Math Review Notes—Mathematical Statistics

Gregory Faletto

Contents

Last updated July 6, 2020

Chapter 1

Mathematical Statistics

These are my notes from taking Math 541A at USC taught by Steven Heilman as well as *Statistical Inference* (2nd edition) by Casella and Berger [Casella and Berger, 2001], *Testing Statistical Hypotheses* by Lehmann and Romano [Lehmann and Romano, 2005], Statistics 100B at UCLA taught by Nicolas Christou, ISE 620 at USC taught by Sheldon Ross, Math 505A at USC taught by Sergey Lototsky, and a few other sources I cite within the text.

1.1 Order Statistics

Definition 1.1.1 (**Order statistics (from Math 541A, more precise)**). Let $X : \Omega \to \mathbb{R}$ be a random variable. Let $X_1, \dots, X_n$ be a random sample of size n from X. Define $X_{(1)} := \min_{1 \le i \le n} X_i$, and for any $2 \le i \le n$, inductively define

$$X_i := \min \Big\{ \{X_1, \dots, X_n\} \setminus \{X_{(1)}, \dots, X_{(i-1)}\} \Big\},$$

so that

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)} = \max_{1 \le i \le n} X_i.$$

The random variables $X_{(1)}, \dots, X_{(n)}$ are called the **order statistics** of $X_1, \dots, X_n$.

Definition 1.1.2 (**Order statistics (from ISE 620, more informal)**). Let $X_1, \dots X_n \sim iid \;\; F$ with $F' = f$. Define $X_{(1)}$ as the smallest among $X_1, \dots, X_n$, $X_{(2)}$ as the 2nd smallest, and so on, up to $X_{(n)}$, the largest of the group. We call $X_{(1)}, \dots, X_{(n)}$ the **order statistics** of $X_1, \dots, X_n$.

Proposition 1.1.0.1 (**Order statistics distribution function; from Math 541A**). Suppose X is a discrete random variable and we can order the values that X takes as $x_1 < x_2 < \cdots$. For any $i \ge 1$, define $p_i := \Pr(X \le x_i)$. Then for any $1 \le i, j \le n$,

$$\Pr(X_{(j)} \le x_i) = \sum_{k=j}^{n} \binom{n}{k} p_i^k (1 - p_i)^{n-k}.$$

Proof. Note that $\{X_{(j)} \le x_i\}$ is equivalent to the event that j or more of the X_i are less than or equal to x_i regardless of order; that is, x_i is the kth smallest observed value. Let A_k be the event that exactly k of the X_i are less than or equal to x_i regardless of order. Then

$$\{X_{(j)} \leq x_i\} = \bigcup_{k=j}^{n} A_k.$$

Then since (by definition of p_i)

$$\Pr(A_k) = \binom{n}{k} p_i^k (1-p_i)^{n-k}$$

and using the fact that the $\{A_k\}$ are disjoint, we have

$$\Pr(\{X_{(j)} \leq x_i\}) = \Pr(\bigcup_{k=j}^{n} A_k) = \sum_{k=1}^{n} \Pr(A_k) = \sum_{k=1}^{n} \binom{n}{k} p_i^k (1-p_i)^{n-k}.$$

□

Corollary 1.1.0.1.1. If X is a continuous random variable with density f_X and cumulative distribution function F_X, then for any $1 \leq j \leq n$, $F_{X_{(j)}}$ has density

$$f_{X_{(j)}}(x) := \frac{n!}{(j-1)!(n-j)!} f_X(x)(F_X(x))^{j-1}(1-F_X(x))^{n-j}, \qquad \forall\, x \in \mathbb{R}.$$

Proof. This follows by differentiating the identity from Proposition 1.1.0.1 for the cumulative distribution function.

□

Proposition 1.1.0.2 (Order statistics joint density function; result from ISE 620). The joint density of the order statistics is

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots,x_n) = n! \prod_{i=1}^{n} f(x_i).$$

Proof. Start with $n = 2$. We seek $f_{X_{(1)},X_{(2)}}(x_1, x_2)$. Note that $X_{(1)} = x_1, X_{(2)} = x_2$ if $X_1 = x_1, X_2 = x_2$ or if $X_1 = x_2, X_2 = x_1$. These are mutually exclusive events, so their density is equal to the sums of the two densities. That is,

$$f_{X_{(1)},X_{(2)}}(x_1,x_2) = f_{X_1,X_2}(x_1,x_2) + f_{X_1,X_2}(x_2,x_1) = 2f(x_1)f(x_2)$$

where the last step follows from the i.i.d. distributions. Generalizing, we have

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots,x_n) = n! \prod_{i=1}^{n} f(x_i).$$

□

Proposition 1.1.0.3 (**Distribution of order statistics of uniform random variable; from 541A**). Let X be a random variable uniformly distributed in $[0,1]$. Then for any $1 \leq j \leq n$, $X_{(j)}$ is a beta distributed random variable with parameters j and $n+1-j$.

Proof. Note that for a uniform distribution on $[0,1]$, $f_X(x) = 1, x \in [0,1]$ and $F_X(x) = x, x \in [0,1]$. Therefore by Corollary 1.1.0.1.1 we have

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!}(x)^{j-1}(1-x)^{n-j}, \qquad x \in [0,1]$$

$$= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n+1-j)}x^{j-1}(1-x)^{n-j} = \frac{\Gamma(j+n+1-j)}{\Gamma(j)\Gamma(n+1-j)}x^{j-1}(1-x)^{n+1-j-1}$$

which is the pdf for a beta distribution with parameters j and $n+1-j$.

□

Corollary 1.1.0.3.1. Let X be a random variable uniformly distributed in $[0,1]$. Then $\mathbb{E}X_{(j)} = \frac{j}{n+1}$

Proof. Follows from Proposition 1.1.0.3 since the mean of such a beta distribution is $\frac{j}{n+1}$.

□

Proposition 1.1.0.4 (**Result from 541A**). Let $a, b \in \mathbb{R}$ with $a < b$. Let U be the number of indices $1 \leq j \leq n$ such that $X_j \leq a$. Let V be the number of indices $1 \leq j \leq n$ such that $a < X_j \leq b$. Then the vector $(U, V, n-U-V)$ is a multinomial random variable, so that for any nonnegative integers u, v with $u + v \leq n$, we have

$$\begin{aligned}&\mathbb{P}(U = u, V = v, n - U - V = n - u - v)\\ &\quad = \frac{n!}{u!v!(n-u-v)!}F_X(a)^u(F_X(b) - F_X(a))^v(1 - F_X(v))^{n-u-v}.\end{aligned}$$

Consequently, for any $1 \leq i, j \leq n$,

$$\mathbb{P}(X_{(i)} \leq a, X_{(j)} \leq b) = \mathbb{P}(U \geq i, U + V \geq j) = \sum_{k=i}^{j-1}\sum_{m=j-k}^{n-k}\mathbb{P}(U = k, V = m) + \mathbb{P}(U \geq j).$$

So, it is possible to write an explicit formula for the joint distribution of $X_{(i)}$ and $X_{(j)}$.

Proof. We can define a multinomial distribution as follows (from Sheldon Ross *Stochastic Processes*, see Definition **??**): "Suppose that n independent trials, each of which results in either outcome $1, 2, \ldots, r$ with respective probabilities $p_1, p_2, \ldots, p_r$ (with $\sum_i p_i = 1$), are performed. Let N_i denote the number of trials resulting in outcome i. Then the joint distribution of $N_1, \ldots, N_r$ is called the **multinomial distribution.**" In this case $r = 3$. If we define outcome 1 to be $X_j \leq a$, outcome 2 to be $a < X_j \leq b$, and outcome 3 to be $X_j > b$, then the counts $(U, V, n-U-V)$ meet this definition exactly, with $p_1 = \Pr(X_j \leq a) = F_X(a)$, $p_2 = \Pr(a < X_j \leq b) = F_X(b) - F_X(a)$, $p_3 = \Pr(X_j > b) = 1 - F_X(b)$. Since the pmf of a multinomial distribution with $r = 3$ is

$$\Pr((N_1, N_2, N_3) = (n_1, n_2, n_3)) = \binom{n}{n_1, n_2, n_3}p_1^{n_1}p_2^{n_2}p_3^{n_3} = \frac{n!}{n_1!n_2!n_3!}p_1^{n_1}p_2^{n_2}p_3^{n_3}$$

we have in this case

$$\Pr(U = u, V = v, n - U - V = n - u - v) = \frac{n!}{u!v!(n-u-v)!} F_X(a)^u (F_X(b) - F_X(a))^v (1 - F_X(v))^{n-u-v}$$

as desired.

□

Definition 1.1.3 (**Median; Math 505A defintion**)**.** The real number m is called a **median** of a random variable X if

$$\Pr(X \leq m) \geq 1/2, \qquad \Pr(X \geq m) \geq 1/2.$$

Proposition 1.1.0.5 (**Math 505A homework problem**)**.** (a) Every random variable has at least one median.

(b) The set of all medians is a closed interval of the real line.

Proof. (a) Suppose the cdf of X, $F_X : \mathbb{R} \to [0, 1]$, is continuous. Because $F_X(x) = \Pr(X \leq x)$ is a cdf, it is also monotonically increasing. By the Intermediate Value Theorem, there exists at least one $m \in \mathbb{R}$ such that $F_X(m) = 1/2$. Because $\Pr(X \leq m) = 1/2 \geq 1/2$ and $\Pr(X \geq m) = 1 - \Pr(X < m) = 1 - 1/2 \geq 1/2$, m is a median.

Suppose F_X is not continuous. If it contains $1/2$ in its range, then m such that $F_X(m) = 1/2$ is a median. If there is no $m \in \mathbb{R}$ such that $F_X(m) = 1/2$, then $m = \inf(\{x \mid F_X(x) \geq 1/2\})$ is a median. To see why, note first that $\Pr(X \leq m) = F_X(m) \geq 1/2$. Second,

$$\Pr(X \geq m) = 1 - \Pr(X < m) = 1 - \lim_{x \to m^-} F_X(x) \geq 1 - 1/2 = 1/2$$

because F_X is right continuous. Therefore m is a median of X.

(b) We show that all medians of X must be in one interval by contradiction. Suppose a and b are medians of X but c is not, where $a < c < b$. By the definition of median, $F_X(a) \geq 1/2$ and $F_X(b) \geq 1/2$. Because c is not a median, $F_X(c) < 1/2$. This implies that F_X is decreasing on the interval from a to c, which contradicts the fact that the distribution function F_X monotonically increases.

Finally we prove that all medians of X are in a closed interval. Let $\mathcal{A}$ be the set of all medians of X; that is, $\mathcal{A} = \{x \mid \Pr(X \leq x) \geq 1/2, \Pr(X \geq x) \geq 1/2\} = \{x \mid F_X(x) \geq 1/2, \lim_{y \to x^-} F_X(y) \leq 1/2\}$. We will show that $\mathcal{A}$ contains its infimum and its supremum. The argument above shows that $a = \inf(\{x \mid F_X(x) \geq 1/2\})$ satisfies $\lim_{y \to a^-} F_X(y) \leq 1/2$; that is, $a \in \mathcal{A}$. Since there is no lower value k which satisfies $F_X(k) \geq 1/2$, $a = \inf(\mathcal{A})$, so $\mathcal{A}$ contains its infimum.

Let $b = \sup\{x \mid \lim_{y \to x^-} F_X(y) \leq 1/2\}$. Because b is the supremum of a set containing a, $b \geq a$. Therefore because F_X is nondecreasing, $F_X(b) \geq F_X(a) \geq 1/2$, which shows that $b \in \mathcal{A}$. Since b is the supremum of the set of all values satisfying $\lim_{y \to x^-} F_X(y) \leq 1/2$, b is the supremum of $\mathcal{A}$. Therefore $\mathcal{A}$ contains its infimum and supremum, and the set of all medians of X is closed.

