phi(|z_{obs}|))$$

where

$$z_{obs} = \frac{\sqrt{n}}{\sigma_0}(\bar{x} - \mu_0)$$

- One-sided z test: if instead, we wish to test

  $$H_0 : \mu \leq \mu_0 \text{ vs } H_1 : \mu > \mu_0$$

  Recall that the likelihood function L is increasing on $(\infty, \bar{x}]$ and decreasing on $[\bar{x}, \infty)$. Hence,

  $$\hat{\mu_0} = \min(\bar{x}, \mu_0).$$

  $$\psi(X_1, \cdots, X_n) = 1(\frac{\sqrt{n}}{\sigma_0}(\bar{X} - \mu_0) \geq z_\alpha).$$

  p-value

  $$1 - \Phi(z_{obs})$$

- One-sided z test:

  $$H_0 : \mu \geq \mu_0 \text{ vs } H_1 : \mu < \mu_0$$

  $$\psi(X_1, \cdots, X_n) = 1(\frac{\sqrt{n}}{\sigma_0}(\bar{X} - \mu_0) \leq -z_\alpha).$$

  p-value

  $$\Phi(z_{obs})$$

Exmaple 4.12 (T test).
Suppose that both μ and σ^2 are unknown. (Note that σ^2 is a nuisance parameter.)

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0$$

The LRT has the form

$$\psi(X_1, \cdots, X_n) = 1(\frac{\sqrt{n}}{S}|\bar{X} - \mu_0| \geq c^*)$$

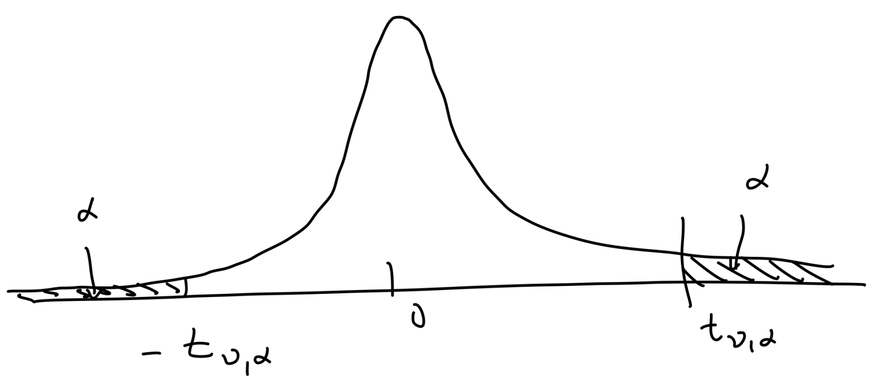
Recall from Theorem 1.26 that under H_0,

$$\text{T statistic} = \frac{\sqrt{n}}{S}(\bar{X} - \mu_0) \sim t_{n-1}$$

Definition 4.13 ($(1-\alpha)100\%$ quantile from the student t distribution) The $(1-\alpha)\cdot 100\%$ quantitle from the student t distribution with ν dof is $t_{\nu,\alpha}$ such that

$$P(T \geq t_{\nu,\alpha}) = \alpha$$

where $T \sim t_\nu$.



- Two-sided T-test:

$$\psi(X_1, \cdots, X_n) = 1(\frac{\sqrt{n}}{S}|\bar{X} - \mu_0| \geq t_{n-1,\alpha/2})$$

$$p - value = P(|T| \geq |t_{obs}|)$$

$$t_{obs} = \frac{\sqrt{n}}{s}(\bar{x} - \mu_0)$$

$$T \sim t_{n-1}$$

- One-sided T-test:

$$H_0 : \mu \leq \mu_0 \text{ vs } H_1 : \mu > \mu_0$$

The level-α LRT is

$$\psi(X_1, \cdots, X_n) = 1(\frac{\sqrt{n}}{S}(\bar{X} - \mu_0) \geq t_{n-1,\alpha})$$

$$p - value = P(T \geq t_{obs})$$

- One-sided T-test:

$$H_0 : \mu \geq \mu_0 \text{ vs } H_1 : \mu < \mu_0$$

The level-α LRT is

$$\psi(X_1, \cdots, X_n) = 1(\frac{\sqrt{n}}{S}(\bar{X} - \mu_0) \leq -t_{n-1,\alpha})$$

