MATH 357 - Honours Statistics

Aram-Alexandre Pooladian
Department of Mathematics
McGill University
Montreal, QC, Canada

September 22, 2018

Contents

Chapter 1

Odds and Ends

1.1 Multinomial Distribution

Let $\vec{X} = (X_1, \ldots, X_k)$ be a vector of random variables and $x_1, \ldots, x_k$ be non-negative integers such that $\sum_{i=1}^k x_i = n$. We say that $X_1, \ldots, X_k$ have a **multinomial distribution** with associated parameters $n, p_1, \ldots, p_k$ if

$$\mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k] = \frac{n!}{x_1!x_2! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

where $p_i \in (0,1)$ and $\sum_{i=1}^k p_i = 1$

Note

1. Once $p_1, \ldots, p_{k-1}$ are specified, we have that $p_k = 1 - \sum_{i=1}^{k-1} p_i$
2. Notation: $\frac{n!}{x_1!x_2! \cdots x_k!} = \binom{n}{x_1 x_2 \ldots x_k}$ is called the *multinomial coefficient*; it's the number of ways we can partition n distinct objects into subsets of size $x_1, x_2, \ldots, x_k$
3. The random variables $X_1, \ldots, X_k$ are **not** independent since $\sum_{i=1}^k x_i = n$
4. Multinomial Distribution can become the Binomial Distribution if $k = 2, x_1 = x, x_2 = n - x$; $\mathbb{P}[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$

The Multinomial Setup

1. There are n independent trials
2. Each trial can result in exactly one of the k possible outcomes
3. the probability of outcome i is p_i for all of the n trials (

Let $X_1, \ldots, X_k$ represent the number of Type 1, …, Type k outcomes that are observed after n trials. Then the distribution of $\vec{X}$ is multinomial with parameters n and $p_1, \ldots, p_k$

Proof. Start with any configuration with $x_1, \ldots, x_k$ of Type 1, …, Type k. This has probability $p_1^{x_1} \cdots p_k^{x_k}$, by the three assumptions above. Then, we sum over all possible configuations; thus we get the multinomial coefficient too. □

1.2 Order Statistics

Definition 1.2.1. Let $X_1, \dots, X_n$ be n random variables. Call $Y_1 \leq Y_2 \leq \cdots \leq Y_n$ the **order statistics** of $X_1, \dots, X_n$ if

$$\begin{aligned} Y_1 &= \min\{X_1, \dots, X_n\} \\ Y_2 &= 2^{\text{nd}} \min\{X_1, \dots, X_n\} \\ &\cdots \\ Y_n &= \max\{X_1, \dots, X_n\} \end{aligned}$$

where we have that the Y_i are also random but not independent. Our main interest falls under the setting where $X_1, \dots, X_n$ are i.i.d continuous random variables with pdf $f_X(x)$

Theorem 1.2.1. *Let $f_{Y_1,\dots,Y_n}(y_1, \dots, y_n)$ be the joint pdf of $Y_1, \dots, Y_n$, then*

$$f_{Y_1,\dots,Y_n}(y_1, \dots, y_n) = n! \prod_{i=1}^{n} f_X(y_i) \qquad \textit{for } y_1 < y_2 < \cdots < y_n$$

and 0 otherwise

Proof. First, recall the monotone transformation theorem for a random vector. If we wish to find the distribution of $\vec{W} = g(\vec{V})$, where the domain of $\vec{V}$ may be partitioned into sets $A_1, \dots, A_k$ such that $g(\cdot)$ is monotone on each A_j. Let $g_j^{-1}(\cdot)$ be the inverse of g on A_j, then:

$$f_{\vec{W}}(\vec{w}) = \sum_{j=1}^{k} f_{\vec{v}}(g_j^{-1}(\vec{w})) \left| \frac{d}{d\vec{w}} g_j^{-1}(\vec{w}) \right|$$

In the context of this question, we see that the different possible orderings (permutations) of the Xs form the disjoint sets A_j, so

$$f_{Y_1,\dots,Y_n}(y_1, \dots, y_n) = \sum_{\sigma_n} f_{X_1,\dots,X_n}(g^{-1}(\vec{y})) \left| \frac{d}{d\vec{y}} g^{-1}(\vec{y}) \right|$$

where σ_n denotes all possible orderings of $X_1, \dots, X_n$. To simplify this expression, we examine a single permutation, say $X_1 < X_2 < \cdots < X_n$. In this case, we have that $y_1 = x_1, \dots, y_n = x_n$ so

$$\begin{aligned} f_{X_1,\dots,X_n}(g^{-1}(\vec{y})) &= f_{X_1}(y_1) \times f_{X_2}(y_2) \times \cdots f_{X_n}(y_n) \\ &= f_X(y_1) \times f_X(y_2) \times \cdots f_X(y_n) = \prod_{i=1}^{n} f_X(y_i) \end{aligned}$$

where the last line follows because the X_i's are identically distributed with common pdf $f_X(x)$. In this case, we have that the Jacobian is the determinant of the identity matrix, which is 1. Now we consider any other permutation. Since the random variables are identically distributed, we see that $f_{X_1,\dots,X_n}(g^{-1}(\vec{y}))$ is the same for all other permutations and the jacobian will also always be 1 (since sometimes the rows of the matrix of derivatives will be transposed but we're taking the absolute value so this worry goes away). Since there are $n!$ possible permutations, we have that

$$f_{Y_1,\dots,Y_n}(y_1, \dots, y_n) = n! \prod_{i=1}^{n} f_X(y_i)$$

□

Note

To find the joint density of only some of the order statistics, a powerful trick is to create a multinomial setup and use it to our advantage

Example 1.2.1. Let $X_1, \dots, X_4$ be i.i.d. with pdf f_X; find the joint pdf of Y_1 and Y_4, that is $f_{Y_1,Y_4}(y_1, y_4)$

$\Rightarrow$ As stated above, the idea is to use the multinomial setup to get the right pattern to efficiently solve the problem. Of the n order statistics, we're only looking at a joint density of $k = 2 < n$ of them so we need $2k + 1$ "boxes". $B_0 = (-\infty, y_1)$, $B_1 = [y_1, y_1 + dy_1)$, $B_2 = [y_1 + dy_1, y_4)$, $B_3 = [y_4, y_4 + dy_4)$, $B_4 = [y_4 + dy_4, \infty)$

We have that $B_0 = \emptyset$, $X_1 \in B_1$, $X_2, X_3 \in B_2$, $X_4 \in B_3$ and $B_4 = \emptyset$ so we have that

$$\begin{aligned}\mathbb{P}[Y_1 \in B_1, Y_4 \in B_3] &= \mathbb{P}[0 \in B_0, 1 \in B_1, 2 \in B_2, 1 \in B_3, 0 \in B_4] \simeq f_{Y_1,Y_4}(y_1, y_4)dy_1dy_4 \\ &= \frac{4!}{0!1!2!1!0!}[F_X(y_1)]^0[f_X(y_1)dy_1]^1[F_X(y_4) - F_X(y_1)]^2[f_X(y_4)dy_4]^1[1 - F_X(y_4)]^0 \\ &= \frac{4!}{2!}f_X(y_1)f_X(y_4)[F_X(y_4) - F_X(y_1)]^2 \quad \text{for } y_1 < y_4\end{aligned}$$

and 0 otherwise.

1.3 Sampling Distributions

Theorem 1.3.1. *Let $X_1, \dots, X_n$ be $\overset{iid}{\sim} N(\mu, \sigma^2)$, then*

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

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

Proof. Use the following properties of mgfs: $M_{AX+B}(t) = e^{Bt}M_X(At)$ and $M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$ and the uniqueness of mgfs □

Theorem 1.3.2. ***(Central Limit Theorem)*** *Let $X_1, \dots, X_n$ be i.i.d random variables with mean μ and variance σ^2, then*

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z \sim N(0, 1)$$

Note in the CLT, the convergence (in distribution) of the fraction that goes to Z is uniform in x, i.e

$$\mathbb{P}\left[\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq x\right] \to \mathbb{P}[Z \leq x]$$

uniformly on $x \in (-\infty, \infty)$

Theorem 1.3.3. ***(Slutsky's Theorem)*** *If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{p,D} c$ then $X_nY_n \xrightarrow{D} cX$*

Lemma 1.3.3.1.

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

Proof.

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} - \overline{X}^2) \\ &= \sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i - n\overline{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - 2\overline{X} \cdot n\overline{X} - n\overline{X}^2 \\ &= \left(\sum_{i=1}^{n} X_i^2\right) - n\overline{X}^2\end{aligned}$$

□

Theorem 1.3.4. *Suppose $X_1, \ldots, X_n$ are i.i.d $N(\mu, \sigma^2)$ (even works without normality, just need finite variance and mean). Let*

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

then $s^2 \xrightarrow{p} \sigma^2$ as $n \to \infty$

Proof.

$$\begin{aligned}s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2\right) - \frac{n}{n-1}\overline{X}^2 \\ &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \frac{n}{n-1}\overline{X}^2 \xrightarrow{P} \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sigma^2\end{aligned}$$

since have that $\mathbb{E}[X_i^2] < \infty$ so by WLLN,

$$\frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n} X_i^2 \to \mathbb{E}[X^2]$$

and, by Slutsky's Theorem and WLLN,

$$\frac{n}{n-1}\overline{X}^2 \to \mathbb{E}[X]^2$$

□

Theorem 1.3.5. *If $X_1, \ldots, X_n$ are i.i.d with finite mean and variance, then*

$$\frac{\overline{X} - \mu}{s/\sqrt{n}} \xrightarrow{D} Z \sim N(0, 1)$$

Proof.

$$\frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{s} \xrightarrow{D} Z \cdot 1$$

via Slutsky's Theorem and by CLT □

Theorem 1.3.6. *Let $X_1, \ldots, X_n$ be i.i.d normally distributed with mean μ and variance σ^2. Then $\overline{X}$ is independent of s^2; the sample mean is independent of the sample variance*

Proof. Omitted □

Theorem 1.3.7. *(From Assignment 1) Let $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$*