□

Remark 1. One example of a random variable which has a median of length L: X is a discrete random variable with the following mass function:

$$\Pr(X = 0) = 0.5$$

$$\Pr(X = L) = 0.5$$

Then $m \in [0, L]$ are medians.

1.2 Random Samples

Definition 1.2.1 (**Random Sample**)**.** Let $n > 0$ be an integer. A **random sample** of size n is a sequence $X_1, \ldots, X_n$ of independent identically distributed random variables.

Definition 1.2.2 (**Statistic**)**.** Let n, k be positive integers. Let $X_1, \ldots, X_n$ be a random sample of size n. Let $f : \mathbb{R}^n \to \mathbb{R}^k$ be a (measureable) function. A **statistic** is a random variable of the form $Y := f(X_1, \ldots, X_n)$. The distribution of Y is called a **sampling distribution.**

Definition 1.2.3 (**Sample mean**)**.** The **sample mean** of a random sample $X_1, \ldots, X_n$ of size n, denoted $\overline{X}$, is the following statistic:

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

Proposition 1.2.0.1. Suppose we have a random sample of size n from an i.i.d. distribution $X_1, X_2, \ldots, X_n$ with $\mathbb{E}(X_1) = \mu \; in \mathbb{R}$, $\text{Var}(X_1) = \sigma^2 < \infty$. Then

(a) $\mathbb{E}(\overline{X}) = \mathbb{E}(X_1)$.

(b) $\text{Var}(\overline{X}) = \sigma^2/n$.

Proof. (a)

$$\mathbb{E}(\overline{X}) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}X_i = \frac{1}{n} \cdot n\mu = \mu$$

(b) Using the independence of the X_i,

$$\text{Var}(\overline{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) = \frac{1}{n^2} \cdot n\sigma^2 = \boxed{\frac{\sigma^2}{n}}$$

□

Proposition 1.2.0.2 (**Stats 100B homework problem**)**.** Suppose that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are two samples, with $X \sim \mathcal{N}(\mu_1, \sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2)$. The difference between the sample means, $\bar{X} - \bar{Y}$, is then a linear combination of $m + n$ normal random variables. Then

a. $\mathbb{E}(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$.

b. $\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$.

c. The distribution of $\bar{X} - \bar{Y}$ is normal with mean and variance equal to the previous results.

Proof. a.

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i,\ \bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

$$\mathbb{E}(\bar{X} - \bar{Y}) = \mathbb{E}\Big(\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{n}\sum_{j=1}^{n} Y_j\Big) = \frac{1}{m}\mathbb{E}\Big(\sum_{i=1}^{m} X_i\Big) - \frac{1}{n}\mathbb{E}\Big(\sum_{j=1}^{n} Y_j\Big)$$

$$= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(X_i) - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}(Y_j) = \frac{1}{m}\sum_{i=1}^{m}\mu_1 - \frac{1}{n}\sum_{j=1}^{n}\mu_2 = \frac{1}{m}m\cdot\mu_1 - \frac{1}{n}n\cdot\mu_2$$

$$\implies \mathbb{E}(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$$

b. Since X and Y are independent,

$$\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y})$$

$$= \mathbb{E}\big[(\bar{X} - \mathbb{E}[\bar{X}])^2\big] + \mathbb{E}\big[(\bar{Y} - \mathbb{E}[\bar{Y}])^2\big]$$

$$= \mathbb{E}\Big(\frac{1}{m}\sum_{i=1}^{m} X_i - \mu_1\Big)^2 + \mathbb{E}\Big(\frac{1}{n}\sum_{j=1}^{n} Y_j - \mu_2\Big)^2$$

$$= \mathbb{E}\Big(\frac{1}{m}\sum_{i=1}^{m}\big(X_i - m\frac{1}{m}\mu_1\big)\Big)^2 + \mathbb{E}\Big(\frac{1}{n}\sum_{j=1}^{n}\big(Y_j - n\frac{1}{n}\mu_2\big)\Big)^2$$

$$= \frac{1}{m^2}\mathbb{E}\Big(\sum_{i=1}^{m}\big(X_i - \mu_1\big)\Big)^2 + \frac{1}{n^2}\mathbb{E}\Big(\sum_{j=1}^{n}\big(Y_j - \mu_2\big)\Big)^2$$

Since X_i and X_j are independent for $i \neq j$ (and likewise for Y), $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$, so

$$\mathbb{E}[(X_i - \mu_1)(X_j - \mu_1)] = 0$$

for $i \neq j$ (and likewise for Y). Therefore the above equation can be written as

$$\frac{1}{m^2}\mathbb{E}\Big(\sum_{i=1}^{m}\big(X_i - \mu_1\big)^2\Big) + \frac{1}{n^2}\mathbb{E}\Big(\sum_{j=1}^{n}\big(Y_j - \mu_2\big)^2\Big)$$

$$\frac{1}{m^2}\sum_{i=1}^{m}\mathbb{E}\big(X_i - \mu_1\big)^2 + \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\big(Y_j - \mu_2\big)^2$$

$$= \frac{1}{m^2}\Big(\sum_{i=1}^{m}\sigma_1^2\Big) + \frac{1}{n^2}\Big(\sum_{j=1}^{n}\sigma_2^2\Big) = \frac{1}{m^2}m\cdot\sigma_1^2 + \frac{1}{n^2}n\cdot\sigma_2^2$$

$$\implies \text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$$

c.

$$M_{X_i}(t) = \exp\left(\mu_1 t + \frac{t^2\sigma_1^2}{2}\right),\ M_{Y_i}(t) = \exp\left(\mu_2 t + \frac{t^2\sigma_2^2}{2}\right)$$

Since individual observations from X and Y are independent,

$$M_{\bar{X}}(t) = \prod_{i=1}^{m} M_{X_i}\left(\frac{1}{m}t\right),\ M_{\bar{Y}}(t) = \prod_{j=1}^{n} M_{Y_j}\left(\frac{1}{n}t\right)$$

and

$$M_{\bar{X}-\bar{Y}}(t) = M_{\bar{X}}(t)M_{-\bar{Y}}(t) = M_{\bar{X}}(t)M_{\bar{Y}}(-t) = \prod_{i=1}^{m} M_{X_i}\left(\frac{1}{m}t\right)\prod_{j=1}^{n} M_{Y_j}\left(\frac{-1}{n}t\right)$$

$$= \left[M_{X_i}\left(\frac{t}{m}\right)\right]^m\left[M_{Y_j}\left(\frac{-t}{n}\right)\right]^n = \left[\exp\left(\frac{\mu_1 t}{m} + \frac{t^2\sigma_1^2}{2m^2}\right)\right]^m\left[\exp\left(\frac{-\mu_2 t}{n} + \frac{(-t)^2\sigma_2^2}{2n^2}\right)\right]^n$$

$$= \exp\left(\frac{m\mu_1 t}{m} + \frac{mt^2\sigma_1^2}{2m^2}\right)\exp\left(\frac{-n\mu_2 t}{n} + \frac{nt^2\sigma_2^2}{2n^2}\right)$$

$$\implies \boxed{M_{\bar{X}-\bar{Y}}(t) = \exp\left[(\mu_1-\mu_2)t + \frac{1}{2}t^2\left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)\right]}$$

This is the moment generating function of a normal distribution with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$, consistent with the results from parts (a) and (b).

□

Definition 1.2.4 (**Sample variance**)**.** Let $n > 1$. The **sample variance** of a random sample $X_1, \ldots, X_n$ of size n, denoted S^2, is the following statistic:

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

The **sample standard deviation** of a random sample of size n is $\sqrt{S^2}$.

Proposition 1.2.0.3 (**Unbiasedness of sample variance**)**.** Suppose we have a random sample of size n from an i.i.d. distribution $X_1, X_2, \ldots, X_n$ with $\mathbb{E}(X_1) = \mu\ in\mathbb{R}$, $\text{Var}(X_1) = \sigma^2 < \infty$. Then $\mathbb{E}(S^2) = \sigma^2$. Further, S^2 is a consistent estimator of σ^2.

Proof. We have

$$\mathbb{E}(S^2) = \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n} X_i^2 - 2X_i\overline{X} + \overline{X}^2\right)$$

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

$$= \frac{n}{n-1}\big(\mathbb{E}(X_i^2) - \mathbb{E}\overline{X}^2\big) = \frac{n}{n-1}\big(\text{Var}(X_i) + \mathbb{E}(X_i)^2 - [\text{Var}(\overline{X}) + \mathbb{E}(\overline{X})^2]\big)$$

Using the results from Proposition 1.2.0.1, we have

$$\mathbb{E}(S^2) = \frac{n}{n-1}\big(\sigma^2 + \mu^2 - [\sigma^2/n + \mu^2]\big) = \frac{n}{n-1} \cdot \frac{(n-1)\sigma^2}{n} = \boxed{\sigma^2}$$

□

Alternative proof from Stats 100B homework.

$$\mathbb{E}\bigg(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\bigg) = \frac{1}{n}\mathbb{E}\bigg(\sum_{i=1}^{n}(x_i - \mu)^2\bigg)$$

Assuming independence of samples, this can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}((x_i - \mu)^2) = \frac{1}{n}n\sigma^2 = \boxed{\sigma^2}$$

Since S^2 is unbiased, it is a consistent estimator if we can show $\text{Var}(S^2) \to 0$ as $n \to \infty$. We have

$$\text{Var}(\frac{(n-1)S^2}{\sigma^2}) = 2(n-1)$$

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

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

$$\implies \lim_{n\to\infty}\text{Var}(S^2) = \lim_{n\to\infty}\frac{2\sigma^4}{n-1} = \boxed{0}$$

Therefore S^2 is a consistent estimator of σ^2.

□

Proposition 1.2.0.4 (**Example 11.2.6 in Lehmann and Romano [2005]**).

$$S^2 \xrightarrow{p} \sigma^2.$$

Proof. By the Weak Law of Large Numbers, $\overline{X}_n \xrightarrow{p} \mu$ and $n^{-1}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathbb{E}(X_1)^2 = \mu^2 + \sigma^2$. Therefore by one of Slutsky's convergence theorems (Theorem **??**) and the Continuous Mapping Theorem,

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X}_n)^2 \xrightarrow{p} \sigma^2 \iff \frac{n-1}{n} S^2 \xrightarrow{p} \sigma^2 \iff S^2 \xrightarrow{p} \sigma^2.$$

□

Lemma 1.2.0.5. Let $X := (X_1, \ldots, X_n)$ be i.i.d. mean zero, variance 1 Gaussian random variables. Let $v_1, \ldots, v_n \in \mathbb{R}^n$. Then $\langle X, v_1\rangle, \ldots, \langle X, v_n\rangle$ are independent if and only if $v_1, \ldots, v_m$ are pairwise orthogonal; that is, $\langle v_i, v_j\rangle = 0 \; \forall \; 1 \leq i < j \leq m$.