$$p - value = P(T \leq t_{obs})$$

Example 4.14 (F test). *Two independent random samples:*

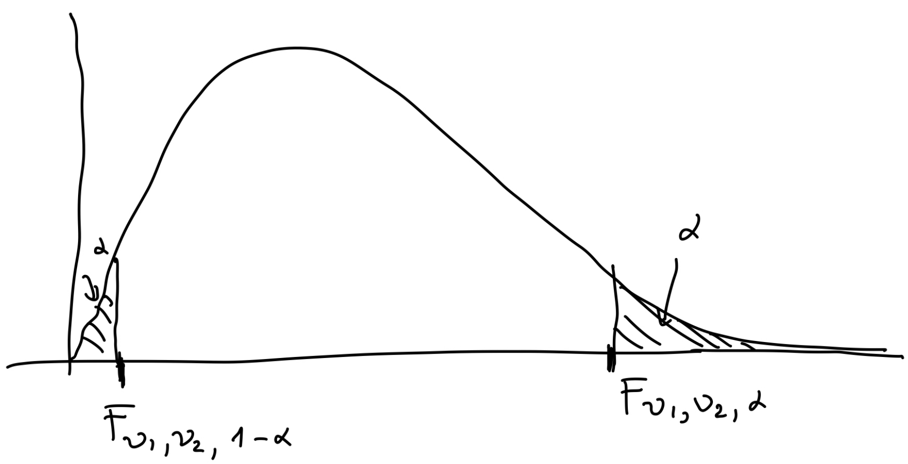
$$\underbrace{X_1, \cdots, X_n}_{\text{random sample from } N(\mu_1, \sigma_1^2))} \qquad \& \qquad \underbrace{Y_1, \cdots, Y_n}_{\text{random sample from } N(\mu_2, \sigma_2^2))}$$

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad vs \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

Definition 4.15. *The* $(1-\alpha) \cdot 100\%$ *quantile of the* F_{ν_1,ν_2} *distribution is* $F_{\nu_1,\nu_2,\alpha}$ *so that*

$$P(W \geq F_{\nu_1,\nu_2,\alpha}) = \alpha$$

where $W \sim F_{\nu_1,\nu_2}$*.*



The level-α LRT (F-test)

Assumptions:

- The samples are independent;
- The population distributions are normal for both samples.

$\psi(X_1, \cdots, X_m, Y_1, \cdots, Y_n) = 1\left(S_X^2/S_Y^2 \in (0, F_{m-1,n-1,1-\alpha/2}] \cup [F_{m-1,n-1,\alpha/2}, \infty)\right)$

p-values: $W_{obs} = S_X^2/S_Y^2$, $W \sim F_{m-1,n-1}$

$$p - value = \begin{cases} 2P(W \geq w_{obs}), \ w_{obs} > 1 \\ 2P(W \leq w_{obs}), \ w_{obs} \leq 1 \end{cases}$$

Remark 4.15 Other classical tests for normla samples that can be derived as LRTs:

(1) Chi-squared test: $X_1, \cdots, X_n$ random sample from $N(\mu, \sigma^2)$

$H_0: \sigma^2 \leq / \geq / = \sigma_0^2$ vs. $H_1: \sigma^2 > / < / \neq \sigma_0^2$

$$\psi = 1\left(\underbrace{\frac{(n-1)S^2}{\sigma_0^2}}_{\sim \chi^2_{n-1} \text{ when } \sigma^2 = \sigma_0^2} \begin{array}{l} \geq \chi^2_{n-1,\alpha} \\ \leq \chi^2_{n-1,1-\alpha} \\ \in (0, \chi^2_{n-1,1-\frac{\alpha}{2}}] \cup [\chi^2_{n-1,\frac{\alpha}{2}}, \infty) \end{array} \right)$$

$\left((1-\alpha)\cdot 100\%\text{ quantile of } \chi^2_{n-1}\right)$

(2) Two-sample t test: Assumptions:

- The samples are independent;
- The population distributions are normal for both samples, with the same variance

(and possibly different means). $X_1, \cdots, X_m$ & $Y_1, \cdots, Y_n$

two independent samples; $X_i \sim N(\mu, \sigma^2)$

$Y_i \sim N(\nu, \sigma^2)$

$$H_0 : \mu \;\; \le \,/\, \ge \,/\, = \;\; \nu \quad \text{vs.} \quad H_1 : \mu \;\; > \,/\, < \,/\, \neq \;\; \nu$$

$$\psi = 1\left(\boxed{\frac{\sqrt{\frac{mn}{(m+n)}}\,(\bar{X} - \bar{Y})}{\sqrt{\frac{1}{m+n-2}\left((m-1)S_X^2 + (n-1)S_Y^2\right)}}} \;\; \begin{array}{l} \ge t_{\cancel{n-1}\; n+m-2,\,\alpha} \\ \le -t_{\cancel{n-1}\; m+n-2,\,\alpha} \\ \in (-\infty, -t_{\cancel{n-1}\; n+m-2,\,\frac{\alpha}{2}}] \\ \quad \cup\, [t_{\cancel{n-1}\; m+n-2,\,\frac{\alpha}{2}}, \infty) \end{array} \right)$$

$\sim t_{n-2}$ when $\mu = \nu$

4.5 Uniformly most powerful tests

Recall the power of a test ψ:

$$\begin{aligned} B_\psi :& \Theta \to [0, 1] \\ & \theta \to B_\psi(\theta) = E_\theta \psi = P_\theta(\underset{\sim}{X} \in R) \end{aligned}$$

So far, we were controlling the type 1 error (level-α test):

$$\sup_{\theta \in \Theta_0} B_\psi(\theta) \le \alpha.$$

Now we can try to minimize the type 2 error, i.e. maximize $B_\psi(\theta)$, $\theta \in \Theta_1$, but we cannot minimize both types of error at the same time.