Theorem 1.3.8. *(From Assignment 1) If $X_1, X_2, \ldots, X_n$ are independent $\chi^2_{\alpha_i}$ random variables then*

$$\sum_{i=1}^{n} X_i \sim \chi^2_\alpha$$

where $\alpha = \sum_{i=1}^n \alpha_i$

Theorem 1.3.9. *(From Assignment 1) Let $X \sim N(0,1)$ be independent of $Y \sim \chi^2_\nu$, then*

$$T = \frac{X}{\sqrt{Y/\nu}} \sim t_\nu$$

Theorem 1.3.10. *(From Assignment 1) Let $W \sim \chi^2_{\nu_1}$ be independent of $Z \sim \chi^2_{\nu_2}$ then*

$$U = \frac{W/\nu_1}{Z/\nu_2} \sim F_{\nu_1,\nu_2}$$

Theorem 1.3.11. *Let $X_1, \ldots, X_n$ be i.i.d normally distributed with mean μ and variance σ^2. Then*

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

Proof. Start with

$$\begin{aligned}
\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X} + \overline{X} - \mu)^2 \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + 2\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(\overline{X}-\mu\right) + \sum_{i=1}^{n}\left(\overline{X}-\mu\right)^2\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + 2(\overline{X}-\mu)\underbrace{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)}_{=0} + \sum_{i=1}^{n}\left(\overline{X}-\mu\right)^2\right] \\
&= \frac{n-1}{\sigma^2}\underbrace{\left[\frac{1}{(n-1)}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right]}_{s^2} + \frac{n}{\sigma^2}\left(\overline{X}-\mu\right)^2 \\
\Rightarrow \underbrace{\sum_{i=1}^{n}\underbrace{\left(\overbrace{\frac{X_i-\mu}{\sigma}}^{\sim N(0,1)}\right)^2}_{\sim\chi^2_1}}_{A\sim\chi^2_n} &= \frac{n-1}{\sigma^2}s^2 + \underbrace{\left(\overbrace{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}^{\sim N(0,1)}\right)^2}_{C\sim\chi^2_1}
\end{aligned}$$

then using the mgfs of each of the terms, we have that

$$\frac{1}{(1-2t)^{n/2}} = M_{B+C}(t) \overset{\star}{=} M_B(t)\cdot M_C(t) = M_B(t)\cdot\frac{1}{(1-2t)^{1/2}} \Rightarrow M_B(t) = \frac{1}{(1-2t)^{(n-1)/2}}$$

where $\star$ holds because $\overline{X}$ and s^2 are independent; by uniqueness of mgfs, we're done! □

Theorem 1.3.12. *Let $X_1, \ldots, X_n$ be iid normal random variables with mean μ and variance σ^2; then*

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

Proof. We use the formulation of the t-distribution from the assignment and mess around with the given inequality;

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} = \underbrace{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}_{\sim N(0,1)} \Big/ \sqrt{\frac{(n-1)s^2}{\sigma^2}\Big/(n-1)}$$

and since $\overline{X}$ and s^2 are independent, we have that the normal distribution in the numerator and the χ^2 variable in the denominator are independent (normality is very crucial here). Thus we have the right setup for a random variable that is distributed as t_{n-1} □

Notes about Student's t Distribution

- Symmetric about 0, looks like the standard normal distribution except with "heavier tails" that don't approach zero as quickly
- A t_ν distribution has $\nu - 1$ moments; i.e. t_1 does not have a first moment (a mean)
- $t_\nu \to N(0,1)$ as $\nu \to \infty$

Theorem 1.3.13. *Let $X_1, \ldots, X_{n_1}$ be i.i.d $N(\mu_1, \sigma^2)$ random variables and $Y_1, \ldots, Y_{n_2}$ be i.i.d $N(\mu_2, \sigma^2)$ random variables (where $\sigma_1^2 = \sigma_2^2 = \sigma^2$). Let $\vec{X}$ and $\vec{Y}$ be independent and let*

$$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(X_i - \overline{X}\right)^2 \quad \textit{and} \quad s_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}\left(Y_i - \overline{Y}\right)^2$$

then $(s_1^2/s_2^2) \sim F_{n_1-1,n_2-1}$

Proof. We have that

$$\frac{s_1^2}{s_2^2} = \left(\underbrace{\frac{(n_1-1)s_1^2}{\sigma^2}}_{\sim\chi^2_{n_1-1}}\frac{1}{(n_1-1)}\right)\Big/\left(\underbrace{\frac{(n_2-1)s_2^2}{\sigma^2}}_{\chi^2_{n_2-1}}\frac{1}{(n_2-1)}\right)$$

and since $\vec{X}$ is independent of $\vec{Y}$, this implies independence of $\frac{(n_1-1)s_1^2}{\sigma^2}$ and $\frac{(n_2-1)s_2^2}{\sigma^2}$; thus the two χ^2 random variables are independent thus the formula from the assignment holds and we have the result. □

Example 1.3.1. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli random variables and let $X = \sum_{i=1}^n X_i$. It is easy to show using MGFs that $X \sim \text{Binomial}(n,p)$ and we have that

$$\frac{X - np}{\sqrt{np(1-p)}} \xrightarrow{D} Z \sim N(0,1)$$

as $n \to \infty$ via CLT. Equivalently, this can be written as

$$\frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\frac{X}{n} - p}{\sqrt{[(\frac{X}{n})(1-\frac{X}{n})]n}} \xrightarrow{D} Z \sim N(0,1)$$

Continuity Corrections

Gamma Distribution

If $X \sim \text{Gamma}(\alpha, \beta)$ then

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta} \quad \text{for } x \geq 0$$

and 0 otherwise. $\mathbb{E}[X] = \alpha\beta$ and $\text{Var}[X] = \alpha\beta^2$. Note the following important relationships

1. If $\alpha = 1$, then $X \sim \text{Gamma}(1, \beta) = \exp[\beta]$
2. If $\alpha = \nu/2$ and $\beta = 2$, then $X \sim \text{Gamma}(\nu/2, 2) = \chi^2_\nu$

Chapter 2

Statistical Inference

2.1 Estimation - Formally

We have a distribution specified by a cdf $F_{\vec{X}}(\vec{x})$ where $\vec{X} = (X_1, \dots, X_n)$ is a random vector. There are two main types of estimation problems. We draw a sample $\vec{X}_1, \dots, \vec{X}_n$ (suppose i.i.d) and we must use it to guess at $F_{\vec{X}}$

1. The form of $F_{\vec{X}}$ is left unspecified; in this case we need to estimate (or guess at) $F_{\vec{X}}$ for every $\vec{x}$. If the set of possible values of $\vec{x}$ is *infinite*, then we have what is called a **distribution free** or **non-parametric estimation problem**

2. We assume a form for $F_{\vec{X}}$ i.e. we suppose that $F_{\vec{X}}$ comes from some specified family of distributions. What is unknown is one or more parameters that specify exactly which member of the family describes our distribution. In other words, we have $F_{\vec{X}}(\vec{x}, \theta)$, where θ is unknown (could be a vector of parameters). This is called a **parametric estimation problem**

Definition 2.1.1. A **statistic** $T = T(\vec{X})$ is a(ny) measurable function of the set of observations, $\vec{X}$, as long as T is not a function of any of the unknown parameters. Examples: $T(\vec{X}) = \overline{X}, = 3.7$,etc

Definition 2.1.2. If a statistic T is used to "estimate" an unknown parameter θ, then we call T an **estimator** or θ and we denote it by $\hat{\theta} = \hat{\theta}(\vec{X})$

Notes

1. An estimator $\hat{\theta} = \hat{\theta}(\vec{X})$ is a random variable, i.e. before we carry out our experiment, we will not know the value of $\hat{\theta}$ we will get after we observed our data and obtained the realized values of $\vec{X}$, denoted by $\vec{x} = (x_1, \dots, x_n)$. To this end, we call $\hat{\theta}(\vec{x})$ an **estimate**

2. Since the pre-experiment $\hat{\theta}$ is a random variable, it has a distribution called a **sampling distribution**, denoted $F_{\hat{\theta}}$

3. **The Big Idea:** Statisticians spend a lot of time devising estimators that have "good" probabilistic properties. For example, we suspect that $\overline{X}$ is better than X_1 for estimating the unknown mean μ (we shall see why this is the case). Likewise, we suspect that s^2 should always be used to estimate an unknown variance σ^2 if μ is also unknown; this is NOT always the case.

4. In order to devise "good" estimators, we need to devise properties of estimators that are desirable. Once we have defined an estimator $\hat{\theta}$ with these good properties, we can only justify its use because, on repeated uses, it will "do well" according to certain specifications. Good properties are properties of the *procedure*; we cannot say how well we have done AFTER substituting our data

2.2 Desirable Properties for Estimators

2.2.1 Unbiasedness

Definition 2.2.1. Let θ be an unknown parameter. $\hat{\theta}$ is said to be an **unbiased estimator** of θ if $\mathbb{E}[\hat{\theta}] \overset{\theta}{\equiv} \theta$ i.e. $\mathbb{E}_\theta[\hat{\theta}] = \theta$ for all $\theta \in \Theta$, where Θ is the *parameter space* of θ

Definition 2.2.2. The **bias** of $\hat{\theta}$ is $b(\theta) = \mathbb{E}[\hat{\theta}] - \theta$; we have that $\hat{\theta}$ is unbiased if $b(\theta) = 0$ for all $\theta \in \Theta$

Note

- It is absolutely necessary that $\mathbb{E}_\theta(\hat{\theta}) = \theta$ for **ALL** θ; prevents us from saying that $\hat{\mu} = 3.2$ is unbiased estimator since it is only unbiased for a very specific incident
- Unbiasedness if a family property or a model property, because the expected value is taken with respect to a specific family

Theorem 2.2.1. *(Not sure if theorem) $\overline{X}$ is always unbiased for μ (the mean) regardless of the family of distribution and holds even without independence*

Proof.

$$\mathbb{E}[\overline{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n X_i\right] = \frac{1}{n}n\mu \overset{\mu}{\equiv} \mu$$

□

Theorem 2.2.2. *Let $X_1, \ldots, X_n$ be i.i.d random variables with finite variance. Then s^2 as previously defined is unbiased for σ^2 (again, regardless of the distribution)*

Proof.

$$\begin{aligned}
\mathbb{E}[s^2] &= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^n \left(X_i - \overline{X}\right)^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^n (X_i - \mu + \mu - \overline{X})^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^n \left((X_i-\mu)^2 + 2(\mu - \overline{X})(X_i - \mu) + (\mu - \overline{X})^2\right)\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\left(\sum_{i=1}^n (X_i-\mu)^2\right) + n(\mu - \overline{X})^2 + 2(\mu - \overline{X})\sum_{i=1}^n (X_i - \mu)\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\left(\sum_{i=1}^n (X_i-\mu)^2\right) - n(\mu - \overline{X})^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\left(\sum_{i=1}^n (X_i-\mu)^2\right)\right] - \frac{n}{n-1}\underbrace{\mathbb{E}\left[\left(\overline{X}-\mu\right)^2\right]}_{=\text{Var}(\overline{X})=\sigma^2/n} \\
&= \frac{1}{n-1}\sum_{i=1}^n \underbrace{\mathbb{E}\left[\left(X_i-\mu\right)^2\right]}_{=\text{Var}(X_i)=\sigma^2} - \frac{n}{n-1}\frac{\sigma^2}{n} \\
&= \frac{1}{n-1}\left(n\sigma^2 - \sigma^2\right) = \sigma^2
\end{aligned}$$