Proof. By Theorem **??**, we have that for any $v \in \mathbb{R}^n$, $\langle X, v\rangle$ is a mean zero Gaussian with variance $\langle v, v\rangle$. For notational convenience, let $\langle X, v_k\rangle = A_k$. Because all the A_k are Gaussian random variables by Theorem **??**, the A_k are uncorrelated if and only if they are independent. That is, we would like to show that their covariances

$$\mathbb{E}[(A_k - \mathbb{E}A_k)(A_\ell - \mathbb{E}A_\ell)]$$

equal zero for all $\{(k, \ell) : k, \ell \in \{1, 2, \ldots, m\}, k \neq \ell\}$ if and only if the vectors $v_1, \ldots, v_m$ are pairwise orthogonal; that is, $\langle v_k, v_\ell\rangle = 0$ for all $\{(k, \ell) : k, \ell \in \{1, 2, \ldots, m\}, k \neq \ell\}$. Note that since $A_k = \sum_{i=1}^n X_i v_{ki}$, $\mathbb{E}(A_k) = \sum_{i=1}^n v_{ki}\mathbb{E}(X_i)$. So for any $\{(k, \ell) : k, \ell \in \{1, 2, \ldots, m\}, k \neq \ell\}$ we have

$$\mathbb{E}[(A_k - \mathbb{E}A_k)(A_\ell - \mathbb{E}A_\ell)] = \mathbb{E}\Big[\Big(\sum_{i=1}^n X_i v_{ki} - \sum_{i=1}^n v_{ki}\mathbb{E}(X_i)\Big)\Big(\sum_{i=1}^n X_i v_{\ell i} - \sum_{i=1}^n v_{\ell i}\mathbb{E}(X_i)\Big)\Big]$$

$$= \mathbb{E}\Big[\Big(\sum_{i=1}^n X_i v_{ki}\Big)\Big(\sum_{i=1}^n X_i v_{\ell i}\Big) - \Big(\sum_{i=1}^n X_i v_{ki}\Big)\Big(\sum_{i=1}^n v_{\ell i}\mathbb{E}(X_i)\Big)$$

$$-\Big(\sum_{i=1}^n v_{ki}\mathbb{E}(X_i)\Big)\Big(\sum_{i=1}^n X_i v_{\ell i}\Big) + \Big(\sum_{i=1}^n v_{ki}\mathbb{E}(X_i)\Big)\Big(\sum_{i=1}^n v_{\ell i}\mathbb{E}(X_i)\Big)\Big]$$

$$= \mathbb{E}\Big(\sum_{i=1}^n X_i^2 v_{ki} v_{\ell i} + \sum_{\{a,b\in\{1,\ldots,n\}, a\neq b\}} X_a X_b v_{ka} v_{\ell b}\Big) - 2\mathbb{E}\Big(\sum_{i=1}^n X_i \mathbb{E}(X_i) v_{ki} v_{\ell i} + \sum_{\{a,b\in\{1,\ldots,n\}, a\neq b\}} X_a \mathbb{E}(X_b) v_{ka} v_{\ell b}\Big)$$

$$+\mathbb{E}\Big(\sum_{i=1}^n \mathbb{E}(X_i)^2 v_{ki} v_{\ell i} + \sum_{\{a,b\in\{1,\ldots,n\}, a\neq b\}} \mathbb{E}(X_a)\mathbb{E}(X_b) v_{ka} v_{\ell b}\Big)$$

Recall that $\mathbb{E}(X_i) = 0$ for all i. Also, due to independence of the X_i, all of the terms that involve $\mathbb{E}(X_a X_b),\ a \neq b$ disappear. This leaves only

$$= \mathbb{E}\Big(\sum_{i=1}^n X_i^2 v_{ki} v_{\ell i}\Big) = \sum_{i=1}^n \mathbb{E}(X_i^2) v_{ki} v_{\ell i} = \mathbb{E}(X_1^2)\sum_{i=1}^n v_{ki} v_{\ell i} \tag{1.1}$$

where the last step follows from the i.i.d. distributions of X_i. Recall

$$\langle v_k, v_\ell \rangle = 0 \iff \sum_{i=1}^{n} v_{ki} v_{\ell i} = 0.$$

Since $\mathbb{E}(X_i^2) \neq 0$, (1.1) equals 0 for all $\{(k,\ell) : k, \ell \in \{1, 2, \ldots, m\}, k \neq \ell\}$ if and only if $\langle v_k, v_\ell \rangle = 0$ for all $\{(k,\ell) : k, \ell \in \{1, 2, \ldots, m\}, k \neq \ell\}$. Therefore the random variables $\langle X, v_1 \rangle, \ldots, \langle X, v_m \rangle$ are independent if and only if the vectors $v_1, \ldots, v_m$ are pairwise orthogonal.

□

Proposition 1.2.0.6 (**Proposition 4.7 in 541A notes**)**.** Let $n \geq 2$ be an integer. Let $X_1, \ldots, X_n$ be a random sample from the Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Let $\overline{X}$ be the sample mean and let S be the sample standard deviation. Then

(i) $\overline{X}$ and S are independent random variables.

(ii) $\overline{X}$ is a Gaussian random variable with mean μ and variance σ^2/n.

(iii) $(n-1)S^2/\sigma^2$ is a χ^2-distributed random variable with $n-1$ degrees of freedom.

Proof. (i) Replace $X_1, \ldots, X_n$ with $X_1 - \mu, \ldots, X_n - \mu$ so that $\mu = 0$. Also divide by σ so that $\sigma = 1$. Note that $\overline{X}$ is independent of all random variables $X_2 - \overline{X}, \ldots, X_n - \overline{X}$ by Lemma 1.2.0.5 because for example

$$X_2 - \overline{X} = \langle X_2, e_2 - \frac{1}{n}(1, 1, \ldots, 1) \rangle$$

where the second vector in the inner product is orthogonal to $(1, 1, \ldots, 1)$ (in fact, $(1, \ldots, 1)$ is orthogonal to anything in the span of these vectors). Likewise for all the remaining vectors you could use to construct X_i. (Note that the other random variables [e.g. $X_2 - \overline{X}$ and $X_3 - \overline{X}$] are not independent.)

So the proof will be complete if we can write S as a function of $X_2 - \overline{X}, \ldots, X_n - \overline{X}$. Observe

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 = (X_1 - \overline{X})^2 + \sum_{i=2}^{n}(X_i - \overline{X})^2 = \left(n\overline{X} - \left[\sum_{i=2}^{n} X_i\right] - \overline{X}\right)^2 + \sum_{i=2}^{n}\left(X_i - \overline{X}\right)^2$$

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

(ii) Follows from Proposition 1.45, Example 1.108, and Exercise 1.58 in 541A notes (condense later?)

If $n = 2$, we have

$$S^2 = \frac{1}{2-1}\sum_{i=1}^{2}(X_i - \overline{X}_2)^2 = \left(X_1 - \frac{X_1 + X_2}{2}\right)^2 + \left(X_2 - \frac{X_1 + X_2}{2}\right)^2$$

$$= \left(\frac{X_1 - X_2}{2}\right)^2 + \left(\frac{X_2 - X_1}{2}\right)^2 = 2 \cdot \frac{1}{4} \cdot (X_1 - X_2)^2 = \left[\frac{1}{\sqrt{2}}(X_1 - X_2)\right]^2$$

Since $\{X_1 - X_2\} \sim \mathcal{N}(0, 2\sigma^2)$, we have

$$\frac{1}{\sqrt{2}\sigma}(X_1 - X_2) \sim \mathcal{N}(0, 1),$$

so

$$S^2 = \frac{1}{\sigma^2}\left[\frac{1}{\sqrt{2}}(X_1 - X_2)\right]^2 \sim \chi^2_1.$$

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

Next we will induct on n. Some tedious algebra shows that

$$\frac{nS^2_{n+1}}{\sigma^2} = \frac{n-1}{\sigma^2}S^2_n + \frac{1}{\sigma^2}\frac{n}{n+1}(X_{n+1} - \overline{X}_n)^2, \qquad \forall n \geq 2,$$

where S^2_n is the sample variance for n independent Gaussian random variables. It can also be shown through simple algebra that

$$\{(X_{n+1} - \overline{X}_n)\} \sim \mathcal{N}(0, \sigma^2(n+1)/n),$$

so

$$\frac{1}{\sigma^2}\frac{n}{n+1}(X_{n+1} - \overline{X}_n)^2 \sim \chi^2_1.$$

It can be shown by first principles (i.e., without Basu's Theorem) that $\overline{X}_n \perp\!\!\!\perp S^2_n$, and clearly $X_{n+1} \perp\!\!\!\perp S^2$. Therefore we have shown that if $(n-1)S^2_n/\sigma^2 \sim \chi^2_{n-1}$, it follows that $nS^2_{n+1}/\sigma^2 \sim \chi^2_n$, concluding the proof that

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

To show that S^2 is ancillary for μ, simply observe that

$$S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1};$$

that is, the distribution of S^2 depends only on n and σ^2, not μ. Therefore by definition S^2 is ancillary for μ.

(iii) Like above, replace $X_1, \ldots, X_n$ with $X_1 - \mu, \ldots, X_n - \mu$ so that $\mu = 0$. Also divide by σ so that $\sigma = 1$. We will prove by induction. Let $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ and let $S^2_n := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. In the case $n = 2$ we have

$$S^2_2 = \left(X_1 - \frac{1}{2}(X_1 + X_2)\right)^2 + \left(X_2 - \frac{1}{2}(X_1 + X_2)\right)^2 = \frac{1}{4}(X_2 - X_1)^2 + \frac{1}{4}(X_2 - X_1)^2 = \frac{1}{2}(X_2 - X_1)^2$$

$$= \left(\frac{1}{\sqrt{2}}(X - 2 - X_1)\right)^2$$

Note that $1/\sqrt{2}(X_2 - X_1)$ is a mean zero Gaussian random variable with variance 1 (see example 1.108 in 541A notes for details). So S_2^2 is χ_1^2 by Definition 1.33 in 541A notes.

We now induct on n. From Lemma 4.8 in 541A notes (will prove later),

$$nS_{n+1}^2 = (n-1)S_n^2 + \frac{n}{n+1}(X_{n+1} - \overline{X}_n)^2, \quad \forall n \geq 2$$

From the first item, S_n is independent of $\overline{X}_n$. Also, X_{n+1} is independent of S_n by Proposition 1.61 in Math 541A notes, since S_n is a function of $X_1, \ldots, X_n$, which are independent of X_{n+1}. So S_n is independent of $(X_{n+1} - \overline{X}_n)^2$. By the inductive hypothesis, $(n-1)S_n^2$ is a χ_{n-1}^2 random variable. From Example 1.108 in Math 541A notes, $X_{n+1} - \overline{X}_n$ is a Gaussian random variable with mean zero and variance $1 + 1/n = (n+1)/n$ so that $\sqrt{n/(n+1)}(X_{n+1} - \overline{X}_n)$ is a mean zero Gaussian with variance 1, implying $n/(n+1)(X_{n+1} - \overline{X}_n)^2$ is χ^2. Definition 1.33 in 541A notes then implies that nS_{n+1} is a χ_n^2 random variable, completing the inductive step.

□

Lemma 1.2.0.7 (Lemma 4.8 in 541A notes.).

Let $X_1, X_2, \ldots$ be random variables. For any $n \geq 2$, let $\overline{X}_n := (1/n)\sum_{i=1}^n X_i$ and let $S_n^2 := 1/(n-1)\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Then

$$nS_{n+1}^2 - (n-1)S_n^2 = \frac{n}{n+1}(X_{n+1} - \overline{X}_n)^2.$$

Proof.

$$nS_{n+1}^2 - (n-1)S_n^2 = \sum_{i=1}^{n+1}(X_i - \overline{X}_{n+1})^2 - \sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Note:

$$(a-b)^2 - (a-c)^2 = a^2 - 2ab + b^2 - a^2 - c^2 + 2ac = b^2 - c^2 + 2a(c-b)$$

$$= (b-c)[(b+c) - 2a] = (b-c)(b+c-2a)$$

for all real a, b, c. Using $a = X_n, b = \overline{X}_{n+1}, c = \overline{X_n}$ we have

$$= (X_{n+1} - \overline{X}_{n+1})^2 + \sum_{i=1}^{n}(\overline{X}_{n+1} - \overline{X}_n)(\overline{X}_{n+1} + \overline{X}_n - 2X_i)$$

$$= (X_{n+1} - \overline{X}_{n+1})^2 + (\overline{X}_{n+1} - \overline{X}_n)\sum_{i=1}^{n}(\overline{X}_{n+1} + \overline{X}_n - 2X_i)$$

$$= (X_{n+1} - \overline{X}_{n+1})^2 + (\overline{X}_{n+1} - \overline{X}_n) \cdot n(\overline{X}_{n+1} + \overline{X}_n - 2\overline{X}_n)$$

$$= (X_{n+1}(1 - 1/(n+1)) - \frac{n}{n++1}\overline{X}_n)^2 + n(\overline{X}_{n+1} - \overline{X}_n)^2$$

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

Algebra: $1/(n+1) - 1/n = \frac{n-(n+1)}{n(n+1)} = -\frac{1}{n(n+1)}$. So we have

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

$$= \frac{n^2}{(n+1)^2}(X_{n+1} - \overline{X}_n)^2 + \frac{n}{(n+1)^2}(X_{n+1} - \overline{X}_n)^2$$

$$= \frac{n^2+n}{(n+1)^2}(X_{n+1} - \overline{X}_n)^2 = \frac{n}{n+1}(X_{n+1} - \overline{X}_n)^2$$

□

Proposition 1.2.0.8 (**Proposition 4.9 in 541A notes**)**.** Let X be a standard Gaussian random variable. Let Y be a χ^2_p random variable. Assume that X and Y are independent. Then $X/\sqrt{Y/p}$ has the following density, known as **Student's t-distribution** with $p =$ degrees of freedom: ($p = n + 1$?)

$$f_{X/(Y/\sqrt{p})}(t) := \frac{\Gamma((p+1)/2)}{\sqrt{p}\sqrt{\pi}\Gamma(p/2)}\left(1 + \frac{t^2}{p}\right)^{-(p+1)/2}, \quad \forall t \in \mathbb{R}$$

(should have $p + 1$ in a bunch of the expressions above? that's what was written on board, not in notes.)