Definition 4.16 (UMP Test). *A test ψ is called a uniformly most powerful(UMP) level-α test if its power satistifes*

(a)

$$\sup_{\theta \in \Theta_0} B_\psi(\theta) \le \alpha$$

(b) *For any other level-α test ψ^* with B^*_ψ, we have that*

$$\forall \theta \in \Theta_1 : B_\psi(\theta) \geq B_{\psi^*}(\theta)$$

(i.e. ψ minimizes the type 2 error uniformly over Θ_1)

Definition 4.17. *H_i, $i \in \{0,1\}$ is called simple if Θ_i is a singleton, i.e. $|\Theta_i| = 1$. Otherwise, H_i is called composite.*

We will start developing a theory for finding UMP tests. We will begin by considering the case of testing a simple H_0 vs a simple H_1.

- $$\Theta = \{\theta_0, \theta_1\}$$
- $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$
- KNAPSACK Problem

Theorem 4.18 (Neyman-Pearson Lemma)**.** *Consider $\Theta = \{\theta_0, \theta_1\}$, $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$. Suppose that*

$$f(x_1, \cdots, x_n; \theta_i),\ i \in \{0,1\}$$

is the PDF/PMF of $(X_1, \cdots, X_n)$ when $\theta = \theta_i$. Define the so-called NP test ψ_k, $k \in [0, \infty]$:

$$\psi_k(x_1, \cdots, x_n) = \begin{cases} 1, & f(x_1, \cdots, x_n; \theta_1) \geq k \cdot f(x_1, \cdots, x_n; \theta_0) \\ 0, & f(x_1, \cdots, x_n; \theta_1) < k \cdot f(x_1, \cdots, x_n; \theta_0) \end{cases}$$

Then ψ_k is a UMP test for H_0 vs H_1 at level

$$\alpha = P_{\theta_0}(\psi_k(X_1, \cdots, X_n) = 1).$$

Remark 4.19. *If ψ_k is randomized test:*

$$\psi_k(\underset{\sim}{x}) = \begin{cases} 1, & f(\underset{\sim}{x}; \theta_1) > k \cdot f(\underset{\sim}{x}; \theta_0) \\ \gamma, & f(\underset{\sim}{x}; \theta_1) = k \cdot f(\underset{\sim}{x}; \theta_0) \\ 0, & f(\underset{\sim}{x}; \theta_1) < k \cdot f(\underset{\sim}{x}; \theta_0) \end{cases}$$

Example 4.20. *$X_1, \cdots, X_n$ from $N(\mu, \sigma_0^2)$, σ_0^2 is assumed to be known, so the parameter space is $\mathbb{R}$. Consider testing:*

$$H_0 : \mu \leq \mu_0 \text{ vs } H_1 : \mu > \mu_0$$

Fix an arbitrary $\mu_1 > \mu$. Consider testing the auxiliary problem:

$$H_0^* : \mu = \mu_0 \text{ vs } H_1^* : \mu = \mu_1$$

If we simply set $k^* = z_\alpha$,

$$\begin{aligned}\psi_{NP}(X_1, \cdots, X_n) &= 1(\frac{\sqrt{n}}{\sigma_0}(\bar{X} - \mu_0) \geq z_\alpha) \\ &= \psi_z(X_1, \cdots, X_n)\end{aligned}$$

which is a one-sided z test. Note that the test ψ_{NP} has nothing to do with μ_1. Hence, ψ_z is UMP for $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$.

Definition 4.21. *A family*

$$P = \{P_\theta : \theta \in \Theta \subset \mathbb{R}\}$$

of distribution with PMF/PDF $f(;\theta), \theta \in \Theta$ is said to have a monotone likelihood ratio(MLR) is a statistic $T : \chi \to \mathbb{R}$ if

(1)

$$\begin{aligned}\Theta &\to P \\ \theta &\to P_\theta\end{aligned}$$

is injective.

(2) *For every $\theta_1, \theta_2 \in \Theta$, $\theta_1 < \theta_2$,, there exists version of $f(\ ;\theta_1)$ $f(\ ;\theta_2)$ and a non-decreasing mapping $h(\ ;\theta_1,\theta_2) : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ so that*

$$\frac{f(\underset{\sim}{x};\theta_2)}{f(\underset{\sim}{x};\theta_1)} = h(T(\underset{\sim}{x});\theta_1,\theta_2)$$

on the set $\{x \in \mathcal{X} : f(\underset{\sim}{x};\theta_1) > 0 \text{ or } f(\underset{\sim}{x};\theta_1) > 0\}$; here "$\frac{a}{\infty} = 0$" if $a > 0$.

Example 4.22. *In the setup of Example 4.20,,*

$$P = \{P_\mu, \mu \in \mathbb{R}\}$$

has a MLR in $T = \bar{X}$.