□

Definition 2.2.3. Let $X_1, \ldots, X_n$ be random variables with distribution that depend on some unknown parameter θ. We call $\hat{\theta}$ a **uniform minimum variance unbiased estimator** or **UMVUE** for θ if

1. $\mathbb{E}[\hat{\theta}] \overset{\theta}{\equiv} \theta$ i.e. unbiased
2. If $\hat{\theta}^\star$ is any other unbiased estimator for θ, then $\text{Var}(\hat{\theta}) \leq \text{Var}(\hat{\theta}^\star)$ for all $\theta \in \Theta$

Theorem 2.2.3. ***(Cramer Rao Inequality)*** *Let $(X_1, \ldots, X_n) = \vec{X} \sim f_{\vec{X}}(\vec{x}, \theta)$ (or $p_{\vec{X}}(\vec{x}, \theta)$, in the discrete case). Let $g(\theta)$ be some function of θ; let $T(\vec{X})$ be an unbiased estimator of $g(\theta)$ i.e. $\mathbb{E}[T(\vec{X})] \overset{\theta}{\equiv} g(\theta)$. Then, under certain regularity conditions (existence of derivatives, ability to interchange derivatives and integrals, and the range of $f_{\vec{X}}(\vec{x}, \theta)$ cannot depend on θ)*

$$Var(T(\vec{X})) \geq \frac{[g'(\theta)]^2}{\mathbb{E}\Big[\big(\frac{\partial}{\partial\theta} \ln[f_{\vec{X}}(\vec{x}, \theta)]\big)^2\Big]}$$

Proof.

Lemma 2.2.3.1. $\mathbb{E}[\frac{\partial}{\partial\theta} \ln[f_{\vec{X}}(\vec{x}, \theta)]] = 0$

Proof.

$$\begin{aligned}
\mathbb{E}[\frac{\partial}{\partial\theta} \ln[f_{\vec{X}}(\vec{x}, \theta)]] &= \int \Big(\frac{\partial}{\partial\theta} \ln[f_{\vec{X}}(\vec{x}, \theta)]\Big) f_{\vec{X}}(\vec{x}, \theta) d\vec{x} \\
&= \int \frac{f'_{\vec{X}}(\vec{x}, \theta)}{f_{\vec{X}}(\vec{x}, \theta)} f_{\vec{X}}(\vec{x}, \theta) d\vec{x} \quad \text{by chain rule} \\
&= \int f'_{\vec{X}}(\vec{x}, \theta) d\vec{x} = \frac{\partial}{\partial\theta} \int f_{\vec{X}}(\vec{x}, \theta) d\vec{x} \quad \text{regularity conditions} \\
&= \frac{\partial}{\partial\theta}(1) = 0
\end{aligned}$$

□

$$\begin{aligned}
g'(\theta) &= \frac{\partial}{\partial\theta}\mathbb{E}[T(\vec{X})] = \frac{\partial}{\partial\theta} \int T(\vec{x}) f_{\vec{X}}(\vec{x}, \theta) d\vec{x} = \int T(\vec{X}) \frac{\partial}{\partial\theta} f_{\vec{X}}(\vec{x}, \theta) d\vec{x} \quad \text{regularity conditions} \\
&= \int T(\vec{x}) \frac{\partial}{\partial\theta}\big(\ln[f_{\vec{X}}(\vec{x}, \theta)]\big) f_{\vec{X}}(\vec{x}, \theta) d\vec{x} \quad \text{chain rule} \\
&= \int \underbrace{(T(\vec{x}) - g(\theta))}_{X} \underbrace{\frac{\partial}{\partial\theta}\big(\ln[f_{\vec{X}}(\vec{x}, \theta)]\big)}_{Y} f_{\vec{X}}(\vec{x}, \theta) d\vec{x} \quad \text{previous lemma} \\
&\Rightarrow g'(\theta) = \mathbb{E}[XY]
\end{aligned}$$

and recall the **Cauchy-Schwarz Inequality**; $\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$

$$\begin{aligned}
[g'(\theta)]^2 &\leq \int [T(\vec{x}) - g(\theta)]^2 f_{\vec{X}}(\vec{x}, \theta) d\vec{x} \cdot \int \Big[\frac{\partial}{\partial\theta}\big(\ln[f_{\vec{X}}(\vec{x}, \theta)]\big)\Big]^2 f_{\vec{X}}(\vec{x}, \theta) d\vec{x} \\
&= \text{Var}[T(\vec{X})] \cdot \mathbb{E}\Big[\big(\frac{\partial}{\partial\theta} \ln[f_{\vec{X}}(\vec{x}, \theta)]\big)^2\Big]
\end{aligned}$$

□

Notes on CR

1. The lower bound is called the **Cramer Rao Lower Bound**

2. Suppose we have an unbiased estimator for $g(\theta)$; if $\mathrm{Var}(T)$ attains the Cramer-Rao Lower Bound, then we know that T is UMVUE

3. It is possible to have a UMVUE although the variance does not equal the lower bound

Theorem 2.2.4. *If $X_1, \dots, X_n$ are i.i.d. from the same distribution, then the denominator of the Cramer-Rao Lower Bound can be written as follows*

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln[f_{\vec{X}}(\vec{x},\theta)]\right)^2\right] = n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln[f_X(x,\theta)]\right)^2\right]$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln[f_{\vec{X}}(\vec{x},\theta)]\right)^2\right] &= \mathrm{Var}\left[\frac{\partial}{\partial\theta}\ln[f_{\vec{X}}(\vec{x},\theta)]\right] = \mathrm{Var}\left[\frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f_{X_i}(x_i,\theta)\right] \quad \text{by independence}\\ &= \mathrm{Var}\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln[f_{X_i}(x_i,\theta)]\right] = \sum_{i=1}^{n}\mathrm{Var}\left[\frac{\partial}{\partial\theta}\ln[f_{X_i}(x_i,\theta)]\right]\\ &= n\mathrm{Var}\left[\frac{\partial}{\partial\theta}\ln[f_{X_i}(x_i,\theta)]\right] = n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln[f_X(x,\theta)]\right)^2\right]\end{aligned}$$

□

Theorem 2.2.5. *Under the "usual" regularity conditions, the denominator of the CR lower bound can be written as*

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln[f_{\vec{X}}(\vec{x},\theta)]\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln[f_X(x,\theta)]\right]$$

Proof.

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln[f_X(x,\theta)]\right] &= -\int\left[\frac{\partial^2}{\partial\theta^2}\ln[f_X(x,\theta)]\right]f_{\vec{X}}(\vec{x},\theta)d\vec{x} = -\int\frac{\partial}{\partial\theta}\frac{f'_{\vec{X}}(\vec{x},\theta)}{f_{\vec{X}}(\vec{x},\theta)}f_{\vec{X}}(\vec{x},\theta)d\vec{x}\\ &= -\int\frac{f''_{\vec{X}}(\vec{x},\theta)f_{\vec{X}}(\vec{x},\theta)}{f^2_{\vec{X}}(\vec{x},\theta)}f_{\vec{X}}(\vec{x},\theta)d\vec{x} + \int\frac{[f'_{\vec{X}}(\vec{x},\theta)]^2 f_{\vec{X}}(\vec{x},\theta)}{f^2_{\vec{X}}(\vec{x},\theta)}d\vec{x}\\ &= -\int f''_{\vec{X}}(\vec{x},\theta)d\vec{x} + \int\left[\frac{f'_{\vec{X}}(\vec{x},\theta)}{f_{\vec{X}}(\vec{x},\theta)}\right]^2 f_{\vec{X}}(\vec{x},\theta)d\vec{x}\\ &= 0 + \int\left[\frac{\partial}{\partial\theta}\ln[f_{\vec{X}}(\vec{x},\theta)]\right]^2 f_{\vec{X}}(\vec{x},\theta)d\vec{x} = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln[f_{\vec{X}}(\vec{x},\theta)]\right)^2\right]\end{aligned}$$

□

2.2.2 Mean Squared Error

Most users of statistics prefer to use estimators that have tolerably small variances and biases that are not too large.

Definition 2.2.4. The **Mean Square Error** of an estimator, $\hat{\theta}$, of a parameter θ, is given by $\mathrm{MSE}(\hat{\theta}) := \mathbb{E}[(\hat{\theta}-\theta)^2]$

Theorem 2.2.6.

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + (b(\hat{\theta}))^2$$

Proof.

$$\begin{aligned}\mathbb{E}[(\hat{\theta}-\theta)^2] &= \mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}]+\mathbb{E}[\hat{\theta}]-\theta)^2] = \mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2+2(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta)-(\mathbb{E}[\hat{\theta}]-\theta)^2] \\ &= \mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2]+\mathbb{E}[(\mathbb{E}[\hat{\theta}]-\theta)^2]+2(\mathbb{E}[\hat{\theta}]-\theta)\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])] \\ &= \mathrm{Var}[\hat{\theta}]+(b(\hat{\theta}))^2+2(\mathbb{E}[\hat{\theta}]-\theta)(\mathbb{E}[\hat{\theta}]-\mathbb{E}[\hat{\theta}]) = \mathrm{Var}[\hat{\theta}]+(b(\hat{\theta}))^2\end{aligned}$$

□

Example 2.2.1. Let $X_1,\ldots,X_n$ be i.i.d normally distributed random variables with mean μ and variance σ^2. Let

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

then $\mathrm{MSE}(\hat{\sigma}^2) < \mathrm{MSE}(s^2)$ for all $\sigma^2 > 0$ for all $n > 1$

$\Rightarrow$ We have that $\mathbb{E}[s^2] = \sigma^2$ so the bias is zero; then to calculate $\mathrm{Var}(s^2)$:

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

because $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$ has variance $2(n-1)$; thus the $\mathrm{MSE}(s^2) = 2\sigma^4/(n-1)$. For the MSE of the other estimator, we need to do a little bit more work;