Proof. Let $Z := \sqrt{Y/p}$. We find the density of Z as follows. Let $t > 0$. Then

$$f_Z(y) = \frac{\mathrm{d}}{\mathrm{d}y}\big|_{y=0} \Pr(Z \leq y) = \frac{\mathrm{d}}{\mathrm{d}y}\big|_{y=0} \Pr(Y \leq y^2 p)$$

$$= \frac{\mathrm{d}}{\mathrm{d}y}\big|_{y=0} \int_0^{y^2 p} \frac{x^{(p/2)-1}e^{-x/2}}{2^{p/2}\Gamma(p/2)}dx = 2yp \cdot p^{(p/2)-1}y^{p-2}e^{-y^2 p/2} \cdot \frac{1}{2^{p/2}\Gamma(p/2)}$$

$$= p^{p/2}y^{p-1}e^{-y^2 p/2} \cdot \frac{1}{2^{p/2-1}\Gamma(p/2)}$$

$\vdots$ skipped this stuff in class proof

$$\Pr(X/Z \leq t) = \Pr(X \leq tZ)$$

$$= \text{ (by definition of joint density) } \int\int_{\{(x,y)\in\mathbb{R}^2 : x\leq ty\}} f_X(x) f_Z(y) dxdy$$

We use the change of variables formula:

$$\int\int_{\phi(U)} f(x,y)dxdy = \int\int_U f(\phi(a,b))|\operatorname{Jac}\phi(a,b)|dadb$$

$$\phi : \mathbb{R}^2 \to \mathbb{R}^2$$

$$\phi(a,b) = (ab, a)$$

$$\phi^{-1}(x,y) = (y, x/y)$$

We chose x/y as the second variable so that an upper limit of the variable will end up being t after the transformation. We need the Jacobian of ϕ:'

$$|\operatorname{Jac}\phi(a,b)| = \left|\det\begin{pmatrix} b & a \\ 1 & 0 \end{pmatrix}\right| = |a|$$

By the change in variables formula,

$$\int\int_{\phi(U)} f(x,y)dxdy = \int\int_U f(\phi(a,b))\,|\operatorname{Jac}\phi(a,b)|\,dadb$$

$$= \int\int_{\{(a,b)\in\mathbb{R}^2 : a\geq 0, b\leq t\}} f_X(ab) f_Z(a)\,|a|\,dadb$$

$$\implies \Pr(X/Z \leq t) = \int_{-\infty}^{t}\int_0^{\infty} |a| f_X(ab) f_Z(a) dadb$$

By the Fundamental Theorem of Calculus,

$$f_{X/Z}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\Pr(X/Z \leq t) = \int_0^{\infty} |a| f_X(at) f_Z(a)) da = \int_0^{\infty} a f_X(at) f_Z(a)) da$$

By the definitions of X and Z,

$$= \frac{1}{2^{-/2-1}\Gamma(p/2)}\int_0^{\infty} a \cdot \frac{1}{\sqrt{2\pi}} e^{-(a^2t^2)/2} \cdot p^{p/2} a^{p-1} e^{-a^2 p/2} da$$

$$= \frac{p^{p/2}}{2^{-/2-1}\Gamma(p/2)\sqrt{2\pi}}\int_0^{\infty} e^{-[a^2(t^2+p)]/2} \cdot a^p da$$

Change of variables: let $x = a^2, dx = 2ada, da = \frac{1}{2a}dx = 1/(2\sqrt{x})dx$. Then this integral is

$$= c\int_0^\infty e^{-[x(t^2+p)]/2} \cdot x^{p/2-1/2}da, \quad \text{where } c = \frac{p^{p/2}}{2^{p/2}\sqrt{2\pi}\Gamma(p/2)}$$

So the integrand is a Gamma density function with parameters α, β: $\alpha - 1 = p/2 - 1/2 \iff \alpha = p/2 + 1/2$, $\beta = 2/(t^2+p)$. So if we multiply and divide $\beta^\alpha\Gamma(\alpha)$

So

$$f_{X/Z}(t) = \frac{p^{p/2}}{2^{p/2}\sqrt{2\pi}\Gamma(p/2)} \cdot \beta^\alpha\Gamma(\alpha) \cdot 1 = \frac{p^{p/2}\Gamma((p_1)/2)}{2^{p/2}\sqrt{2\pi}\Gamma(p/2)} \cdot \left(\frac{2}{t^2+p}\right)^{(p-1)/2}$$

$$= \frac{p^{p/2}\Gamma((p+1)/2)}{\sqrt{\pi}\Gamma(p/2)} \cdot (t^2+p)^{-(p+1)/2} = \frac{\Gamma((p+1)/2)}{\sqrt{\pi p}\Gamma(p/2)} \cdot (1+t^2/p)^{-(p+1)/2}$$

□

Remark 2 (Remark 4.10 in 541A notes)**.** If $X_1, \ldots, X_n$ is a random sample from a Gaussian distribution with mean $\mu \in \mathbb{R}$, standard deviation $\sigma < 0$, then

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

also has Student's t distribution. ($\overline{X} := n^{-1}\sum_{i=1}^n X_i, S = \sqrt{(n-1)^{n-1}\sum_{i=1}^n (X_i - \overline{X})^2}$.)

Proposition 1.2.0.9 (**Stats 100B homework 3 problem**)**.** Let X_1, X_2 be a random sample from a normal distribution with a mean μ and standard deviation σ. Then $(n-1)s^2/\sigma^2$ has a χ_1^2 distribution.

Proof.

$$s^2 = \frac{1}{2-1}\sum_{i=1}^2 (X_i - \bar{X})^2 = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 = (X_1 - \frac{X_1+X_2}{2})^2 + (X_2 - \frac{X_1+X_2}{2})^2$$

$$= X_1^2 - 2X_1(\frac{X_1+X_2}{2}) + (\frac{X_1+X_2}{2})^2 + X_2^2 - 2X_2(\frac{X_1+X_2}{2}) + (\frac{X_1+X_2}{2})^2$$

$$= X_1^2 + X_2^2 - X_1(X_1+X_2) - X_2(X_1+X_2) + 2(\frac{X_1+X_2}{2})^2$$

$$= X_1^2 + X_2^2 - (X_1+X_2)(X_1+X_2) + \frac{(X_1+X_2)^2}{2}$$

$$= \frac{1}{2}(2X_1^2 + 2X_2^2) - \frac{1}{2}(X_1^2 + 2X_1X_2 + X_2^2)$$

$$= \frac{1}{2}(X_1^2 - 2X_1X_2 + X_2^2)$$

$$\boxed{s^2 = \frac{1}{2}(X_1 - X_2)^2}$$

$$\implies \frac{(n-1)s^2}{\sigma^2} = (2-1)\frac{1}{2\sigma^2}(X_1 - X_2)^2 = \left(\frac{X_1 - X_2}{\sigma\sqrt{2}}\right)^2$$

Since X_1 and X_2 are normal,

$$X_1 - X_2 \sim \mathcal{N}(\mu - \mu, \sqrt{\sigma^2 + \sigma^2}) = \mathcal{N}(0, \sigma\sqrt{2}) \implies \frac{X_1 - X_2}{\sigma\sqrt{2}} \sim \mathcal{N}(0,1)$$

$$\implies \left(\frac{X_1 - X_2}{\sigma\sqrt{2}}\right)^2 = \boxed{\frac{(n-1)s^2}{\sigma^2} \sim \chi_1^2}$$

□

Proposition 1.2.0.10 (**Stats 100B homework problem**)**.** Suppose two independent random samples of n_1 and n_2 observations are selected from two normal populations. Further, assume that the populations possess a common variance σ^2 which is unknown. Let the sample variances be S_1^2 and S_2^2 and assume they are unbiased. Then the pooled estimator for σ^2

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

is unbiased and has variance $\frac{2\sigma^4}{n_1+n_2-2}$.

Proof. First we show S^2 is unbiased.

$$\mathbb{E}(S^2) = \mathbb{E}\left(\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\right) = \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbb{E}(S_1^2) + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbb{E}(S_2^2)$$

$$= \frac{n_1 - 1}{n_1 + n_2 - 2}\sigma^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}\sigma^2 = \frac{(n_1 + n_2 - 2)\sigma^2}{n_1 + n_2 - 2} = \boxed{\sigma^2}$$

Now we derive its variance.

$$\text{Var}(S^2) = \text{Var}\left(\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\right)$$

Since S_1 and S_2 are independent, this can be written as

$$\frac{1}{(n_1+n_2-2)^2}\bigg(\text{Var}[(n_1-1)S_1^2]+\text{Var}[(n_2-1)S_2^2]\bigg)$$

Since the populations are normal, we know

$$\frac{(n_i-1)S_i^2}{\sigma^2}\sim\chi^2_{n_i-1} \implies \text{Var}\bigg(\frac{(n_i-1)S_i^2}{\sigma^2}\bigg)=2(n_i-1)$$

$$\text{Var}(S^2)=\frac{\sigma^4}{(n_1+n_2-2)^2}\bigg(\text{Var}\bigg[\frac{(n_1-1)S_1^2}{\sigma^2}\bigg]+\text{Var}\bigg[\frac{(n_2-1)S_2^2}{\sigma^2}\bigg]\bigg)$$

$$\frac{\sigma^4}{(n_1+n_2-2)^2}(2(n_1-1)+2(n_2-1))=\sigma^4\frac{2(n_1+n_2-2)}{(n_1+n_2-2)^2}$$

$$=\frac{2\sigma^4}{n_1+n_2-2}$$

□

Proposition 1.2.0.11 (Stats 100B Homework problem). Suppose that $X_1,\dots,X_m$ and $Y_1,\dots,Y_n$ are two samples, with $X\sim\mathcal{N}(\mu_1,\sigma_1)$ and $Y\sim\mathcal{N}(\mu_2,\sigma_2)$. The difference between the sample means, $\bar{X}-\bar{Y}$, is then a linear combination of $m+n$ normal random variables.

a. $\mathbb{E}(\bar{X}-\bar{Y})$.

b. $\text{Var}(\bar{X}-\bar{Y})$

c. The distribution of $\bar{X}-\bar{Y}$ is normal.