Theorem 4.23 (Karlin-Rubin)**.** *Let $X_1, \cdots, X_n$ be a random sample and P the family of distribution of $(X_1, \cdots, X_n)$. Suppose*

$$P = \{P_\theta, \theta \in \Theta \subset \mathbb{R}\},$$

and P has a MLR in a statistic T.

$$H_0 : \theta \underset{\geq}{\leq} \theta_0 \quad \text{vs.} \quad H_1 : \theta \underset{<}{>} \theta_0$$

let $\alpha \in (0,1)$ and ψ_{KR} be a test given by

$$\psi_{KR}(\underset{\sim}{x}) = \begin{cases} 1 \\ \gamma \\ 0 \end{cases}, \quad T(\underset{\sim}{x}) \begin{matrix} \overset{<}{>} \\ = \\ \underset{>}{<} \end{matrix} k$$

where γ and k are such that

$$\boxed{(*) \quad P_{\theta_0}(T \underset{<}{>} k) + \gamma \cdot P_{\theta_0}(T = k) = \alpha .}$$

Then :

(1) ψ_{KR} *minimizes uniformly the type 2 and type 1 error among all tests* ψ *with* $E_{\theta_0}\psi = \alpha$.

(2) ψ_{KR} *is a UMP level* α *test for* H_0 *vs* H_1

(3) $B_{\psi_{KR}}$ *is non-decreasing (non-increasing) in* θ.

Remark 4.24. *Let* F_θ^T *denote the CDF of* T, *i.e.* $F_\theta^T(t) = P_\theta(T \leq t)$,

$$(F_\theta^T)^{-1}(u) = \inf\{x : F_\theta^T(x) \geq u\},\ u \in (0,1).$$

Then: for

$$H_0 : \theta \underset{\geq}{\leq} \theta_0 \quad \text{vs.} \quad H_1 : \theta \underset{<}{>} \theta_0$$

We can set

$$k = \left(F^T_{\theta_0}\right)^{-1} \left(\overset{\alpha}{1 - \alpha}\right)$$

$$\gamma = \begin{cases} \dfrac{\alpha - P_{\theta_0}(T \overset{<}{>} k)}{P_{\theta_0}(T = k)}, & \text{if } P_{\theta_0}(T=k) \neq 0 \\ 1, & \text{if } P_{\theta_0}(T=k) = 0 \end{cases}$$

Example 4.25. $X_1, \cdots, X_n$ *random sample from Poisson*(λ), $\lambda > 0$. P *has a MLR in* $T = \sum_{i=1}^n X_i$.

The UMP test for testing

$$H_0: \lambda \le \lambda_0 \quad vs. \quad H_1: \lambda > \lambda_0$$

is

$$\psi(x_1, \ldots, x_n) = \begin{cases} 1 & \\ \gamma & \sum_{i=1}^{n} x_i \quad \begin{matrix} > \\ = \\ < \end{matrix} \quad k \\ 0 & \end{cases}$$

For example, when $\alpha = 0.05$, $n = 10$, $\lambda_0 = 5$:

$$k = \left(F^T_{\lambda_0}\right)^{-1}(0.95) = \text{qpois}(0.95, 50)$$

$$= 62.$$

W is Poisson (50)

$$\gamma = \frac{0.05 - P(W > 62)}{P(W = 62)} = \frac{0.05 - 1 + \text{ppois}(62, 50)}{\text{dpois}(62, 50)}$$

$$\doteq 0.573$$

Note: if $X \sim Poisson(\lambda_1)$ and $Y \sim Poisson(\lambda_2)$ and X and Y are independent, then $X + Y \sim Poisson(\lambda_1 + \lambda_2)$.

Example 4.26. *Consider the setup of Example 4.20. We wish to test*

$$H_0 : \mu = \mu_0 \text{ } vs \text{ } H_1 : \mu \neq \mu_0$$

A UMP level-α test ψ would need to satisfy

- $$E_{\mu_0}\psi \le \alpha$$

- $$E_\mu \psi = \sup\{E_\mu \psi^* : \text{ } \psi^* \text{ is a test such that } E_{\theta_0}\psi^* \le \alpha\}$$

Now for all $\mu > \mu_0 : \psi$ would be UMP for

$$H_0 : \mu = \mu_0 \ vs \ H_1^* : \mu > \mu_0$$

for all $\mu < \mu_0 : \psi$ would be UMP for

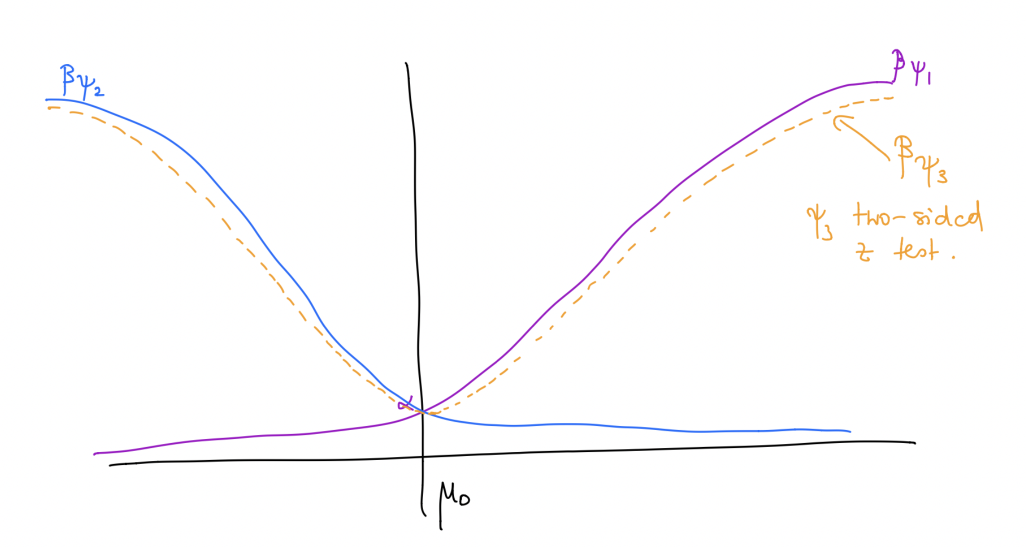
$$H_0 : \mu = \mu_0 \ vs \ H_1^{**} : \mu < \mu_0$$