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

then, for the variance

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

thus the $\mathrm{MSE}(\hat{\sigma}^2) = 2\sigma^4(n-1)n^{-2} + \sigma^4 n^{-}2 < \mathrm{MSE}(s^2)$

2.2.3 Consistency

Definition 2.2.5. Let $\hat{\theta} = \hat{\theta}(x_1,\ldots,x_n)$ be an estimator of the parameter θ. Then we say that $\hat{\theta}$ is **weakly consistent** for θ if $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$. We say that $\hat{\theta}$ is **strongly consistent** for θ if it converges almost surely

Notes

1. We can show that an estimator $\hat{\theta}$ is consistent by showing: $\mathbb{E}[\hat{\theta}] \to \theta$ (asymptotically unbiased) and $\mathrm{Var}[\hat{\theta}] \to 0$

   Proof.

   $$\mathbb{P}[|\hat{\theta}-\theta| > \delta] \leq \frac{MSE[\hat{\theta}]}{\delta^2} = \frac{\mathrm{Var}[\hat{\theta}]}{\delta^2} + \frac{(b(\hat{\theta}))^2}{\delta^2} \to 0$$

   □

2. By showing directly that $\hat{\theta} \xrightarrow{p} \theta$

3. By using the WLLN, $\overline{X} \xrightarrow{p} \mathbb{E}[X]$

Example 2.2.2. Let $X_1, \dots, X_n$ be i.i.d. with cdf F, then let

$$\hat{F}_n(x) = \begin{cases} \frac{k}{n} & \text{if } y_k \leq x < y_{k+1}, \\ 0 & \text{if } x < y_1, \\ 1 & \text{if } x \geq y_n \end{cases}$$

then it is easy to see that $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}[X_i \leq x]$; then we claim that for any x, $\hat{F}_n(x)$ is a consistent estimator for $F(x)$

Proof. We have that $\{\mathbb{1}[X_i \leq x]\}$ is a sequence of i.i.d Bernoulli random variables with

$$\text{“}p\text{”} = \mathbb{E}[\mathbb{1}[X_i \leq x]] = \mathbb{P}[X_i \leq x] = F(x)$$

So since $\hat{p} \xrightarrow{p} p$, for a set of Bernoulli random variables, then $\hat{F}_n(x) \xrightarrow{p} F(x)$ □

Remark Consistency is a limiting property; we can still have a terrible consistent estimator. For example, take $X_1, \dots, X_n$ be i.i.d. with unknown mean μ, then $\hat{\mu} = 0$ if $n < 10^{100}$ and $\overline{X}$ if $n \geq 10^{100}$; thus this is consistent

2.2.4 Sufficiency

Idea Let θ be an unknown parameter. We seek a statistic $T = T(\vec{X})$ such that the information contained about θ in T is the same as the information about θ contained in T is the same as the information about θ contained in $\vec{X}$. The *hope* is that the dimension of T will be less than that of $\vec{X}$, so that even though we have summarized $\vec{X}$ in some fashion, we have not lost any information through the summarization process. Often, it is easier to work with T than all of the observations.

Definition 2.2.6. The function of θ, $f_{\vec{X}}(\vec{x}, \theta)$, for fixed $\vec{x}$, is called the **likelihood function** of $\vec{x}$, denoted by $L(\vec{x}, \theta)$; thus each fixed $\vec{x}$ produces a different likelihood function. If $L(\vec{x}, \theta)$ does not change with θ, i.e. is constant for all θ

Definition 2.2.7. A statistic T is called **sufficient** for θ if $f_{\vec{X}|T(\vec{X})=t}(\vec{x}, \theta|t)$ is independent of θ for all $\vec{x}$, all t such that $f_T(t) \neq 0$, and for all $\theta \in \Theta$

Notes

1. θ could be vector valued, as could T
2. $T(\vec{X}) = \vec{X}$ is trivially a sufficient statistic for θ as $f_{\vec{X}|\vec{X}=\vec{x}}(\vec{x}; \theta|\vec{x}) = 1$ is independent of θ
3. We call the collection of sets $\{A_t : A_t = \{\vec{x}|T(\vec{x}) = t\}\forall t\}$ a **sufficient partition of the sample space** S if $f_{\vec{X}}(\vec{x}, \theta|\vec{x} \in A_t)$ is independent of θ for all t. Also, T is sufficient if it induces a sufficient partition of S
4. If T is sufficient, it may not provide a reasonable estimator for θ

Theorem 2.2.7. *If T is sufficient for θ and g is a 1-to-1 function of T (bijective), then $g(T)$ is sufficient for θ*

Proof.

$$\{\vec{x} : g(T(\vec{X})) = g(t)\} = \{\vec{x} : T(\vec{X}) = g^{-1(}g(t))\} = \{\vec{x} : T(\vec{X}) = t\}$$

□

Example 2.2.3. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli random variables with p unknown. We want to show that $T(\vec{X}) = X = \sum_{i=1}^n X_i$ is sufficient for p

$\Rightarrow$

$$p_{\vec{X}|T(\vec{X})=t}(\vec{x}|t) = \mathbb{P}[\vec{X} = \vec{x}, p|T(\vec{X}) = t]$$

and 0 if $t \neq T(\vec{X})$. Assume the former case, then

$$\begin{aligned}
p_{\vec{X}|T(\vec{X})=t}(\vec{x}|t) &= \mathbb{P}[\vec{X} = \vec{x}, p|T(\vec{X}) = t] \\
&= \frac{\mathbb{P}[\vec{X} = \vec{x}, p]}{\mathbb{P}[X = t]} \quad t = sum_{i=1}^n x_i \\
&= \frac{\mathbb{P}[X_1 = x_1, \ldots, X_n = x_n; p]}{\binom{n}{t} p^t (1-p)^{n-t}} \\
&= \frac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{1}{\binom{n}{t}}
\end{aligned}$$

which holds for all $\vec{x}$, thus $T(\vec{X}) = X$ is sufficient for p

Simulation Problem

Suppose $\vec{X} \sim f_{\vec{X}}(\vec{x}, \theta)$ with θ unknown. Suppose we have a sample $\vec{X}$ and you lose it. We hope to get $\vec{X}^\star \overset{D}{=} \vec{X}$. We cannot simulate $\vec{X}^\star$ from $f_{\vec{X}}(\vec{x}, \theta)$ because θ is unknown. However, even though you lost $\vec{X}$, if you retained $T(\vec{X})$, you can recover an $\vec{X}^\star \overset{D}{=} \vec{X}$, if T is sufficient for θ. Let $T = \vec{X}$ and $Y = \vec{X}^\star$, then have a ready-made T. Now consider $f_{\vec{X}|T=t}(\vec{x}, \theta|t)$ is actually independent of θ to extract a pair (x, y); thus we get $\vec{X}^\star$ from this

Theorem 2.2.8. ***(Neyman-Fisher Factorization)*** *Let $\vec{X} \sim f_{\vec{X}}(\vec{x}, \theta)$ where θ is unknown. Then $T = T(\vec{X})$ is sufficient for θ if we can write*

$$f_{\vec{X}}(\vec{x}, \theta) = g(T(\vec{X}), \theta) h_{\vec{X}}(\vec{x})$$

where g depends on θ and on $\vec{x}$ only through $T(\vec{x})$ and h depends only on $\vec{x}$ and is independent of θ

Proof.
$\Rightarrow$ Let T be sufficient according to the definition; we want to show that $p_{\vec{X}}(\vec{x}, \theta)$ factors

$$\begin{aligned}
p_{\vec{X}}(\vec{x}, \theta) = \mathbb{P}[\vec{X} = \vec{x}, \theta] &= \mathbb{P}[\vec{X} = \vec{x}, T(\vec{X}) = t, \theta] = 0 \text{ if } t \neq T(\vec{x}) \text{ otherwise, let } t = T(\vec{x}) \\
&= \mathbb{P}[\vec{X} = \vec{x}, T(\vec{X}) = T(\vec{x}); \theta] \\
&= \mathbb{P}[\vec{X} = \vec{x}, \theta | T(\vec{X}) = T(\vec{x})] \cdot \mathbb{P}[T(\vec{X}) = T(\vec{x}), \theta] \\
&= \underbrace{\mathbb{P}[\vec{X} = \vec{x} | T(\vec{X}) = T(\vec{x})]}_{h(\vec{x})} \cdot \underbrace{\mathbb{P}[T(\vec{X}) = T(\vec{x}), \theta]}_{g(T(\vec{x}), \theta)}
\end{aligned}$$

$\Leftarrow$ Let $p_{\vec{X}}(\vec{x}_0, \theta)$ be factorized as above. WTS T is sufficient for θ; let $\vec{x} = \vec{x}_0$

$$\begin{aligned}
p_{\vec{X}|T(\vec{X})=t}(\vec{x}_0, \theta|t) &= \mathbb{P}[\vec{X} = \vec{x}_0, \theta | T(\vec{X}) = t] = \mathbb{P}[\vec{X} = \vec{x}_0, \theta | T(\vec{X}) = T(\vec{x}_0)] = \mathbb{P}[\vec{X} = \vec{x}_0, \theta] \Big/ \mathbb{P}[T(\vec{X}) = T(\vec{x}_0)] \\
&= \Big[g(T(\vec{x}_0), \theta) h(\vec{x}_0) \Big] \Big/ \sum \mathbb{P}[\vec{x} : T(\vec{x}) = T(\vec{x}_0)] \\
&= \Big[g(T(\vec{x}_0), \theta) h(\vec{x}_0) \Big] \Big/ \sum_{\vec{x} : T(\vec{x}) = T(\vec{x}_0)} g(T(\vec{x}), \theta) h(\vec{x}) \\
&= \Big[g(T(\vec{x}_0), \theta) h(\vec{x}_0) \Big] \Big/ \sum_{\vec{x} : T(\vec{x}) = T(\vec{x}_0)} g(T(\vec{x}_0), \theta) h(\vec{x})
\end{aligned}$$

and since the g's cancel, thus we get something that is independent of θ, thus T is sufficient! □

2.2.5 Minimal Sufficiency

Definition 2.2.8. We say that a statistic T is **minimal sufficient** for θ if T is sufficient and is a function of every other sufficient statistic

Theorem 2.2.9. *Let T be some statistic. Suppose that, for a given $\vec{x}$ and $\vec{y}$, then*

$$(\star) \quad \frac{f_{\vec{X}}(\vec{x},\theta)}{f_{\vec{X}}(\vec{y},\theta)}$$

is independent of θ if and only if $T(\vec{x}) = T(\vec{y})$; then T is minimal sufficient for θ