Proof. a.

$$\bar{X}=\frac{1}{m}\sum_{i=1}^m X_i,\ \bar{Y}=\frac{1}{n}\sum_{j=1}^n Y_j$$

$$\mathbb{E}(\bar{X}-\bar{Y})=\mathbb{E}\bigg(\frac{1}{m}\sum_{i=1}^m X_i-\frac{1}{n}\sum_{j=1}^n Y_j\bigg)=\frac{1}{m}\mathbb{E}\bigg(\sum_{i=1}^m X_i\bigg)-\frac{1}{n}\mathbb{E}\bigg(\sum_{j=1}^n Y_j\bigg)$$

$$=\frac{1}{m}\sum_{i=1}^m\mathbb{E}(X_i)-\frac{1}{n}\sum_{j=1}^n\mathbb{E}(Y_j)=\frac{1}{m}\sum_{i=1}^m\mu_1-\frac{1}{n}\sum_{j=1}^n\mu_2=\frac{1}{m}m\cdot\mu_1-\frac{1}{n}n\cdot\mu_2$$

$$\boxed{\mathbb{E}(\bar{X}-\bar{Y})=\mu_1-\mu_2}$$

b. Since X and Y are independent,

$$\text{Var}(\bar{X}-\bar{Y})=\text{Var}(\bar{X})+\text{Var}(\bar{Y})$$

$$=\mathbb{E}\big[(\bar{X}-\mathbb{E}[\bar{X}])^2\big]+\mathbb{E}\big[(\bar{Y}-\mathbb{E}[\bar{Y}])^2\big]$$

$$= \mathbb{E}\Big(\frac{1}{m}\sum_{i=1}^{m} X_i - \mu_1\Big)^2 + \mathbb{E}\Big(\frac{1}{n}\sum_{j=1}^{n} Y_j - \mu_2\Big)^2$$

$$= \mathbb{E}\Big(\frac{1}{m}\sum_{i=1}^{m} \big(X_i - m\frac{1}{m}\mu_1\big)\Big)^2 + \mathbb{E}\Big(\frac{1}{n}\sum_{j=1}^{n} \big(Y_j - n\frac{1}{n}\mu_2\big)\Big)^2$$

$$= \frac{1}{m^2}\mathbb{E}\Big(\sum_{i=1}^{m} \big(X_i - \mu_1\big)\Big)^2 + \frac{1}{n^2}\mathbb{E}\Big(\sum_{j=1}^{n} \big(Y_j - \mu_2\big)\Big)^2$$

Since X_i and X_j are independent for $i \neq j$ (and likewise for Y), $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$, so

$$\mathbb{E}[(X_i - \mu_1)(X_j - \mu_1)] = 0$$

for $i \neq j$ (and likewise for Y). Therefore the above equation can be written as

$$\frac{1}{m^2}\mathbb{E}\Big(\sum_{i=1}^{m} \big(X_i - \mu_1\big)^2\Big) + \frac{1}{n^2}\mathbb{E}\Big(\sum_{j=1}^{n} \big(Y_j - \mu_2\big)^2\Big)$$

$$\frac{1}{m^2}\sum_{i=1}^{m} \mathbb{E}\big(X_i - \mu_1\big)^2 + \frac{1}{n^2}\sum_{j=1}^{n} \mathbb{E}\big(Y_j - \mu_2\big)^2$$

$$= \frac{1}{m^2}\Big(\sum_{i=1}^{m} \sigma_1^2\Big) + \frac{1}{n^2}\Big(\sum_{j=1}^{n} \sigma_2^2\Big) = \frac{1}{m^2} m \cdot \sigma_1^2 + \frac{1}{n^2} n \cdot \sigma_2^2$$

$$\boxed{\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

c.

$$M_{X_i}(t) = \exp\Big(\mu_1 t + \frac{t^2\sigma_1^2}{2}\Big),\ M_{Y_i}(t) = \exp\Big(\mu_2 t + \frac{t^2\sigma_2^2}{2}\Big)$$

Since individual observations from X and Y are independent,

$$M_{\bar{X}}(t) = \prod_{i=1}^{m} M_{X_i}\Big(\frac{1}{m}t\Big),\ M_{\bar{Y}}(t) = \prod_{j=1}^{n} M_{Y_j}\Big(\frac{1}{n}t\Big)$$

and

$$M_{\bar{X}-\bar{Y}}(t) = M_{\bar{X}}(t)M_{-\bar{Y}}(t) = M_{\bar{X}}(t)M_{\bar{Y}}(-t) = \prod_{i=1}^{m} M_{X_i}\Big(\frac{1}{m}t\Big)\prod_{j=1}^{n} M_{Y_j}\Big(\frac{-1}{n}t\Big)$$

$$= \Big[M_{X_i}\Big(\frac{t}{m}\Big)\Big]^m\Big[M_{Y_j}\Big(\frac{-t}{n}\Big)\Big]^n = \Big[\exp\Big(\frac{\mu_1 t}{m} + \frac{t^2\sigma_1^2}{2m^2}\Big)\Big]^m\Big[\exp\Big(\frac{-\mu_2 t}{n} + \frac{(-t)^2\sigma_2^2}{2n^2}\Big)\Big]^n$$

$$= \exp\Big(\frac{m\mu_1 t}{m} + \frac{mt^2\sigma_1^2}{2m^2}\Big)\exp\Big(\frac{-n\mu_2 t}{n} + \frac{nt^2\sigma_2^2}{2n^2}\Big)$$

$$\implies \boxed{M_{\bar{X}-\bar{Y}}(t) = \exp\Big[(\mu_1 - \mu_2)t + \frac{1}{2}t^2\Big(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\Big)\Big]}$$

This is the moment generating function of a normal distribution with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$, consistent with the results from parts (a) and (b).

□

1.2.1 The Delta Method

Theorem 1.2.1.1 (Delta Method, Theorem 4.14 in 541A notes, 5.5.24 in Casella and Berger [2001],). Let $\theta \in \mathbb{R}$. Let $Y_1, Y_2, \ldots$ be random variables such that $\sqrt{n}(Y_n - \theta)$ converges in distribution to a mean zero Gaussian random variable with variance $\sigma^2 > 0$. Let $f : \mathbb{R} \to \mathbb{R}$. Assume that f' exists and is continuous, and $f'(\theta) \neq 0$. Then

$$\sqrt{n}(f(Y_n) - f(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(f'(\theta))^2).$$

Proof from new class notes. Since $f'(\theta)$ exists, $\lim_{y \to \theta} \frac{f(y)-f(\theta)}{y-\theta}$ exists. That is, there exists $h : \mathbb{R} \to \mathbb{R}$ such that $\lim_{z \to 0} \frac{h(z)}{z} = 0$ and for all $y \in \mathbb{R}$,

$$f'(\theta) = \frac{f(y) - f(\theta)}{(y - \theta)} + h(y - \theta)$$

$$\iff f(y) = f(\theta) + f'(\theta)(y - \theta) + h(y - \theta).$$

In particular,

$$\sqrt{n}[f(Y_n) - f(\theta)] = \underbrace{f'(\theta)}_{\text{(constant)}} \underbrace{\sqrt{n}(Y_n - \theta)}_{\implies \mathcal{N}(0,\sigma^2)} + \underbrace{\sqrt{n}h(Y_n - \theta)}_{?}. \tag{1.2}$$

where we note that $\sqrt{n}(Y_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ by assumption. Since it is multiplied by $f'(\theta) \in \mathbb{R}$, the product of these two terms converges to $\mathcal{N}(0, \sigma^2[f'(\theta)]^2)$ by Slutsky's Theorem (Theorem **??**(b)). We seek to show what happens to the third term of (1.2) as $n \to \infty$ (the result follows if the term converges in probability to 0). Note that for any $n \geq 1$ and for any $t > 0$,

$$\Pr(\sqrt{n}|h(Y_n - \theta)| > t) = \Pr\left(\sqrt{n}|h(Y_n - \theta)| > t \cap |Y_n - \theta| > \frac{t}{\sqrt{n}}\right) + \Pr\left(\sqrt{n}|h(Y_n - \theta)| > t \cap |Y_n - \theta| \leq \frac{t}{\sqrt{n}}\right)$$

$$\iff \Pr(\sqrt{n}|h(Y_n - \theta)| > t) \leq \Pr(|Y_n - \theta| > t/\sqrt{n}) + \Pr(\sqrt{n}|h(Y_n - \theta)| > t \cap |Y_n - \theta| \leq t/\sqrt{n}). \tag{1.3}$$

Since we already have by assumption $\sqrt{n}(Y_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, it follows that $|Y_n - \theta| \xrightarrow{p} 0$. (For completeness, a detailed argument is included in the below lemma.) It then follows that the second term converges in probability to 0 if $\lim_{n \to \infty} \Pr(|Y_n - \theta| > t/\sqrt{n}) = 0$ because $\lim_{z \to 0} h(z)/z = 0$. Therefore for any $t > 0$,

$$\lim_{n\to\infty} \Pr(\sqrt{n}|h(Y_n - \theta)| > t) = 0 \iff \sqrt{n}|h(Y_n - \theta)| \xrightarrow{p} 0$$

which yields the result by (1.2).

□

Theorem 1.2.1.2 (Delta Method (GSBA 604 presentation, p. 15 of MLE notes)). Let X, Y be two random variables with $Y = H(X)$ where H is smooth. Suppose $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$). Then $\mathbb{E}(Y) = H(\mu)$ and $\text{Var}(Y) = [H'(\mu)]^2\sigma^2$. When $x = \hat{\mu}_{MLE}$ and $Y = \hat{\Xi}$,

$$\text{Var}_\eta(\hat{\xi}) = \frac{\left[\frac{\text{d}}{\text{d}\eta}h(\eta)\right]^2}{n\text{Var}_\eta(\hat{\xi})}.$$

Remark 3. Note that this follows from the Delta Method (Theorem 1.2.1.1) and functional equivariance of the MLE (Proposition 1.4.6.7).

Lemma 1.2.1.3. Under the same assumptions and notation as in Theorem 1.2.1.1,

$$\lim_{n\to\infty} \Pr(|Y_n - \theta| > t/\sqrt{n}) = 0$$

Proof. We will examine the behavior of the right side of (1.3) as $n \to \infty$ by looking at the first term and showing that $Y_n - \theta$ converges in probability to 0. If $t > 0$, then $\Pr(|Y_n - \theta| > t) = \Pr(\sqrt{n}|Y_n - \theta| > t\sqrt{n})$, and if $c > 0$ is a constant, then for sufficiently large n, the last quantity is at most $\Pr(\sqrt{n}|Y_n - \theta| > c)$. So we have

$$\Pr(|Y_n - \theta| > t) = \Pr(\sqrt{n}|Y_n - \theta| > t\sqrt{n}) \leq \Pr(\sqrt{n}|Y_n - \theta| > c)$$

But as $n \to \infty$, c can be any constant (arbitrarily large). So

$$\lim_{n\to\infty} \Pr(\sqrt{n}|Y_n - \theta| > t) \leq \int_c^\infty e^{-y^2/2}\frac{1}{\sqrt{2\pi}}dy.$$

Therefore

$$\lim_{n\to\infty} \Pr(|Y_n - \theta| > t/\sqrt{n}) = 0.$$

□

Theorem 1.2.1.4 (Convergence Theorem with Bounded Moment, Theorem 4.16 in 541A notes.). Let $X_1, X_2, \ldots$ be random variables that converge in distribution to a random variable X. Assume $\exists\ \epsilon > 0, c < \infty$ such that $\mathbb{E}(|X_n|^{1+\epsilon}) \leq c, \ \ \forall\ n \geq 1$. Then

$$\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(X_n).$$

Proof. In Heilman's Graduate Probability Notes, Theorem 1.59 and Exercise 3.8(iii).

□

If $f'(\theta) = 0$ in the Delta Method, we can instead use a second order Taylor expansion as follows.

Theorem 1.2.1.5 (**Second Order Delta Method, Theorem 4.17 in Math 541A Notes.**)**.** Let $\theta \in \mathbb{R}$. Let $Y_1, Y_2, \ldots$ be random variables such that $\sqrt{n}(Y_n - \theta)$ converges in distribution to a mean zero Gaussian random variable with variance $\sigma^2 > 0$. Let $f : \mathbb{R} \to \mathbb{R}$. Assume that f'' exists and is continuous, $f'(\theta) = 0$ and $f''(\theta) \neq 0$. Then

$$n(f(Y_n) - f(\theta))$$

converges in distribution to a χ_1^2 random variable multiplied by $\sigma^2 \frac{1}{2}|f''(\theta)|$ as $n \to \infty$.