$$\begin{aligned} \psi = \psi_1 &= 1(\frac{\sqrt{n}}{\sigma_0}(\bar{X} - \mu_0) \geq z_\alpha) \\ = \psi_2 &= 1(\frac{\sqrt{n}}{\sigma_0}(\bar{X} - \mu_0) \leq -z_\alpha) \end{aligned}$$

But

$$\{\underset{\sim}{x} : \psi_1 \neq \psi_2\} = \{\underset{\sim}{x} : \frac{\sqrt{n}}{\sigma_0}(\bar{x} - \mu_0) \geq z_\alpha \ or \ \frac{\sqrt{n}}{\sigma_0}(\bar{x} - \mu_0) \leq -z_\alpha\}$$

does not have probablity 0. So such a test ψ does not exist.



Convention: we can develop a theory of UMP level-α tests for the two-sided theory problems. ($\theta = \theta_0$ vs $\theta \neq \theta_0$) if we restrict attention to unbiased tests:

$$B_\psi(\theta) \geq \alpha \ \forall \theta \neq \theta_0$$

5 Chapter 5: Confidence Sets

5.1 Confidence set

Goal: express uncertainty in parametric estimates

Definition 5.1 (Confidence set)**.** *Consider a parametric model*

$$P = \{P_{\theta,\xi}, (\theta, \xi) \in \mathfrak{L}\}.$$

Here, θ is the parameter of interest and ξ is a nuisance parameter. Let $\Theta = \{\theta : (\theta, \xi) \in \mathfrak{L},$ for at least one $\xi\}$. The mapping

$$\begin{aligned} C :& \chi \to 2^{\Theta} \\ & (x_1, \cdots, x_n) \to c(\underset{\sim}{x}) \end{aligned}$$

is called a confidence set for θ if for all $\theta \in \Theta$ the set $\{\underset{\sim}{x} \in \mathcal{X} : \theta \in c(\underset{\sim}{x})\}$ is measurable.
A confidence set c has confidence level $1 - \alpha$ if $\forall \theta \in \Theta$, $\forall \xi : (\theta, \xi) \in \mathcal{L}$

$$P_{\theta,\xi}(\theta \in C(\underset{\sim}{X})) \geq 1 - \alpha$$

Remark If there are no nuisance parameters, ξ is simply omitted in Def 5.1 and $\mathcal{L} = \Theta$.

Example 5.2 (Constructing confidence sets using pivots)**.** *$X_1, \cdots, X_n$ random sample from the Exponential distribution with density*

$$f(x; \lambda) = \lambda e^{-\lambda x}, \; x > 0$$

$$P = \{Exp(\lambda), \lambda \in (0, \infty)\}$$

Goal: construct CS for λ.

Note:

$$\sum_{i=1}^{n} X_i \sim Gamma(n, \lambda)$$

Define

$$Q = 2(\sum_{i=1}^{n} X_i) \cdot \lambda = Q(\underset{\sim}{X}, \lambda) \sim \chi^2_{2n} \textit{ does not depend on } \lambda$$

The MGF of Q is

$$E_\lambda\left(e^{tQ}\right) = E_\lambda\left(e^{(2t\lambda)\sum_{i=1}^{n} X_i}\right) = \left(E_\lambda e^{(2\lambda t)X_i}\right)^n$$

$$= \left(1 - \frac{2t\lambda}{\lambda}\right)^{-n} = \boxed{(1-2t)^{-n}, \ t < \tfrac{1}{2}.}$$

MGF χ^2_{2n}

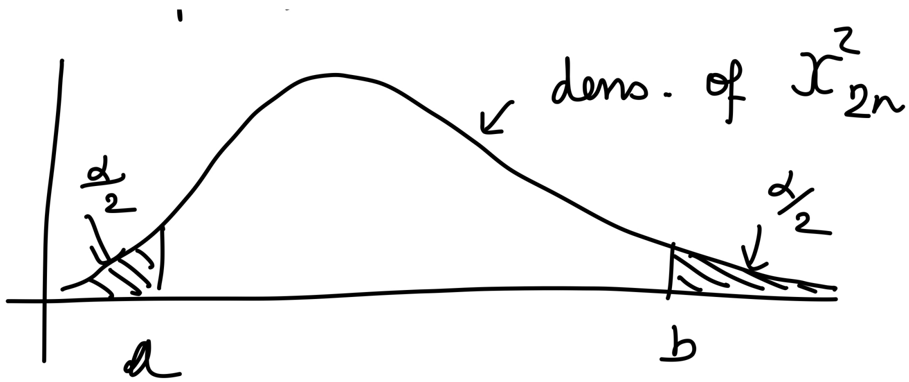
$$\Rightarrow \quad Q = Q(\underset{\sim}{X}, \lambda) \sim \chi^2_{2n} = \text{does not depend on } \lambda.$$

A quantity which depends on $(X_1, \cdots, X_n)$ and the parameter of interest θ, and whose distribution does not depend on θ or ξ is called a ***PIVOT***.