Proof. Let T be the given statistic satisfying $(\star)$. Suppose that T is sufficient and let T' be any other sufficient statistic. We show that $T'(\vec{x}) = T'(\vec{y}) \Rightarrow T(\vec{x}) = T(\vec{y})$ (i.e. $T = g(T')$ for some g). Since T is sufficient:

$$\frac{p_{\vec{X}}(\vec{x},\theta)}{p_{\vec{X}}(\vec{y},\theta)} = \frac{g(T'(\vec{x}),\theta)h(\vec{x})}{g(T'(\vec{y}),\theta)h(\vec{y})} \quad \text{by Neyman-Fisher Factorization}$$

If we assume that $T'(\vec{x}) = T'(\vec{y})$ then

$$\frac{p_{\vec{X}}(\vec{x},\theta)}{p_{\vec{X}}(\vec{y},\theta)} = \frac{h(\vec{x})}{h(\vec{y})} \quad \text{is independent of } \theta$$

but by $(\star)$, the fraction on the LHS is independent of θ iff $T(\vec{x}) = T(\vec{y})$ i.e. T is minimal sufficient. It remains to show that T is sufficient.
Let $\vec{x}_0$ be some point in S, then

$$\begin{aligned}\mathbb{P}[\vec{X} = \vec{x}_0;\theta | T(\vec{X}) = T(\vec{x}_0)] &= \frac{\mathbb{P}[\vec{X} = \vec{x}_0;\theta]}{\mathbb{P}[T(\vec{X}) = T(\vec{x}_0);\theta]} = \frac{\mathbb{P}[\vec{X} = \vec{x}_0;\theta]}{\sum_{\vec{x}:T(\vec{x})=T(\vec{x}_0)} \mathbb{P}[\vec{X} = \vec{x};\theta]} \\ &= \left[\sum_{\vec{x}:T(\vec{x})=T(\vec{x}_0)} \frac{\mathbb{P}[\vec{X} = \vec{x};\theta]}{\mathbb{P}[\vec{X} = \vec{x}_0;\theta]}\right]^{-1}\end{aligned}$$

now the sum is over "$\vec{x},\vec{y}$": $T(\vec{x}) = T(\vec{y})$ which means by $(\star)$ that the fration of the probabilities in the last term is independent of θ; thus the LHS (at the beginning) is independent of θ thus T is sufficient by definition □

2.2.6 Completeness and Return to UMVUEs

Theorem 2.2.10. ***(Rao Blackwell Theorem)*** *Let U be unbiased for θ, let T be sufficient for θ. Then let $\hat{\theta} = \mathbb{E}[U|T]$*

1. $\mathbb{E}[\hat{\theta}] \overset{\theta}{\equiv} \theta$
2. *$Var[\hat{\theta}] \leq Var[U]$ for all θ*

Proof. First note that $\hat{\theta}$ is, indeed, an unbiased estimator of θ since it does not depend on θ. Reason for this: since T is sufficient, then by definition of sufficiency, the conditional distribution of U given T is independent of θ. So, in particular, the conditional mean $\mathbb{E}[U|T]$ is independent of θ. For the proof of the first claim, we have that

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}_T[\mathbb{E}[U|T]] \overset{\text{L.T.E}}{=} \mathbb{E}[U] \equiv \theta \quad \forall \theta \in \Theta$$

and the second

$$\text{Var}[U] \overset{L.T.V}{=} \text{Var}_T[\mathbb{E}[U|T]] + \mathbb{E}_T[\text{Var}[U|T]] = \text{Var}[\hat{\theta}] + \underbrace{\mathbb{E}_T[\text{Var}[U|T]]}_{\geq 0} \Rightarrow \text{Var}[U] \geq \text{Var}[\hat{\theta}]$$

□

1. Process of conditioning on a sufficient statistic is called "Rao-Blackwellization"

2. We have equality (of the variances) if and only if $U = U(T)$, that is U is only a function of a sufficient statistic

Theorem 2.2.11. ***(Generalized Rao Blackwell)*** *Let V be any estimator of θ (not necessarily unbiased). Then let $\hat{\theta} = \mathbb{E}[V|T]$ where T is sufficient for θ, then MSE$(\hat{\theta}) <$ MSE(V)*

Proof.

$$\begin{aligned}\text{MSE}[V] = \text{Var}[V] + (\mathbb{E}[V] - \theta)^2 &= \text{Var}_T[\mathbb{E}[V|T]] + \mathbb{E}_T[\text{Var}[V|T]] + (\mathbb{E}_T[\mathbb{E}[V|T]] - \theta)^2 \\ &\geq \text{Var}[\hat{\theta}] + (\mathbb{E}_T[\hat{\theta}] - \theta)^2 = \text{MSE}[\hat{\theta}]\end{aligned}$$

□

Definition 2.2.9. A family of distributions indexed by a parameter θ is said to be **complete** if for any function g, with $\mathbb{E}_\theta[g(X)] = 0$ for all $\theta \in \Theta$, we have that $g(x) = 0$ with probability one for all θ

Definition 2.2.10. A statistic T is said to be **complete** if the family of distribution that it induces is complete

Theorem 2.2.12. ***(Lehmann Scheffe Theorem)*** *Let U be an unbiased estimator of θ and $U = U(T)$ is a function of T only, where T is complete and sufficient then*

1. *U is the unique unbiased estimator of θ that is a function of T*

2. *U is the UMVUE of θ*

Proof.
(a) Uniqueness: Let $g(T)$ be any other unbiased estimator of θ that is a function of T. Then

$$\mathbb{E}[U(T) - g(T)] \equiv 0 \quad \forall \theta \in \Theta$$

now let $W(T) = U(T) - g(T)$. Since T is complete, then $W(T) = 0$ with probability 1 for all θ, thus $U(T) = g(T)$ with probability 1 for all θ

(b) Now let V be an unbiased estimator (irrespective of whether it is a function of T or not). **Claim:** We cannot have Var$[V] <$ Var$[U(T)]$. Reason: Consider $\mathbb{E}[V|T]$, i.e. Rao-Blackwellized V. We have that $\mathbb{E}[V|T]$ is a function of T and unbiased for θ; thus by Rao-Blackwell, we have that $\mathbb{E}[U|T] = U(T)$ (by the uniqueness from part (a)). Now if

$$\text{Var}[V] < \text{Var}[U(T)] \quad \text{then} \quad \text{Var}[\mathbb{E}[U|T]] \leq \text{Var}[V] < \text{Var}[U(T)]$$

but this is impossible since $\mathbb{E}[U|T] = U(T)$ with probability 1 □

Can use L-S in the following two ways

1. Identify an unbiased estimator that is a function of a complete sufficient statistic

2. Find any unbiased estimator and Rao-Blackwellize it with a complete, sufficient statistic

Chapter 3

Interval Estimation - Confidence Intervals

Instead of finding point estimates for θ, we find an interval which will capture θ with high proabability. Specifically, let θ be our unknown parameter.

Definition 3.0.1. We say that the interval (L, R) is a **100(1-α)-percent confidence interval** for θ if

$$\mathbb{P}[L < \theta < R] = 1 - \alpha \quad \forall \theta \in \Theta$$

here, $L = L(\vec{X})$ and $R = R(\vec{X})$; this is a confidence interval for unknwn variables.

Notes

1. $1 - \alpha$ is called the **confidence coefficient**; usually chosen by the user such that $1 - \alpha$ is large (also called the **coverage probability**)

2. When the realized data $\vec{x}$ are substituted for $\vec{X}$, the interval has fixed (non-random) endpoints. Since θ is also fixed (despite being unknown), we have that

$$\mathbb{P}[L(\vec{x}) < \theta < R(\vec{x})] = \begin{cases} 1 & \text{if } \theta \in (L(\vec{x}), R(\vec{x})), \\ 0 & \text{if } \theta \notin (L(\vec{x}), R(\vec{x})) \end{cases} \tag{3.1}$$

3. $1 - \alpha$ is a *probability*, it **does not** reflect a level of confidence that $\theta \in (L(\vec{x}), R(\vec{x}))$; there is no mathematical definition of **confidence**. However, $1 - \alpha$ does represent the *pre-experiment* probability that θ will be in the *random* interval $(L(\vec{X}), R(\vec{X}))$

Interpretation of the definition

- **Weak Interpretation:** If we were to repeat our experiment a large number of times, each time computing a $100(1 - \alpha)$-percent confidence interval, then IF we could see the true θ, we would observe that roughly $100(1 - \alpha)$-percent of confidence intervals will have captured θ

- **Strong Interpretation** If we were to repeat the experiment with possibly different unknown θ, say θ_1, θ_2, etc. and compute a $100(1 - \alpha)$-percent confidence interval for each θ_i, then roughly $100(1 - \alpha)$-percent of our intervals should have captured their respective θ_is

(did not type up the rest of this section, sorry!)

Chapter 4

Systematically Finding Point Estimators

4.1 Method of Moments

Suppose that $X_1, \ldots, X_n$ are i.i.d coming from some cdf $F_X(x, \theta)$, where $\theta = (\theta_1, \ldots, \theta_r)$ is a vector of unknown parameters. Let $\mu_i = \mathbb{E}[X^i]$ be the moment about zero for $i = 1, \ldots$. In general, if $\mu_1, \mu_2, \ldots, \mu_r$ exist then there will be a function of $(\theta_1, \ldots, \theta_r)$ such that $\mu_j = g_j(\theta_1, \ldots, \theta_r)$. If we can solve for the θ_js in terms of the μ_js, then we get $\theta_j = h_j(\mu_1, \ldots, \mu_r)$

Definition 4.1.1. The **i-th sample moment** is denoted by

$$m_i = \frac{1}{n} \sum_{k=1}^{n} (X_k)^i$$

The idea in the MM is to replace μ_j by the sample moment estimators in the $h_j(\cdot)$ functions; thus we have that $\hat{\theta}_j = h_j(m_1, \ldots, m_r)$

4.1.1 Properties of MM

Only one property!