Proof. Using a second order Taylor expansion of f, there exists a random Z_n between θ and Y_n such that

$$f(Y_n) = f(\theta) + f'(\theta)(Y_n - \theta) + \frac{1}{2}f''(Z_n)(Y_n - \theta)^2 = f(\theta) + \frac{1}{2}f''(Z_n)(Y_n - \theta)^2 \tag{1.4}$$

where the second equality follows because $f'(\theta) = 0$. As in the proof of Theorem 1.2.1.1, $Z_n \xrightarrow{p} \theta$. Since f'' is continuous, $f''(Z_n)$ converges in probability to $f''(\theta)$ by Proposition 2.36 in the Math 541A notes (Theorem **??**, continuous functions conserve convergence in probability). Therefore using (1.4),

$$n(f(Y_n) - f(\theta)) = \frac{1}{2}f''(Z_n) \cdot n(Y_n - \theta)^2$$

Note that $\sqrt{n}(Y_n - \theta)$ converges in distribution to a mean zero Gaussian random variable by assumption, so $n(Y_n - \theta)^2$ converges in distribution to a χ_1^2 random variable by Proposition 2.36 in the Math 541A notes (Theorem **??**). So since $f''(Z_n)$ converges in probability to a constant, by Proposition 2.36 in the Math 541A notes (Slutsky's Theorem, Theorem **??**), the right side converges in probability to $\frac{1}{2}f''(\theta)\sigma$ multiplied by a χ_1^2 random variable.

□

1.2.2 Simulation of Random Variables

Proposition 1.2.2.1. If $X : \Omega \to \mathbb{R}$ is an arbitrary random variable with cumulative distribution function $F : \mathbb{R} \to [0, 1]$, then the function F^{-1} (if it exists) is a random variable on $[0, 1]$ with the uniform probability law on $(0, 1)$ that is equal in distribution to X.

Proof. Starting with the cdf of $F^{-1}(u)$,

$$\Pr(s \in [0,1] : F^{-1}(s) \leq t) = \Pr(s \in [0,1] : F(t) > s) = F(t) = \Pr(\omega \in \Omega : X(\omega) \leq t)$$

where the third equality uses the definition of a uniform probability law on $(0, 1)$.

□

Remark 4. If F^{-1} does not exist, it can still work if you construct a generalized inverse of F as follows:

Proposition 1.2.2.2 (**Exercise 4.20 in Math 541A notes**)**.** Let $X : \Omega \to \mathbb{R}$ be a random variable on a sample space Ω equipped with a probability law $\mathbb{P}$. For any $t \in \mathbb{R}$ let $F(t) := \mathbb{P}(X \leq t)$. For any $s \in (0, 1)$ define

$$Y(s) := \sup\{t \in \mathbb{R} : F(t) < s\}.$$

So Y is a random variable on $(0, 1)$ with the uniform probability law on $(0, 1)$. Then X and Y are equal in distribution. That is, $\mathbb{P}(Y \leq t) = F(t)$ for all $t \in \mathbb{R}$.

Proof. Note that since F is a cumulative distribution function, F is nondecreasing and F is right-continuous. So we have

$$\sup\{t \in \mathbb{R} : F(t) < s\} = \begin{cases} F^{-1}(s) & \text{if } F \text{ is strictly increasing (i.e. invertible) near } s \\ \inf\{x : F(x) = F(s)\} & \text{if } F \text{ is constant near } s \end{cases}$$

That is, the only time this quantity is different from $F^{-1}(s)$ is when $F^{-1}(\cdot)$ is undefined because F is constant on some interval around s. But if that is the case, $F(\sup\{t \in \mathbb{R} : F(t) < s\}) = F(\inf\{x : F(x) = F(s)\}) = s$ anyway. With that in mind we proceed:

$$\mathbb{P}(Y \leq t) = \mathbb{P}(s \in (0, 1) : Y(s) \leq t) = \mathbb{P}(s \in (0, 1) : \sup\{t' \in \mathbb{R} : F(t') < s\} \leq t)$$

$$= \mathbb{P}(s \in (0, 1) : F\big(\sup\{t' \in \mathbb{R} : F(t') < s\}\big) \leq F(t)\big) = \mathbb{P}(s \in (0, 1) : s \leq F(t)\big)$$

$$= \mathbb{P}(s \in (0, 1) : F(t) > s) = F(t) = \Pr(\omega \in \Omega : X(\omega) \leq t).$$

□

Example 1.2.1 (**Example 4.22 in Math 541A notes**)**.** Let X be an exponential random variable with parameter 1.

$$\Pr(X \leq t) = \int_0^t e^{-x} dx = [-e^{-x}]_0^t = 1 - e^{-t} = F(t)$$

We seek $F^{-1}(t)$:

$$1 - e^{-y} = t \iff e^{-y} = 1 - t \iff -y = \log(1 - t) \iff y = -\log(1 - t) \implies F^{-1}(t) = -\log(1 - t)$$

So to simulate an exponential random variable with parameter 1, sample $-\log(1 - U)$ where $U \sim \mathrm{U}(0, 1)$.

Remark 5. What if the cdf is hard to compute? For example, in a Gaussian distribution:

$$F(t) = \int_{-\infty}^{t} (2\pi)^{-1/2} \exp(-x^2/2) dx.$$

F^{-1} cannot be described using elementary formulas, so $F^{-1}(u)$ is not the best way to simulate a Gaussian random variable. When using the Central Limit Theorem approach (see 541A notes for details), Edgeworth expansion says: if we replace $U_1, \dots, U_n$ with i.i.d. $X_1, \dots, X_n$ and the first m moments of X_1 agree with the first m moments of Gaussian random variables, then the error in the CLT approximation to a Gaussian is $n^{-(m-1)/2}$. (See https://en.wikipedia.org/wiki/Edgeworth_series.) But this is still inefficient, because one Gaussian sample requires n uniform samples.

Proposition 1.2.2.3 (**Box-Muller Algorithm**)**.** Let U_1, U_2 be independent random variable distributed in $(0, 1)$. Define

$$R := \sqrt{-2\log(U_1)}$$

t̂his density is something like $e^{-x^2/2}$

$$\Psi := 2\pi U_2$$

$$X := R\cos(\Psi), \quad Y := R\sin(\Psi)$$

Then X, Y are independent standard Gaussian random variables.

Proof. Homework problem.

□

1.3 Data Reduction

Suppose we have some data and an exponential family. We would like to find the parameter θ among the exponential family that fits the data well. Suppose we have a large data set, maybe so large that you can't store all the data in RAM at once. What is the "least memory" or "most efficient" method for finding θ? The answer: try to find a statistic that captures all the relevant information about θ. For example, to find the mean of a Gaussian sample, use the sample mean. You don't have to store all the raw data, you can just store the sample mean. The following is a generalization of this concept:

1.3.1 Sufficient Statistics

Definition 1.3.1 (**Sufficient Statistic; definition 5.1 in Math 541A notes**)**.** Suppose $X_1, \dots, X_n$ is a sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of distributions (such as an

exponential family). Let $g : \mathbb{R}^n \to \mathbb{R}^k$ so that $Y := g(X_1, \ldots, X_n)$ is a statistic. We say that Y is a **sufficient statistic** for θ if for every $y \in \mathbb{R}^k$ and for every $\theta \in \Theta$, the conditional distribution of $(X_1, \ldots, X_n)$ given $Y = y$ (with respect to probabilities given by f_θ) does not depend on θ. That is, Y provides sufficient information to determine θ from $X_1, \ldots, X_n$.

Remark 6. Based on a comment Heilman made on class, this definition assumes independence of the random variables? Basically everything in this class does?

Definition 1.3.2 (**Sufficiency (Math 541B in-class definition)**). $T(X_1, \ldots, X_n)$ is **sufficient** for θ (more precisely, for the statistical model $\{P_\theta, \theta \in \Theta\}$) if

$$\mathbb{P}_\theta((X_1, \ldots, X_n) \in A \mid T(X_1, \ldots, X_n) = t) = \mu_t(A).$$

In particular, this probability does not depend on θ.

Goldstein lecture: Suppose we have a model $\{f_\theta : \theta \in \Theta\}$ which we interpret as a set of densities or mass functions. We have $\Theta \in \mathbb{R}^p$, and we know the model up to p parameters. Example; we have $X_1, X_2, \ldots, X_n \sim$ i.i.d. f_θ where $\theta \in (\mu, \sigma^2), \mu \in \mathbb{R}$, where $f_\theta \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}.$$

Example 1.3.1 (**Example 5.2 in 541A notes**). Let $X_1, \ldots, X_n$ be a random sample of size n from a Bernoulli distribution with parameter $0 < \theta < 1$. Then $Y := X_1 = \ldots + X_n$ is sufficient for θ.

Proposition 1.3.1.1 (**Example 5.2 in 541A notes**). Let $X_1, \ldots, X_n$ be a random sample of size n from a Bernoulli distribution with parameter $0 < \theta < 1$. Let $Y := X_1 = \ldots + X_n$. Then

$$\mathbb{P}_\theta(X = x \mid Y = y) = \begin{cases} 0 & y \neq \sum_i x_i \\ \Big/ \binom{n}{y} & \sum_i x_i = y \end{cases}$$

Remark 7. If a statistic is sufficient for θ, then we can use that sufficient statistic to re-create the data (or re-create an equivalent data set with the same statistical properties as far we are concerned with estimating the parameter of interest).

Proof. Let $x_1, \ldots, x_n \in [0, 1]$. Let $0 \leq y \leq n$ be an integer. Then Y is binomial with parameters n and θ. We may assume $y = x_1 + \ldots + x_n$, otherwise there is nothing to show. Using the definition of conditional probability,

$$\Pr\big((X_1, \ldots, X_n) = (x_1, \ldots, x_n) \mid Y = y\big) = \frac{1}{\Pr(Y = y)} \cdot \Pr\big((X_1, \ldots, X_n) = (x_1, \ldots, x_n) \cap Y = y\big)$$

$$= \frac{1}{\Pr(Y = y)} \cdot \Pr\big((X_1, \ldots, X_n) = (x_1, \ldots, x_n)\big)$$

Using independence and the definition of a binomial distribution, we have

$$= \frac{1}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} \cdot \prod_{i=1}^{n} \Pr\big(X_i = x_i\big) = \frac{1}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} \cdot \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \frac{1}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} \cdot \theta^y(1-\theta)^{n-y} = \frac{1}{\binom{n}{y}}.$$

Since this expression does not depend on θ, Y is sufficient for θ.

□

Example 1.3.2 (Example 5.3 in 541A notes). Let $X_1, \ldots, X_n$ be a sample of size n from a Gaussian distribution with known variance $\sigma^2 > 0$ and unknown mean $\mu \in \mathbb{R}$. Then $Y := (X_1, \ldots, X_n)/n$ is a sufficient statistic for μ.

Proof. Note that Y is a Gaussian random variable with mean μ and variance σ^2/n. Let $x_1, \ldots, x_n \in \mathbb{R}$ and let $y = (x_1 + \ldots + x_n)/n$. Then

$$f_{X_1,\ldots,X_n|Y}(x_1, \ldots, x_n \mid y) = \frac{1}{f_Y(y)} \cdot f_{X_1,\ldots,X_n,Y}(x_1, \ldots, x_n, y) = \frac{1}{f_Y(y)} \cdot f_{X_1,\ldots,X_n}(x_1, \ldots, x_n, y)$$

Since

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

we have

$$= \frac{1}{f_Y(y)} \cdot \prod_{i=1}^n f_{X_i}(x_i) = \frac{1}{f_Y(y)} \cdot \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n \cdot \exp\left(-\frac{1}{2\sigma^2}(x_1^2 + \ldots + x_n^2) - \frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i \right)$$

$$= \frac{\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n \cdot \exp\left(-\frac{1}{2\sigma^2}(x_1^2 + \ldots + x_n^2) - \frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i \right)}{n^{1/2}(\sigma^2 2\pi)^{-1/2} \exp\left(-\frac{n}{2\sigma^2} y^2 - \frac{n}{2\sigma^2}\mu^2 + \frac{n\mu}{\sigma^2} y \right)}$$

$$= \frac{\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n \cdot \exp\left(-\frac{1}{2\sigma^2}(x_1^2 + \ldots + x_n^2) \right)}{n^{1/2}(\sigma^2 2\pi)^{-1/2} \exp\left(-\frac{n}{2\sigma^2} y^2 \right)}$$

Because μ does not appear in this expression, Y is sufficient for μ.