To construct a confidence set for λ from Q, we can simply choose (a, b) so that the CS is at confidence level $1 - \alpha$. Here, we choose $a, b \in \mathbb{R}$, $a < b$, so that

$$P(\chi^2_{2n} \in (a, b)) = 1 - \alpha.$$

For example, we can set $a = \chi^2_{2n,1-\alpha/2}$, $b = \chi^2_{2n,\alpha/2}$



To obtain the CS from (a, b), we can solve for

$$a < Q(\underset{\sim}{X}, \lambda) < b$$

$$\frac{a}{2\sum_{i=1}^n X_i} < \lambda < \frac{b}{2\sum_{i=1}^n X_i}$$

Set

$$C(\underset{\sim}{X}) = \left(\frac{a}{2\sum_{i=1}^n X_i}, \frac{b}{2\sum_{i=1}^n X_i}\right)$$

Then, for any $\lambda > 0$,

$$\begin{aligned} &P_\lambda\left(\lambda \in \left(\frac{a}{2\sum_{i=1}^n X_i}, \frac{b}{2\sum_{i=1}^n X_i}\right)\right) \\ =&P_\lambda\left(a < 2\left(\sum_{i=1}^n X_i\right) < b\right) \\ =&P(\chi^2_{2n} \in (a, b)) = 1 - \alpha \end{aligned}$$

Hence, $C(\underset{\sim}{X})$ above is a confidence set for λ at confidence level $1 - \alpha$.

Example 5.3 (More Pivots)**.** *$X_1, \cdots, X_n$ a random sample from $N(\mu, \sigma^2)$. We wish to construct a confidence set at level $(1 - \alpha)$ for μ (i.e. σ^2 is a nuisance parameter). Define*

$$Q(X_1, \cdots, X_n, \mu) = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$$

Choose (a, b)*, i.e., a,b* $\in \mathbb{R}$ *so that*

$$P(t_{n-1} \in (a, b)) = 1 - \alpha$$

Definition 5.4. *Suppose that* $C(\underset{\sim}{X})$ *is confidence set for* θ *at level* $1 - \alpha$*.*

- *If* $C(\underset{\sim}{X})$ *has the form* $(L(\underset{\sim}{X}), U(\underset{\sim}{X}))$*, then* C *is called a two-sided confidence interval at confidence level* $1 - \alpha$*.*
- *If* $C(\underset{\sim}{X})$ *has the form* $(\infty, U(\underset{\sim}{X})$*, then* C *is called upper one-sided confidence interval at confidence level* $1 - \alpha$*.*
- *If* $C(\underset{\sim}{X})$ *has the form* $(L(\underset{\sim}{X}), \infty)$*, then* C *is called lower one-sided confidence interval at confidence level* $1 - \alpha$*.*

Definition 5.5 (Unbiased confidence set)**.** *For any* $\theta \in \Theta$*, let* k_θ *be a set of undesirable parameters. A confidence set at confidence level* $1 - \alpha$ *is called unbiased if*

$$\forall \theta \in \Theta,\ \forall \xi : (\theta, \xi) \in \mathcal{L},\ \forall \theta^* \in k_\theta, P_{\theta,\xi}(\theta^* \in C(\underset{\sim}{X})) \leq 1 - \alpha$$

Example 5.6 (Ex 5.3 continued)**.** $X_1, \cdots, X_n$ *sample from* $N(\mu, \sigma^2)$*,* μ *of interest,* σ^2 *nuisance,* $k_\mu = (\infty, \mu)$*. For* $\mu^* \in k_\mu$*,*

$$\begin{aligned}
&P_{\mu,\sigma^2}(\mu^* \in (\bar{X} - \frac{t_{n-1,\alpha} \cdot S}{\sqrt{n}}, \infty)) \\
=&P_{\mu,\sigma^2}(\frac{\bar{X} - \mu}{S}\sqrt{n} < t_{n-1,\alpha} + \underbrace{\frac{\mu^* - \mu}{S}\sqrt{n}}_{<0}) \\
\leq&P_{\mu,\sigma^2}\left(\underbrace{\frac{\bar{X} - \mu}{S} \cdot \sqrt{n}}_{\sim t_{n-1}} < t_{n-1,\alpha}\right) = 1 - \alpha.
\end{aligned}$$

- *Similarly, if* $k_\mu = (\mu, \infty)$

$$(-\infty, \bar{X} + \frac{t_{n-1,\alpha} \cdot S}{\sqrt{n}})$$

is unbiased

- *Similarly, if* $k_\mu = \{\mu\}^C$

$$(\bar{X} - \frac{t_{n-1,\alpha} \cdot S}{\sqrt{n}}, \bar{X} + \frac{t_{n-1,\alpha} \cdot S}{\sqrt{n}})$$

is unbiased.