Theorem 4.1.1. *If $h_1, \ldots, h_r$ are continuous then $\hat{\theta}_1, \ldots, \hat{\theta}_r$ are consistent*

Proof. By the WLLN, we have that $m_i \to \mathbb{E}[X^i] = \mu_i$. Hence, by the continuity of the h_is, we havet that

$$\hat{\theta}_i = h_i(m_1, \ldots, m_r) \to h_i(\mu_1, \ldots, \mu_r) = \mu_i$$

□

Example 4.1.1. Let $X_1, \ldots, X_n$ are i.i.d. from a normal distribution with μ and σ^2 unknown. Find MM estimators of μ and σ^2

$\Rightarrow$ We have that $\theta_1 = \mu$ and $\theta_2 = \sigma^2$; also $\mu_1 = g_1(\mu, \sigma^2) = \mu$ and $\mu_2 = g_2(\mu, \sigma^2) = \mu^2 + \sigma^2$. So immediately, we have that

$$\mu = \mu_1 \Rightarrow m_1 = \frac{1}{n} \sum_{k=1}^{n} X_k = \overline{X} \Rightarrow \hat{\mu} = \overline{X}$$

$$\sigma^2 = \mu_2 - \mu^2 \Rightarrow m_2 = \frac{1}{n} \sum_{k=1}^{n} (X_k)^2$$

$$\begin{aligned} \hat{\sigma}^2 = m_2 - (\hat{\mu})^2 = \left(\frac{1}{n} \sum_{i=1}^{n} (X_k)^2 \right) - \frac{n}{n} \overline{X}^2 &= \frac{1}{n} \left[\sum_{k=1}^{n} (X_k)^2 - n\overline{X}^2 \right] \\ &= \frac{1}{n} \sum_{k=1}^{n} \left(X_k - \overline{X} \right)^2 \end{aligned}$$

4.2 Method of Maximum Likelihood

Definition 4.2.1. We say $\hat{\theta}$ is the **maximum likelihood estimate** as $\vec{x}$ is given by

$$L(\hat{\theta}, \vec{x}) = \max_{\theta \in \Theta} L(\theta, \vec{x})$$

We have that the maximization of the likelihood is carried out in one of three different ways

1. Using calculus
2. Through inspection of the likelihood if calculus cannot be used
3. Numerical methods if $\hat{\theta}$ cannot be found in closed form

(1) Using Calculus to find MLE

Let $\theta = (\theta_1, \theta_2, \ldots, \theta_r) \in \Theta$ be a vector of r unknown parameters. Let $l(\theta, \vec{x}) = \ln[L(\theta, \vec{x})]$. Assuming that the partial derivatives below exist, we set up and solve the following equations for $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_r$.

$$\begin{aligned} \frac{\partial l(\theta_1, \ldots, \theta_r)}{\partial \theta_1} &= 0 \\ \frac{\partial l(\theta_1, \ldots, \theta_r)}{\partial \theta_2} &= 0 \\ \cdots \\ \frac{\partial l(\theta_1, \ldots, \theta_r)}{\partial \theta_r} &= 0 \end{aligned}$$

where the MLEs will be amongst the roots of this set of equations, called the *likelihood equations.*

Example 4.2.1. Let $X_1, X_2, \ldots, X_n$ be i.i.d $\mathrm{N}(\mu, \sigma^2)$ random variables with both μ and σ^2 unknown. Find the MLE of μ and σ^2.
Solution: Here, $\theta = (\mu, \sigma^2)$ where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; thus $\Theta = (-\infty, \infty) \times (0, \infty)$. We have that

$$\begin{aligned} L((\mu, \sigma^2), \vec{x}) &= (2\pi)^{-n/2} \sigma^{-n} \exp\left[\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right] \\ l((\mu, \sigma^2), \vec{x}) &= \ln[L((\mu, \sigma^2), \vec{x})] = C - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

where we have that the maximum likelihood equations (and solved) are

$$\frac{\partial l((\mu, \sigma^2), \vec{x})}{\partial \mu} = 0 \Rightarrow \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \Rightarrow \hat{\mu} = \bar{x}$$

$$\frac{\partial l((\mu, \sigma^2), \vec{x})}{\partial \sigma^2} = 0 \Rightarrow -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

where we subbed in $\bar{x}$ for μ in the set of equations. Thus we have that the ML estima**tors** are

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

Example 4.2.2. Let $X_1, X_2, \ldots, X_n$ are i.i.d Bernoulli(p) random variables with p unknown. Find the MLE of p
Solution:

$$l(p, \vec{x}) = \ln[p^x(1-p)^{n-x}] = x\ln(p) + (n-x)\ln(1-p) \qquad \text{where } x = \sum_{i=1}^{n} x_i$$

and the likelihood equations is given (and solved) as follows:

$$\frac{\partial l(p, \vec{x})}{\partial p} = 0 = \frac{x}{p} - \frac{(n-x)}{1-p} \Rightarrow \hat{p} = \frac{x}{n}$$

thus the maximum likelihood estimator for p is $\hat{p} = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Example 4.2.3. Let $X_1, X_2, \ldots$ be i.i.d Bernoulli random variables. Let n denote the observed trial at which the x-th success occurs. Then N is the pre-experiment trial of the x-th success has a negative Binomial distribution with

$$\mathbb{P}[N = n; p] = \binom{n-1}{x-1} p^{x-1}(1-p)^{n-1-(x-1)} \cdot p = \binom{n-1}{x-1} p^x(1-p)^{n-x}$$

Maximizing $\ln(\mathbb{P}[N = n])$ with respect to p we get $\hat{p} = x/n$, same likelihood as we had in the setup where we had n (fixed trials) and observed x successes (the Binomial coefficient, being a constant, does not affect the maximization). Thus we have that in the Binomial setting, the MLE is $\hat{p} = \frac{X}{n}$; in the negative Binomial setting, $\hat{p} = \frac{x}{N}$

Note: In the first case, the number of successes, X, is random (and Binomial) while in the second case, the number of of successes is fixed, x, and the number of trials to get to the x-th success, N, is random. Thus, the *distribution* of the two $\hat{p}$s will be different. In general, the distribution of the MLEs will depend on the mechanism that will generate the data.

(2) Finding MLE by inspection

It could be that the likelihood is not differentiable — maybe not at θ, where the maximum occurs, as the following example shows. We may still be able to find the MLE by inspection.

Example 4.2.4. Let $X_1, X_2, \ldots, X_n$ be i.i.d $U(0, \theta)$ then the likelihood looks like this:...

$$L(\theta, \vec{x}) = \frac{1}{\theta^n}\prod_{i=1}^{n} \mathbb{1}\{0 < x_i < \theta\} \Rightarrow \frac{1}{\theta^n}\mathbb{1}\{0 < x_{(n)} < \theta\}$$

First note that the since the X_i are $U(0, \theta)$ defined on the open interval, the likelihood does not have a maximum. However, a simple fix solves these difficulties. Simply redefine the X_is to be $U(0, \theta]$. This will not affect any probabilities that we compute from the pdf, since we are changing the pdf at a point of Lebseque measure 0. Then, it is clear that $\hat{\theta}(\vec{x}) = y_n$ and the estimator is $\hat{\theta}(\vec{X}) = Y_n$

Numerical Methods Needed

Example 4.2.5. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\Gamma(\alpha, \beta)$ where both α and β are unknown. It is clear from the form of the pdf, and hence the likelihood, that numerical methods are needed to find $\hat{\alpha}$ and $\hat{\beta}$

Interesting Side Example

Let X have a hypergeometric distribution, i.e.

$$\mathbb{P}[X = x]\frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}} \qquad x \leq \min(a, n)$$

Suppose that n is chosen, a is unknown and x is observed. To find the MLE of N, where N is unknown, we need to maximize $L(N; n, x)$ with respect to N. We cannot use calculus since $N = \{1, 2, \ldots, \}$

4.3 Properties of MLEs

Maximum likelihood estimators have certain good properties which justify their use.

1. Invariance
2. Consistency
3. Asymptotic Normality

(1) Invariance

Let $\hat{\theta}$ be the mle of θ; let $\eta = \tau(\theta)$. Then the mle of η, $\hat{\eta} = \widehat{\tau(\theta)} = \tau(\hat{\theta})$ for any function $\tau(\cdot)$. What follows is the proof for when τ is a one-to-one function; on mycourses is the proof for arbitrary function.

Proof. Let $L^\star(\eta, \vec{x})$ be the likelihood parameter according to η. We have that $L^\star(\eta, \vec{x}) = L(\tau^{-1}(\tau(\theta)), \vec{x})$ where $L(\cdot, \vec{x})$ is the likelihood function with parameter θ; $L(\theta, \vec{x})$

$$\begin{aligned}
\max_\eta L^\star(\eta, \vec{x}) &= L^\star(\hat{\eta}, \vec{x}) \\
&= \max_\theta L(\theta, \vec{x}) = L(\hat{\theta}, \vec{x}) \\
&\text{But } L(\hat{\theta}, \vec{x}) = L(\tau^{-1}(\tau(\hat{\theta}))) = L^\star(\tau(\hat{\theta})) \\
&\Rightarrow \tau(\hat{\theta}) = \hat{\eta} = \widehat{\tau(\theta)}
\end{aligned}$$

□

Example 4.3.1. Suppose that the time to recovery, X, from a certain surgical procedure has a gamma distribution with unknown parameters α and β. It is of interest to estimate

$$\frac{\mathbb{P}[X \geq 5]}{\mathbb{P}[X \geq 3]}$$

using maximum likelihood. Solution follows:

$$\frac{\mathbb{P}[X \geq 5]}{\mathbb{P}[X \geq 3]} = \frac{\int_5^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx}{\int_3^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx}$$

Look's terrible, as it would appear that we would have to start with a likelihood that is a function of this ratio. However, by the invariance of mle, all we need is $\hat{\alpha}$ and $\hat{\beta}$, the mle of α and β, respectively. Thus

$$\widehat{\frac{\mathbb{P}[X \geq 5]}{\mathbb{P}[X \geq 3]}} = \frac{\int_5^\infty \frac{1}{\Gamma(\hat{\alpha})\hat{\beta}^{\hat{\alpha}}} x^{\hat{\alpha}-1} e^{-x/\hat{\beta}} dx}{\int_3^\infty \frac{1}{\Gamma(\hat{\alpha})\hat{\beta}^{\hat{\alpha}}} x^{\hat{\alpha}-1} e^{-x/\hat{\beta}} dx}$$

(2) Consistency

Theorem 4.3.1. *If $X_1, \ldots, X_n$ are i.i.d with pdf (or pmf) $f_X(x, \theta_0)$ (or $p_X(x, \theta_0)$) and $\hat{\theta}$ is the mle of θ_0, then we have*

$$\hat{\theta} \xrightarrow{p} \theta_0 \tag{4.1}$$

(in fact, the convergence is strong; i.e. with probability 1) under certain regularity conditions.