□

Theorem 1.3.1.2 (Theorem 6.2.2 in Casella and Berger [2001], not in 541A lecture notes). If $p(x \mid \theta)$ is the joint pdf or pmf of a random sample $\boldsymbol{X} = X_1, \ldots, X_n$ and $q(t \mid \theta)$ is the pdf or pmf of the statistic $T(\boldsymbol{X})$, then $T(\boldsymbol{X})$ is a sufficient statistic for θ if, for every $\boldsymbol{x} = x_1, \ldots, x_n$ in the sample space, the ratio $p(\boldsymbol{x} \mid \theta)/q(T(\boldsymbol{x} \mid \theta)$ is constant as a function of θ.

Theorem 1.3.1.3 (**Neyman-Fisher Factorization Theorem, Theorem 5.4 in 541A notes**). Suppose $X = (X_1, \dots, X_n)$ is a random sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of probability density functions or probability mass functions. Let $t : \mathbb{R}^n \to \mathbb{R}^k$, so $Y := t(X_1, \dots, X_n)$ is a statistic. Then Y is sufficient for θ if and only if there exists a nonnegative $\{g_\theta : \theta \in \Theta\}$, $h : \mathbb{R}^n \to \mathbb{R}$, $g_\theta : \mathbb{R}^k \to \mathbb{R}$ such that

$$f_\theta(x) = g_\theta(t(x))h(x), \quad \forall\ x \in \mathbb{R}^n, \ \ \forall \theta \in \Theta. \tag{1.5}$$

Proof. We will prove only the discrete case to avoid measure theory. For a general case, see Keener Section 6.4.

Suppose Y is sufficient. Let $x \in \mathbb{R}^n$. Note that by definition and using $Y = t(X)$,

$$f_\theta(x) = \mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(X = x \cap t(X) = t(x)) = \mathbb{P}_\theta(Y = t(x))\mathbb{P}_\theta(X = x \mid Y = t(x))$$

By sufficiency, $\mathbb{P}_\theta(X = x \mid Y = t(x))$ does not depend on θ. Therefore we can satisfy (1.5) with $g_\theta(t(x)) = \mathbb{P}_\theta(Y = t(x))$, $h(x) = \mathbb{P}_\theta(X = x \mid Y = t(x))$, so the factorization holds.

Now suppose there exists a nonnegative $\{g_\theta : \theta \in \Theta\}$, $h : \mathbb{R}^n \to \mathbb{R}$, $g_\theta : \mathbb{R}^k \to \mathbb{R}$ such that

$$f_\theta(x) = g_\theta(t(x))h(x), \quad \forall\ x \in \mathbb{R}^n, \ \ \forall \theta \in \Theta.$$

Define $r_\theta(z) := \mathbb{P}_\theta(t(X) = z) \ \ \forall\ z \in \mathbb{R}^k$ (the probability mass function for $t(X)$). Also define $t^{-1}t(x) := \{y \in \mathbb{R}^n; t(y) = t(x)\} \ \ \forall x \in \mathbb{R}^n$. To show sufficiency, we need to show that $\mathbb{P}_\theta(X = x \mid Y = t(x))$ does not depend on θ. Note that

$$\mathbb{P}_\theta(X = x \mid Y = t(x)) = \frac{f_\theta(x)}{f_Y(t(x))} = \frac{f_\theta(x)}{r_\theta(t(x))}$$

Using our assumption and the Total Probability Theorem, we have

$$= \frac{g_\theta(t(x)h(x))}{\mathbb{P}_\theta(t(X) = t(x))} = \frac{g_\theta(t(x))h(x)}{\sum_{z \in t^{-1}t(x)} \mathbb{P}_\theta(X = z)} = \frac{g_\theta(t(x))h(x)}{\sum_{z \in t^{-1}t(x)} f_\theta(z)} = \frac{g_\theta(t(x))h(x)}{\sum_{z \in t^{-1}t(x)} g_\theta(t(z))h(z)}$$

By definition of $t^{-1}t(z)$, we we can write this as

$$= \frac{g_\theta(t(x))h(x)}{\sum_{z \in t^{-1}t(x)} g_\theta(t(x))h(z)} = \frac{g_\theta(t(x))h(x)}{g_\theta(t(x)) \sum_{z \in t^{-1}t(x)} h(z)} = \frac{h(x)}{\sum_{z \in t^{-1}t(x)} h(z)}$$

where the second-to-last step follows since $t(x)$ is constant for all $z \in t^{-1}t(x)$. Since this expression does not contain θ, Y is sufficient for θ.

□

Remark 8. Intuition: data only cares about θ through $t(x)$.

To use the Factorization Theorem (Theorem 1.3.1.3) to find a sufficient statistic, we factor the joint pdf of the sample into two parts, with one part not depending on θ. The other part, the one that depends on θ, usually depends on the sample only through the function $t(x)$, and this function is a sufficient statistic for θ.

Exercise 1. Suppose $X_1, X_2, \ldots, X_n \sim$ i.i.d. $\mathcal{N}(0,1)$. So density is

$$\frac{1}{\sqrt{2\pi}} e^{-1/2(x-\theta)^2}$$

Show that

$$e^{-1/2(x^2-2x\theta+\theta^2)} =$$

$$f_\theta(x) = \left(\frac{1}{2\pi}\right)^{n/2} e^{2/12\sum_i X_i^2} e^{\theta\sum X_i - n\theta^2/2}$$

so if $t(x) = \sum_{i=1}^n X_i$, $h(x) = \left(\frac{1}{2\pi}\right)^{n/2} e^{2/12\sum_i X_i^2}$, $g_\theta(t(x)) = e^{\theta\sum X_i - n\theta^2/2}$, then by the Factorization Theorem (Theorem 1.3.1.3) this $(\overline{x})$ is a sufficient statistic.

Remark 9. In this case, if we deleted the original data we could recreate the original data by sampling from a $\mathcal{N}(0,1)$ distribution, then add the difference between the mean we get and the original sample mean to get an equivalent data set to the original one.

Remark 10. Suppose we define $t(x) := x$, $\forall\, x \in \mathbb{R}^n$. Then $Y = t(X_1, \ldots, X_n) = (X_1, \ldots, X_n)$ is (trivially) sufficient for θ. In general there will be infinitely many sufficient statistics for θ. For instance, in Example 1.3.1.1, $(X_1 + \ldots + X_n)^2$ is also sufficient. So is $(X_1 + \ldots + X_n)^3$, etc. More generally, any invertible function of any sufficient statistic is itself sufficient.

We can see that $(X_1 k \ldots, X_n)$ is sufficient for θ if $(t(x_1, \ldots, x_n) = (x_1, \ldots, x_n)$, $g_\theta = f_\theta, h = 1$. But this is not really helpful. We see we are interested in sufficient statistics that are smaller—reduce the data (in some sense) as much as possible.

1.3.2 Minimal Sufficient Statistics

Proposition 1.3.2.1. Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of probability density functions or probability mass functions. Let $t : \mathbb{R}^n \to \mathbb{R}^k$. Let $Y := t(X_1, \ldots, X_n)$. Assume Y is sufficient of θ. Let $a : \mathbb{R}^n \to \mathbb{R}^m$, let $Z := u(X_1, \ldots, X_n)$. suppose there exists $r : \mathbb{R}^m \to \mathbb{R}^k$ such that $r(u(x)) = t(x)$ for all $x \in \mathbb{R}^n$. That is, suppose $Y = r(Z)$. Then Z is sufficient for θ.

Proof.

$$f_\theta(x) = g_\theta(t(x))h(x) = g_\theta(r(u(x))h(x)$$

there exists $g_\theta : \mathbb{R}^k \to [0, \infty)$. Y is sufficient.

Define

$$\tilde{g}_\theta(y) := g_\theta(r(y)) \quad \forall\, y \in \mathbb{R}^m$$

So

$f_\theta(x) = \tilde{g}_\theta(u(x))h(x) \quad \forall x \in \mathbb{R}^n$. So Z is sufficient for θ by the Factorization Theorem (Theorem 1.3.1.3).

□

Definition 1.3.3 (**Minimal sufficient statistic**). Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of probability density functions or probability mass functions. Let $t : \mathbb{R}^n \to \mathbb{R}^k$. Let $Y := t(X_1, \ldots, X_n)$. Assume Y is sufficient of θ. Then Y is a **minimal sufficient statistic** for θ if for every statistic $Z : \Omega \to \mathbb{R}^m$ that is sufficient for θ there exists a function $\mathbb{R}^m \to \mathbb{R}^k$ such that $Y = r(Z)$.

Remark 11. Minimal sufficient statistics are not in general unique (because if you take any one-to-one function you get another one), but they are unique up to invertible transformations. (This is true because if Y and Z are both minimal sufficient, $Y = r(Z)$ and $Z = s(Y)$, so $Y = r(s(Y)), Z = s(r(Z)))$. They exist under mild assumptions (for a family of densities or probability mass functions).

Proposition 1.3.2.2 (**Proposition Larry Goldstein gave in class; Proposition 5.12 in notes**). Suppose $X_1, \ldots, X_n$ is a random sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$, is a family of probability density functions or probability mass functions ($\Theta \in \mathbb{R}^n$). (In the case of probability mass functions, we also assume that the set $\cup_{\theta \in \Theta}\{x \in \mathbb{R}^n : f_\theta(x) > 0\}$ is countable.) Then there exists a statistic Y that is minimal sufficient for θ.

Proof where θ is countable. By relabeling, let $\Theta = \{1, 2, \ldots\}$. We say for x, y sequences, we define the equivalence relation $x \sim y$ if $\exists\ \alpha \in \mathbb{R}$ such that $x = \alpha y$. Finite

$$t : \mathbb{R}^n \to \mathbb{R}^m / \sim, \qquad \Theta = \{1, \ldots, m\}$$

$$t(x) = (f_1(x), f_2(x), \ldots, f_m(x))$$

these likelihood are multiples of each other where α is a constant. The likelihood ratio is a constant not depending on θ. If they have the same $t(x)$ then we have that.

□

Theorem 1.3.2.3 (**Theorem 5.8 in 541A notes**). Let $\{f_\theta : \theta \in \Theta\}$ be a family of probability density functions or probability mass functions. Let $X_1, \ldots, X_n$ be a random sample from a member of the family.

Let $t : \mathbb{R}^n \to \mathbb{R}^m$ and define $Y := t(X_1, \ldots, X_n)$. Assume that Y is sufficient for θ. Y is minimal sufficient if and only if the following condition holds for every $x, y \in \mathbb{R}^n$:

There exists $c(x, y) \in \mathbb{R}$ that does not depend on θ such that $f_\theta(x) = c(x, y) f_\theta(y) \quad \forall\, \theta \in \Theta$

if and only if

$t(x) = t(y)$.

Proof. We are only considering probability mass functions to make things easier. We first prove sufficiency. We will show that the condition holding implies that Y is minimal sufficient.