5.2 Correspondence between confidence sets and hypothesis tests

Theorem 5.7. *For any confidence set C, there exists a family of non-randomized tests*

$$\{\psi_{\theta_0}, \theta_0 \in \Theta\}$$

with

$$C(\underset{\sim}{x}) = \{\theta_0 \in \Theta : \psi_{\theta_0}(\underset{\sim}{x}) = 0\}$$

is measurable for all θ_0 since θ_0 is measurable.

Example 5.8. *$X_1, \cdots, X_n$ random sample from $N(\mu, \sigma^2)$. In Example 5.3, we derived CI for μ using pivots.*

- *lower one-sided confidence interval for μ:*

$$(\bar{X} - \frac{t_{n-1,\alpha} \cdot S}{\sqrt{n}}, \infty)$$

we can calculate, for $\mu_0 \in \mathbb{R}$,

$$\psi_{\mu_0}(\underset{\sim}{x}) = \begin{cases} 1, & \mu_0 \notin (\bar{X} - \frac{t_{n-1,\alpha} \cdot S}{\sqrt{n}}, \infty) \\ 0, & \mu_0 \in (\bar{X} - \frac{t_{n-1,\alpha} \cdot S}{\sqrt{n}}, \infty) \end{cases}$$

$$= \begin{cases} 1, & \mu_0 \leq \bar{X} - \frac{t_{n-1,\alpha} \cdot S}{\sqrt{n}} \\ 0, & \mu_0 > \bar{X} - \frac{t_{n-1,\alpha} \cdot S}{\sqrt{n}} \end{cases}$$

$$= \begin{cases} 1, & \frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n} \geq t_{n-1,\alpha} \\ 0, & \frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n} < t_{n-1,\alpha} \end{cases}$$

This is the one-sided t-test (Ex 4.12) for

$$H_0 : \mu \le \mu_0 \text{ } vs \text{ } H_1 : \mu > \mu_0$$

- *For the two-sided confidence interval for μ:*

$$(\bar{X} - \frac{t_{n-1,\alpha} \cdot S}{\sqrt{n}}, \bar{X} + \frac{t_{n-1,\alpha} \cdot S}{\sqrt{n}})$$

we can derive the associated family of tests. For any $\mu_0 \in \mathbb{R}$,

$$\psi_{\mu_0} = \begin{cases} 1, & \mu \notin (\bar{x} - \frac{t_{n-1,\alpha/2} \cdot S}{\sqrt{n}}, \bar{x} + \frac{t_{n-1,\alpha/2} \cdot S}{\sqrt{n}}) \\ 0, & \mu \in (\bar{x} - \frac{t_{n-1,\alpha/2} \cdot S}{\sqrt{n}}, \bar{x} + \frac{t_{n-1,\alpha/2} \cdot S}{\sqrt{n}}) \end{cases}$$
$$= \begin{cases} 1, & \sqrt{n} \left| \frac{\bar{x} - \mu_0}{s} \right| \ge t_{n-1,\frac{\alpha}{2}} \\ 0, & \sqrt{n} \left| \frac{\bar{x} - \mu_0}{s} \right| < t_{n-1,\frac{\alpha}{2}} \end{cases}$$

This is the two-sided t test for

$$H_0 : \mu = \mu_0 \text{ } vs \text{ } H_1 : \mu \neq \mu_0.$$

Theorem 5.9. *Consider a confidence set C and the corresponding family of tests $\{\psi_{\theta_0}, \theta_0 \in \Theta\}$ as specified in Theorem 5.7. Let also, for any $\theta \in \Theta$, k_θ be the set of undesirable parameters. For each $\theta_0 \in \Theta$, let*

$$\Theta_1^{\theta_0} = \{\theta \in \Theta : \theta_0 \in k_\theta\}$$

Then the following holds:

(1) C has confidence level $1 - \alpha$ if and only if $\forall (\theta_0, \xi) \in \mathcal{L}$:

$$E_{(\theta_0,\xi)} \psi_{\theta_0}(\underset{\sim}{X}) \le \alpha$$

*(2) C is an unbiased level-$(1 - \alpha)$ confidence set for θ if and only if, for each $\theta_0 \in \Theta$, ψ_{θ_0} is an **unbiased** level-α test of*

$$H_0 : \theta = \theta_0 \text{ } vs \text{ } H_1 : \theta \in \Theta_1^{\theta_0}$$

Note that Theorem 5.9 only guarantees the null hypothesis that $\theta = \theta_0$. ***unbiased*** *means type 2 error $\leq 1 - \alpha$.*