Proof. Omitted (sad); but the sketch of the proof is as follows: The difficulty is that don't necessarily have a closed form expression for $\hat{\theta}$. All we know is that $\hat{\theta}$ maximizes the likelihood. We have that

$$\begin{aligned}
l(\theta, \vec{X}) &= \ln[L(\theta, \vec{X})] \\
&= \sum_{i=1}^n \ln[L(\theta, X_i)] \quad \text{(due to independence)}
\end{aligned}$$

and so we have that $X \overset{D}{=} X_1$ (or $X_2, \dots, X_n$ because they are identically distributed). So we have that EQUATION thus the maximum of $\frac{1}{n} l(\theta, \vec{X})$ converges to the max of $\mathbb{E}_{\theta_0}[L(\theta, \vec{X})]$. Finally, if the functions are smooth, the maximizers should converge to one another. The maximizer of $\frac{1}{n} l(\theta, \vec{X})$ is $\hat{\theta}$ (the mle).

All that remains is to show that $\mathbb{E}_{\theta_0}[l(\theta, \vec{X})]$ is maximized at θ_0. This claim depends on *Jensen's Inequality*:

$$\mathbb{E}[\phi(X)] \leq \phi[\mathbb{E}(X)] \quad \text{where } \phi \text{ is concave}$$

So consider

$$\begin{aligned}
\mathbb{E}\left[\ln \frac{L(\theta, X)}{L(\theta_0, X)}\right] &\leq \ln \mathbb{E}_{\theta_0}\left[\frac{L(\theta, X)}{L(\theta_0, X)}\right] \\
&= \ln \int \frac{\overbrace{L(\theta, x)}^{f(\theta,x)}}{L(\theta_0, x)} \underbrace{f(x, \theta_0)}_{L(\theta_0,x)}\, dx \\
&= \ln \int f(x, \theta) dx = \ln(1) = 0
\end{aligned}$$

this implies that $\mathbb{E}_{\theta_0}[\ln L(\theta, X)] \leq E_{\theta_0}[\ln L(\theta_0, x)]$; i.e. θ_0 is the maximizer of $\mathbb{E}_{\theta_0}[\ln L(\theta, X)]$

□

(3) Asymptotic Normality

Let $X_1, \dots, X_n$ be i.i.d. Let $\hat{\theta}$ be the mle and θ_0 be the true unknown parameter. Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{D}{\to} Z \sim \mathrm{N}(0, I^{-1}(\theta_0))$$

where $I(\theta) = \mathbb{E}_{\theta_0}\left(\left[\frac{\partial}{\partial\theta} \ln f(X, \theta)\Big|_{\theta_0}\right]^2\right)$, which is the denominator in the Cramer-Rao lower bound and since $I^{-1}(\theta_0)$ is the C-R lower bound, this theorem tells us that, in the limit, mles have a normal distribution whose variance is smaller than any other.

Definition 4.3.1. $I(\theta_0)$ is called thet **Fisher Information**

Proof. If we take the derivative of the log-likelihood and do some operations:

$$\begin{aligned}
l'(\theta_0, \vec{x}) &= \frac{\partial}{\partial\theta}\left(\ln f_{\vec{X}}(\vec{x};\theta)\right)\Big|_{\theta_0} \\
&= \sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta}\left(\ln f_X(x_i;\theta)\right)\right]\Big|_{\theta_0} \\
\mathrm{Var}\left[l'(\theta_0,\vec{x})\right] &= n\mathrm{Var}_{\theta_0}\left[\frac{\partial}{\partial\theta}\left(\ln f_X(x;\theta)\right)\Big|_{\theta_0}\right] \\
&= n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_X(x;\theta)\Big|_{\theta_0}\right)^2\right] && \text{since } \mathbb{E}\left[\frac{\partial}{\partial\theta}\ln f_X(x;\theta)\right]=0 \\
\frac{1}{n}\,l''(\theta_0,\vec{x}) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2}{\partial\theta^2}\left(\ln f_X(x_i;\theta)\right)\right)\Big|_{\theta_0} \\
&\overset{probability}{\rightarrow} \mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(x;\theta)\Big|_{\theta_0}\right] && \text{by the Law of Large Numbers}
\end{aligned}$$

We take the first order Taylor Expansion:

$$l'(\hat{\theta},\vec{x}) \approx l'(\theta_0,\vec{x}) + (\hat{\theta}-\theta_0)l''(\theta_0,\vec{x})$$

which is justified since $\hat{\theta}$ is consistent for θ_0.
Since $\hat{\theta}$ is the MLE, then the left hand side is zero, so:

$$l'(\theta_0,\vec{x}) \approx (\hat{\theta}-\theta_0)\cdot(-l''(\theta_0,\vec{x}))$$

If we standardize both sides, then the left-hand side is given by:

$$\frac{l'(\theta_0,\vec{x}) - \mathbb{E}_{\theta_0}\left[l'(\theta_0,\vec{x})\right]}{\sqrt{\mathrm{Var}_{\theta_0}\left[l'(\theta_0,\vec{x})\right]}} \approx \frac{(\hat{\theta}-\theta_0) - l''(\theta_0,x)}{\sqrt{\mathrm{Var}_{\theta_0}\left[l'(\theta_0,\vec{x})\right]}} \qquad \text{since } \mathbb{E}_{\theta}\left[l'(\theta_0,\vec{x})\right]=0$$

Since $l'(\theta_0,\vec{x})$ is the sum of iid random variables, then the left hand side converges in distribution to $Z \sim N(0,1)$. If we standardize the right-hand side:

$$\begin{aligned}
\frac{(\hat{\theta}-\theta_0)\cdot(-l''(\theta_0,\vec{x}))}{\sqrt{n\mathbb{E}_{\theta_0}\left[\left[\frac{\partial}{\partial\theta}\left(\ln f_X(x,\theta)\right)\right]\right]^2}} &= \frac{(\hat{\theta}-\theta_0)\cdot\left(-\frac{n\cdot l''(\theta_0,\vec{x})}{n}\right)}{\sqrt{n\mathbb{E}_{\theta_0}\left[\left[\frac{\partial}{\partial\theta}\left(\ln f_X(x,\theta)\right)\right]\right]^2}} \\
&= \sqrt{n}\frac{-\frac{1}{n}\cdot l''(\theta_0,\vec{x})}{\sqrt{\mathbb{E}\left[\left[\frac{\partial}{\partial\theta}\ln f_X(\vec{x},\theta)\right]\Big|_{\theta_0}\right]^2}}
\end{aligned}$$

As $n\to\infty$, then:

$$-\frac{1}{n}\cdot l''(\theta_0,\vec{x}) \overset{probability}{\rightarrow} -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(x,\theta)\Big|_{\theta_0}\right] = \mathbb{E}\left[\left[\frac{\partial}{\partial\theta}\ln f(x,\theta)\Big|_{\theta_0}\right]^2\right]$$

Finally, the right hand side converges to a $Z \sim N(0,1)$ distribution so:

$$\sqrt{n}(\hat{\theta}-\theta_0)\sqrt{I(\theta_0)} \overset{distribution}{\rightarrow} Z \sim N(0,1)$$

which implies that:

$$\sqrt{n}(\hat{\theta}-\theta_0) \overset{distribution}{\rightarrow} N\left(0,\frac{1}{I(\theta_0)}\right)$$

where:

$$I(\theta_0) = \mathbb{E}\left[\left[\frac{\partial}{\partial\theta}\ln f_X(x,\theta)\Big|_{\theta_0}\right]^2\right]$$

is the Fisher Information of X. □

Chapter 5

Hypothesis Testing

5.1 Hypothesis Testing

Up to now, we have been concerned with problems of estimator (either point or interval) of an unknown parameter. Now consider the following problem:

Let Θ be our parameter space, and suppose that we can write Θ as the disjoint union of Θ_0 and Θ_1, so $\Theta = \Theta_0 \cup \Theta_1$ such that $\Theta_0 \cap \Theta_1 = \emptyset$.

The hypothesis testing problem is to decide on the basis of observations $X_1, X_2, \ldots, X_n$ whether $\theta \in \Theta_1$ or $\theta \in \Theta_2$.

Example 5.1.1. Suppose that artificial hips were previously manufactured from ceramic material. It was known from data collected that their expected time to failure is $\mu = 15$ years. Manufacturers decide to start manufacturing using titanium. We are interested to see if these new hips have an expected lifetime $\mu > 15$ years. The hypothesis testing problem here is:

Based on the data collected of the time to failure for the titanium hips, decide if:

$$\theta = \theta_0 \text{ (so } \mu = 15) \quad \text{or} \quad \theta \in \left]15, \infty\right[\text{ (so } \mu > 15)$$

so we define $\Theta = \{x \in \mathbb{R} : x \geq 15\} = \Theta_0 \cup \Theta_1$ where $\mu = 15$ as Θ_0 and $\mu > 15$ as Θ_1.

Definition 5.1.1. The two choices for θ are called the **null hypothesis**, written as $H_0 : \theta \in \Theta_0$, and the **alternative hypothesis** as $H_a : \theta \in \Theta_1$

Remark. We shall use observations $\vec{X}$ from $f_{\vec{X}}(\vec{x}; \theta)$ to help us make this decision. To this end, we will use what we call a test statistic $T = T(\vec{X})$ which is a function of $\vec{X}$ and (usually) a hypothesized θ.

Remark. We will have to come up with the rules that will enable us to decide between H_0 and H_a. The procedure is as follows:

1. Partition the range of $T(\vec{X})$ into two regions A (the acceptance region) and R (the rejection region) such that:
   - If $T(\vec{X}) \in R$ then we will reject the null hypothesis H_0 and accept H_a.
   - If $T(\vec{X}) \in A$ then we "accept" the null hypothesis H_0
     (there are quotes on "accept" cause it's apparently not that easy)

5.1.1 Components for Hypothesis Testing

1. $H_0 : \theta \in \Theta_0$ versus $H_a : \theta \in \Theta_1$; example are $H_0 : \mu = 15$ and $H_a : \mu > 15$, etc
2. Test Statistic $T = T(\vec{X})$
3. Rejection Rule (or Rejection Region): A and R

We agree before seeing our data that if $T \in R$, we **reject** H_0 and **accept** H_a and if $T \in A$ we will "**accept**" H_0

Jargon (Definitions)

If Θ_i, for $i = 0, 1$ consists of a single point (i.e. $H_0 : \mu = 15$) then we say that H_i is a **simple** (or a point) hypothesis. If Θ_i contains more than one point, we call it a **composite** hypothesis (i.e. $H_a : \mu > 15$). Hypothesis can come in various combinations of these two types. Until further notice, consider the "simple versus simple" setting; $H_0 : \theta = \theta_0$ and $H_a : \theta = \theta_1$

How do we decide on our rejection rule?