Example 5.10. *From 5.6, we know that if $k_\mu = (-\infty, \mu)$, then*

$$(\bar{X} - \frac{t_{n-1,\alpha} \cdot S}{\sqrt{n}}, \infty)$$

is an unbiased level-$(1-\alpha)$ CI for μ. For $\mu_0 \in \mathbb{R}$:

$$\{\mu \in \mathbb{R} : \mu_0 \in (-\infty, \mu)\} = (\mu_0, \infty).$$

Hence, from Theorem 5.9, the one-sided t-test

$$\psi_{\mu_0} = \begin{cases} 1, & \sqrt{n}\frac{\bar{X}-\mu_0}{S} \geq t_{n-1,\alpha} \\ 0, & \sqrt{n}\frac{\bar{X}-\mu_0}{S} < t_{n-1,\alpha} \end{cases}$$

is unbiased, level-α test for

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

- $K_\mu = \{\mu\}^c \rightsquigarrow$ two-sided CI for μ.

$$\left(\bar{X} - \frac{t_{n-1,\frac{\alpha}{2}} \cdot S}{\sqrt{n}}, \bar{X} + \frac{t_{n-1,\frac{\alpha}{2}} \cdot S}{\sqrt{n}}\right)$$

(unbiased, level-$(1-\alpha)$)

$$\{\mu \in \mathbb{R} : \mu_0 \in \{\mu\}^c\} = \{\mu \in \mathbb{R} : \mu \neq \mu_0\}$$

Two-sided t-test is an unbiased, level-α test for $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$.

Example 5.11 (Constructing CS from tests)**.** *$X_1, \cdots, X_n$ random sample from $N(\mu, \sigma^2)$, μ nuisance; our goal is to construct confidence sets for σ^2. Recall chi-square test*

Remark 4.15 :

$H_0 : \sigma^2 \leq \sigma_0^2$ $\quad$ ($\geq$, $=$) $\quad$ vs. $\quad H_1 : \sigma^2 > \sigma_0^2$ $\quad$ ($<$, $\neq$)

$$\Psi_{\sigma_0^2}(\underset{\sim}{x}) = 1\left(\frac{(n-1)s^2}{\sigma_0^2} \begin{array}{l} \geq \chi^2_{n-1,\alpha} \\ \leq \chi^2_{n-1,1-\alpha} \end{array} \right)$$

$\in (0, \chi^2_{n-1,1-\alpha/2}] \cup [\chi^2_{n-1,\alpha/2}, \infty)$

- $$k_{\sigma^2} = (0, \sigma^2) \to H_1 : \sigma_0^2 < \sigma^2$$
  $$C(\underset{\sim}{x}) = \left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha}}, \infty \right)$$

- $$k_{\sigma^2} = \{\sigma^2\}^C \to H_1 : \sigma_0^2 \neq \sigma^2$$
  $$C(\underset{\sim}{x}) = \left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \right)$$

Remark 5.12. *The correspondence between the tests and CS can also be used to develop uniformly most accurate CSs (these correspond to UMP classes of tests.)*

5.3 Interpretation of Confidence Sets

Example 5.13. *Generate a sample of size $n = 10$ from $N(1, 2)$. Suppose for this sample, we observed*

$$\bar{x} = 1.1, \quad s^2 = 1.5$$

two-sided CI for μ at CL 95%:

$$\left(\bar{x} - \boxed{\frac{\underbrace{t_{9,\,0.025}}_{2.262}\cdot\sqrt{1.5}}{\sqrt{10}}},\ \bar{x} + \boxed{\ -\ \text{''}\ -\ }\right)$$

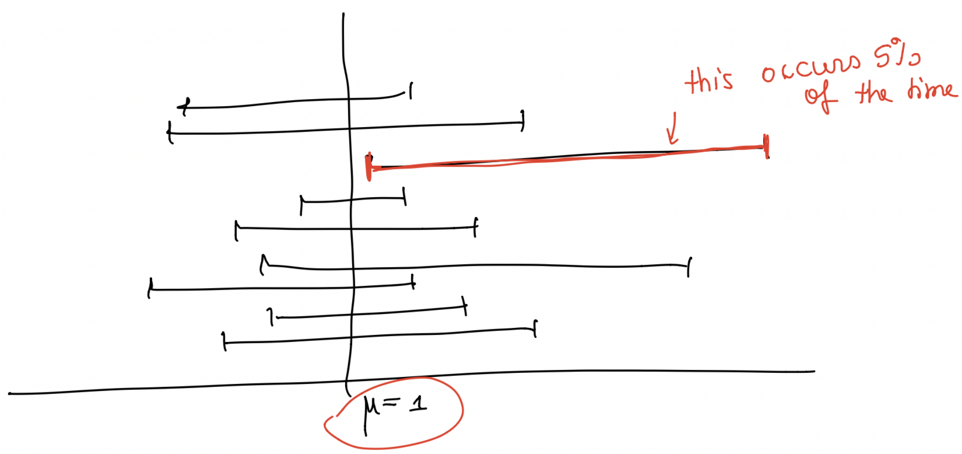
- $\Rightarrow (0.224,\ 1.976)$.

Test: $\mu = 1$ vs. $\mu \neq 1$.

Since $1 \in (0.224,\ 1.976) \Rightarrow$ do not reject at the 5% level

- Interpreting $\boxed{(0.224,\ 1.976)}$?

" ~~This is the interval in which the true μ lies with probability 95%~~ " No!



- set of "plausible values of μ".