First, we realize that whatever rule we decide on there is always the possibility that when we apply the rule, we will make an error. These are the two possible types of error that could occur (prior to our experiment).

1. Type 1: Rejecting H_0 when, in fact, H_0 is true. This will happen when $T \in R$ but H_0 is true i.e. $\theta = \theta_0$
2. Type 2: We "accept" H_0 when, in fact, H_a is true. This will happen when $T \in A$ but H_a is true i.e. $\theta = \theta_1$

The idea is that we want to choose a rejection rule that makes the probability of each of the two types of errors **small**; i.e. don't want to make errors

More Jargon (More Definitions)

We call the probability of a Type 1 error the **significance level** of the test (i.e. of the rejection rule) and denoted by α. Thus $\mathbb{P}_{\theta_0}[T \in R] = \alpha$. The probability of a Type 2 error is denoted by β; thus $\beta = \mathbb{P}_{\theta_1}[T \in A]$. We call $1 - \beta = \mathbb{P}_{\theta_1}[T \in R]$ (i.e. probability of rejecting when you SHOULD be rejecting) the **power** of the test that $\theta = \theta_1$

As was said, we would like both α and β to be small. By convention, $\alpha \leq 0.05$ is considered "small". Unfortunately it is usually very difficult to ensure that β will aos be small (0.05 or less). Thus, as $\alpha \downarrow, \beta \uparrow$

What do we do to combat this? We worry about the Type 1 error to start with and we set up our rejection rile so that α is small (often chosen to be 0.05).

Comment: We *can* reduce β (with the given reject rule, R, and α). A β of 0.2 is considered attainable (i.e. $1 = \beta = 0.8$). If $T \in R$ then you reject $H_0 : \theta = \theta_0$ and accept $H_a : \theta = \theta_1$. If R has a Type 1 error with probability of α and $\alpha = 0.05$, you are happy with $T_{\text{obs}} \in R$ to reject H_0, only because the pre-experiment probability that this rejection region would falsely reject H_0 is 0.05 i.e. in repeated use, the rejection rule would falsely reject H_0 on roughly 5-percent of occasions. Clearly, if T_{obs} falls in R, we need not worry at this time about β (or $1 - \beta$) because we cannot make a Type 2 error of $T \notin A$. If $T_{\text{obs}} \in A$, knowing that β will be often greater than or equal to 0.2, we should not feel comfortable accepting H_0 since our procedure would have at least a probability of 0.2 (large) of falsely accepting H_0. Thus, in practice, we use the following statement: "If $T_{\text{obs}} \in A$, there is not enough evidence to

reject H_0 at the 100α-percent level of significance". This is a no-decision statement.
If you could ensure that $\beta = 0.05$, then there is no reason to avoid saying "We accept H_0". Thus, because your rejection rule will only, on 5-percent of occasions, falsely accept H_0. Most often, β will not be small enough.

Summary

- If $T \in R$, then say "We reject H_0 at the 100α-percent level of significance
- If $T \in A$, then say "There is no evidence to reject the null hypothesis at the 100α-percent level of significance "

Since we determine our rejection rule by considering only the significance level (i.e. the Type 1 Error probability), what we choose as H_0 and H_a is important.

Rule: Consider which of the two types of error is more serious within the context of the applied. Choose H_0 so that the more "serious" of the two types of error is the Type 1 Error. Often, this means that H_0 represents the status quo i.e. we are looking to overthrow the status quo

Example 5.1.2. To make things less abstract, consider the following example. Let $X_1, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables with μ and σ^2 unknown. Consider the hypothesis:

1. $H_0 : \mu = \mu_0$ versus $H_a : \mu = \mu_1$ (simple versus simple), with $\mu_1 > \mu_0$
2. Let the test statistic for the hypothesis be $T = (\overline{X} - \mu)/(s/\sqrt{n})$
3. **Rejection Rule** Reject the null and accept the alternative if $T > t_{n-1}(\alpha)$

Rational The significance level of the test is $\mathbb{P}_{\mu_0}[T \in R] = \mathbb{P}_{\mu_0}[T > t_{n-1}(\alpha)]$ where $T \sim t_{n-1}$; thus $\mathbb{P}_{\mu_0}[T \in R] = \alpha$ by definition of $t_{n-1}(\alpha)$

Note α is just the proportion of times this test statistic would be greater than $t_{n-1}(\alpha)$ if $H_0 : \mu = \mu_0$ were true

Generalization Now let's generalize the hypothesis to something more interesting.

1. $H_0 : \mu = \mu_0$ and $H_a : \mu > \mu_0$
2. Test statistic: $T = (\overline{X} - \mu)/(s/\sqrt{n})$
3. Rejection Rejection: Reject at the 100α-percent level of significance if $T > t_{n-1}(\alpha)$

Note! The rejection rule for the two types of alternative hypotehsis do NOT depend on μ_1 at all. Does this rejection rule have "good" properties? To then end, we need to define what we mean by "good" property

Definition 5.1.2. Consider $H_0 : \theta = \theta_0$ and $H_a : \theta = \theta_1$. We say that a test with rejection region (rule) R is a most powerful α-level test (rejection rule) if

1. $\mathbb{P}_{\theta_0}[T \in R] = \alpha$
2. If $T^\star$ is any other α-level test with rejection region $R^\star$ then

$$\underbrace{\mathbb{P}_{\theta_1}[T \in R]}_{1-\beta} \geq \mathbb{P}_{\theta_1}[T^\star \in R^\star]$$

and

$$1 - \mathbb{P}_{\theta_1}[T \in R^C] = 1 - \mathbb{P}_{\theta_1}[T \in A]$$

i.e. R gives the α-level test with power at least that of any other α-level test of $\theta = \theta_0$ versus $\theta = \theta_1$

Definition 5.1.3. Suppose we have $H_0 : \theta \in \theta_0$ and $H_a : \theta \in \Theta_1$, where Θ_1 could be a non-singleton set. An α-leve test with statistic T and rejection region R is called **uniformly most powerful** α-level test of H_0 versus H_a if for any other α-level test, $T^\star$ with rejection region $R^\star$, we have that

$$\mathbb{P}_{\theta_1}[T \in R] \geq \mathbb{P}_{\theta_1}[T^\star \in R^\star]$$

for all $\theta \in \Theta_1$

Theorem 5.1.1. ***(Neyman-Pearson Lemma)*** *Let $\vec{X} = X_1, X_2, \dots, X_n$ have pdf $f_{\vec{X}}(\vec{x}, \theta)$ where $\theta \in \Theta = \{\theta_0, \theta_1\}$. Suppose we wish to test $H_0 : \theta = \theta_0$ versus $H_a : \theta = \theta_1$. Then the most powerful α-level test of H_0 versus H_a is given by the rejection rule:*

$$\textit{Reject } H_0 \textit{ if } f_{\vec{X}}(\vec{X}, \theta_0) < c_\alpha f_{\vec{X}}(\vec{X}, \theta_1)$$

for some c_α, which is determined by the requirement that

$$\mathbb{P}_{\theta_0}[f_{\vec{X}}(\vec{x}, \theta_0) < c_\alpha f_{\vec{X}}(\vec{x}, \theta_1)] = \alpha$$

i.e. we reject H_0 in favor of H_a if the data under H_a are more likely than the data under H_0. How much more, is determined so that with this rule, we will 100α-percent of the time falsely reject H_0

Example 5.1.3. Let $X_1, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables with μ unknown and σ^2 known (otherwise, it would be a *composite* hypothesis, which we do not want; we're looking at a simple versus simple case). Find the most powerful α-level test of $H_0 : \mu = \mu_0$ versus $H_a : \mu = \mu_1$ where $\mu_1 > \mu_0$

⇒ Since we have a simple-versus-simple hypothesis, we can use the Neyman-Pearson Lemma (NP). The most powerful test is of the form H_0 if $f_{\vec{X}}(\vec{X}, \theta_0) < c_\alpha f_{\vec{X}}(\vec{X}, \theta_1)$ i.e.

$$
\begin{aligned}
&(2\pi)^{n/2} \frac{1}{\sigma^n} \exp\left[\frac{-1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2\right] < c_\alpha (2\pi)^{n/2} \frac{1}{\sigma^n} \exp\left[\frac{-1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2\right] \\
&\exp\left[\frac{-1}{2\sigma^2} \sum_{i=1}^n X_i^2 + \mu_0 \sum_{i=1}^n X_i - \mu_0^2\right] < c_\alpha \exp\left[\frac{-1}{2\sigma^2} \sum_{i=1}^n X_i^2 + \mu_1 \sum_{i=1}^n X_i - \mu_1^2\right] \\
&\frac{-1}{2\sigma^2} \sum_{i=1}^n X_i^2 + \frac{\mu_0}{2\sigma^2} \sum_{i=1}^n X_i - \frac{\mu_0^2}{2\sigma^2} + \frac{1}{2\sigma^2} \sum_{i=1}^n X_i^2 - \frac{\mu_1}{2\sigma^2} \sum_{i=1}^n X_i + \frac{\mu_1^2}{2\sigma^2} < c_\alpha^\star \\
&\mu_0 \sum_{i=1}^n X_i - \mu_1 \sum_{i=1}^n X_i < c_\alpha^{\star\star} \Rightarrow \sum_{i=1}^n X_i(\mu_0 - \mu_1) < c_\alpha^{\star\star} \Rightarrow \sum_{i=1}^n X_i > c_\alpha^{\star\star\star} \Rightarrow \overline{X} > c_\alpha^{4\star}
\end{aligned}
$$

Now we must determine this constant so that

$$\mathbb{P}_{\mu_0}[\overline{X} > c_\alpha^{4\star}] = \alpha \Rightarrow \mathbb{P}_{\mu_0}\left[\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > c^{5\star}\right] = \alpha$$

and since under the null, the X_i are i.i.d normal with mean μ_0 and variance σ^2, the fraction in the probability is distributed as standard normal. So we set $c^{5\star} = z_\alpha$. Thus, the most powerful α-level test says to reject if

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$$

Now consider the more interesting $H_0 : \mu = \mu_0$ and $H_a : \mu > \mu_0$; well we have just seen the most powerful α-level test of $H_0 : \mu = \mu_0$ versus $H_a : \mu = \mu_1$ has a rejection rule that is independent of μ_1. Thus the rejection rule for this more interesting case is the same as before and is uniformly most powerful (UMP)