Recall the likelihood ratio:

$$\frac{f_\theta(x)}{f_\theta(y)}$$

Note that the condition is equivalent to the likelihood ratio not depending on θ if and only if $t(x) = t(y)$. Consider the range $R = \{t(x) : x \in \mathbb{R}^n\}$ jnd then for $t \in R$ let $S_t = \{y : S(y) = t\}$. If t is in R, then there must be some z so that $t(z)$ is that z. This ensure that S_t is nonempty (there is at least one z so that $t(z) = t$. Let $t(x) \in R$, then $S_{t(x)}$ is nonempty (in particular it contains x). Pick any y you like in $t(x)$: $y \in S$. S depends on $t(x)$ so we can index it by $t(x)$: $y_{t(x)} \in S_{t(x)}$. let $y_t \in S_t$. Note that

$$t(y_{t(x)}) = t(x)$$

But now by the assumption, we have

There exists $c(x, y_{t(x)}) \in \mathbb{R}$ that does not depend on θ such that $f_\theta(x) = c(x, y_{t(x)}) f_\theta(y_{t(x)}) \quad \forall\, \theta \in \Theta$

Then note that if $h(x) = c(x, y_{t(x)})$, $g_\theta(t) = f_\theta(y_t) \iff g_\theta(t(x)) = f_\theta(y_{t(x)})$, we meet the conditions for the Factorization Theorem (Theorem 1.3.1.3). So using the Factorization Theorem, Y is sufficient.

$\vdots$

Part we did in class on Friday 02/15: evidently (according to Goldstein) this shows that the statistic is minimal but not necessarily sufficient. Let $Z = u(X_1, \ldots, X_n)$ be any other sufficient statistic. We need to eventually show that Y is a function of Z. By the Factorization Theorem (Theorem 1.3.1.3), there exists $h : \mathbb{R}^m \to \mathbb{R}, g_\theta; \mathbb{R}^n \to \mathbb{R}$ such that for all $\theta \in \Theta$,

$$f_\theta(x) = g'_\theta(u(x))h'(x), \quad \forall\, x \in \mathbb{R}^n.$$

Let $y \in \mathbb{R}^n$. If $h'(y) = 0$, then $f_\theta(y) = 0$ for all $\theta \in \Theta$. So, $\mathbb{P}_\theta(y \in \mathbb{R}^n : h'(y) = 0) = 0$ for all $\theta \in \Theta$. So we can ignore this possibility since it's a probability 0 event and assume $h'(y) > 0, \forall y \in \mathbb{R}^n$.

Now let $x, y \in \mathbb{R}^n$ such that $u(x) = u(y)$. **By an exercise we're going to do later**, if $t(x) = t(y)$ then t is a function of u, so we will be done if we can show that $t(x) = t(y)$. Note that since $u(x) = u(y)$, for any $\theta \in \Theta$

$$f_\theta(x) = g'_\theta(u(x))h'(x) = \frac{g'_\theta(u(y))h'(x) = g'_\theta(u(y))h'(y)}{f_\theta(y)} \frac{h'(x)}{h'(y)} = f_\theta(y)\frac{h'(x)}{h'(y)}, \quad forall\theta \in \Theta$$

So define $c(x, y) = h'(x)/h'(y)$, we have

$$f_\theta(x) = f_\theta(y)c(x, y), \quad \forall\theta \in \Theta$$

Therefore $t(x) = t(y)$, so we're done showing that if the condition holds then Y is minimal sufficient.

Then next thing to show is that if Y is minimal sufficient then the condition holds.

$$\vdots$$

For any $z \in \{t(x) : x \in \mathbb{R}^n\}$, let x_z be any element of $t^{-1}(z)$

□

Proposition 1.3.2.4 (**Exercise 5.10 in Math 541A notes**). Let $\{f_\theta : \theta \in \Theta\}$ be a k-parameter exponential family $\{f_\theta : \theta \in \Theta,\, a(w(\theta)) < \infty\}$ of probability density functions or probability mass functions, where

$$f_\theta(x) := h(x) \exp\Big(\sum_{i=1}^{k} w_i(\theta)t_i(x) - a(w(\theta))\Big), \qquad \forall\, x \in \mathbb{R}.$$

For any $\theta \in \Theta$, let $w(\theta); = (w_1(\theta), \ldots, w_k(\theta))$. Assume that the following subset of $\mathbb{R}^k$ is k-dimensional:

$$\{(w_1(\theta), \ldots, w_k(\theta)) \in \mathbb{R}^k : \theta \in \Theta\}.$$

That is, if $x \in \mathbb{R}^k$ satisfies $\langle x, y\rangle = 0$ for all y in this set, then $x = 0$.

Let $X = (X_1, \ldots, X_n)$ be a random sample of size n from f_θ. Define $t : \mathbb{R}^n \to \mathbb{R}^n$ by

$$t(X) := \sum_{j=1}^{n} \left(t_1(X_j), \ldots, t_n(X_j)\right).$$

Then $t(X)$ is minimal sufficient for θ.

Proof. First note that $t(X)$ is sufficient by the Factorization Theorem (Theorem 1.3.1.3) because we have for any $x = (x_1, \dots, x_n) \in \mathbb{R}^n$

$$f_\theta(x) = \prod_{j=1}^n \left[h(x_j) \exp\left(\sum_{i=1}^k w_i(\theta) t_i(x_j) - a(w(\theta)) \right) \right] = \exp\left(\sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(x_j) - n \cdot a(w(\theta)) \right) \prod_{j=1}^n h(x_j)$$

$$= g_\theta(t(x)) h(x), \qquad \forall\, x \in \mathbb{R}^n$$

where

$$h(x) = \prod_{j=1}^n h(x_j), \qquad g_\theta(t(x)) = \exp\left(\sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(x_j) \right).$$

Now we will show minimal sufficiency using Theorem 5.8 from the lecture notes (Theorem 1.3.2.3). We seek to show that for every $x, y \in \mathbb{R}^n$, the likelihood ratio $\frac{f_\theta(x)}{f_\theta(y)}$ is constant (the constant may depend on x and y) if and only if $t(x) = t(y)$. Let $x, y \in \mathbb{R}^n$. Suppose there is some constant $c(x, y) > 0$ that may depend on x and y but not θ such that

$$\exp\left(\sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(x_j) - n \cdot a(w(\theta)) \right) \prod_{j=1}^n h(x_j)$$

$$= c(x, y) \exp\left(\sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(y_j) - n \cdot a(w(\theta)) \right) \prod_{j=1}^n h(y_j)$$

$$\iff \exp\left(\sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(x_j) \right) = c_1(x, y) \exp\left(\sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(y_j) \right)$$

(where cancellation of the h functions is permissible since they depend only on x and y, so we can let $c_1(x, y) = c(x, y) \cdot \prod_{j=1}^n h(y_j) / \prod_{j=1}^n h(x_j) > 0$)

$$\iff \sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(x_j) = c_2(x, y) + \sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(y_j)$$

where $c_2(x, y) = \log\big(c_1(x, y)\big)$. Then if θ_0, θ_1 are any two points in Θ,

$$\sum_{j=1}^n \sum_{i=1}^k w_i(\theta_0) t_i(x_j) - \sum_{j=1}^n \sum_{i=1}^k w_i(\theta_1) t_i(x_j) = \sum_{j=1}^n \sum_{i=1}^k w_i(\theta_0) t_i(y_j) - \sum_{j=1}^n \sum_{i=1}^k w_i(\theta_1) t_i(y_j)$$

(where the $c_2(x, y)$ terms cancel since (x, y) is held fixed.)

$$\iff \sum_{j=1}^{n}\sum_{i=1}^{k}[w_i(\theta_0) - w_i(\theta_1)]t_i(x_j) = \sum_{j=1}^{n}\sum_{i=1}^{k}[w_i(\theta_0) - w_i(\theta_1)]t_i(y_j)$$

$$\iff \sum_{j=1}^{n}\sum_{i=1}^{k}[w_i(\theta_0) - w_i(\theta_1)][t_i(x_j) - t_i(y_j)] = 0.$$

This equation holds for all $\theta \in \Theta$ if and only if $t_i(x_j) - t_i(y_j) = 0, i = 1, \dots, k, j = 1, \dots, n$; that is, it holds if and only if $t(x) = t(y)$. Therefore we have shown that $t(\cdot)$ is minimal sufficient by Theorem 1.3.2.3.

□

Remark 12. Note that the assumption of the exercise is always satisfied for an exponential family in canonical form. From this proposition we can conclude that if we sample from a Gaussian with unknown mean μ and variance $\sigma^2 > 0$, then $\overline{X}$ is minimal sufficient for θ and $(\overline{X}, S)$ is minimal sufficient for (μ, σ^2).

Proposition 1.3.2.5 (**Exercise 5.13 in Math 541A notes**)**.** Let $\mathbb{P}_1, \mathbb{P}_2$ be two probability laws on the sample space $\Omega = \mathbb{R}$. Suppose these laws have densities $f_1, f_2 : \mathbb{R} \to [0, \infty)$ so that

$$\mathbb{P}_i(A) = \int_A f_i(x)dx, \qquad \forall\, i = 1, 2, \qquad \forall\, A \subseteq \mathbb{R}.$$

Then

(a)

$$\sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \frac{1}{2}\int_{\mathbb{R}} |f_1(x) - f_2(x)|dx.$$

(b) If $\mathbb{P}_1, \mathbb{P}_2$ are probability laws on $\Omega = \mathbb{Z}$

$$\sup_{A \subseteq \mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \frac{1}{2}\sum_{z \in \mathbb{Z}} |\mathbb{P}_1(z) - \mathbb{P}_2(z)|.$$

Proof. (a) Note that $\sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$ returns the difference in areas under f_1 and f_2 in the region $A \subset \mathbb{R}$ where that difference is positive. Suppose without loss of generality that

$$\sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \sup_{A \subseteq \mathbb{R}} \{\mathbb{P}_1(A) - \mathbb{P}_2(A)\}. \tag{1.6}$$

(There is no loss of generality because in the case that $\sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \sup_{A \subseteq \mathbb{R}} \{\mathbb{P}_2(A) - \mathbb{P}_1(A)\}$, we can simply switch the names of $\mathbb{P}_1$ and $\mathbb{P}_2$ to get the desired result.) Then the region A which maximizes the quantity on the right side of (1.6) is the suggested region, $A := \{x \in \mathbb{R} : f_1(x) > f_2(x)\}$. That is,

$$\sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \int_A (f_1(x) - f_2(x))dx.$$

Note that

$$\int_{\mathbb{R}} |f_1(x) - f_2(x)|dx = \int_A |f_1(x) - f_2(x)|dx + \int_{\mathbb{R} \setminus A} |f_1(x) - f_2(x)|dx$$

$$\iff \int_{\mathbb{R}} |f_1(x) - f_2(x)|dx = \int_A (f_1(x) - f_2(x))dx + \int_{\mathbb{R}\setminus A} (f_2(x) - f_1(x))dx. \tag{1.7}$$

Since we already have $\int_A (f_1(x)-f_2(x))dx = \sup_{A\subseteq\mathbb{R}} |\mathbb{P}_1(A)-\mathbb{P}_2(A)|$, if it is also true that $\int_{\mathbb{R}\setminus A}(f_2(x)-f_1(x))dx = \sup_{A\subseteq\mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$ we are done by (1.7), so this is what we will show next. Let $\int_A f_2(x)dx = a_2$ and let $\int_A f_1(x)dx = a_1$, so that

$$\sup_{A\subseteq\mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \int_A (f_1(x) - f_2(x))dx = a_1 - a_2.$$

Note that

$$1 = \int_{\mathbb{R}} f_2(x)dx = \int_A f_2(x)dx + \int_{\mathbb{R}\setminus A} f_2(x)dx = a_2 + \int_{\mathbb{R}\setminus A} f_2(x)dx \iff \int_{\mathbb{R}\setminus A} f_2(x)dx = 1 - a_2,$$

and similarly $\int_{\mathbb{R}\setminus A} f_1(x)dx = 1 - a_1$. Therefore

$$\int_{\mathbb{R}\setminus A}(f_2(x)-f_1(x))dx = \int_{\mathbb{R}\setminus A} f_2(x)dx - \int_{\mathbb{R}\setminus A} f_1(x)dx = 1-a_2-(1-a_1) = a_1-a_2 = \sup_{A\subseteq\mathbb{R}} |\mathbb{P}_1(A)-\mathbb{P}_2(A)|.$$

So by (1.7), we have

$$\int_{\mathbb{R}} |f_1(x) - f_2(x)|dx = 2 \sup_{A\subseteq\mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| \iff \sup_{A\subset\mathbb{R}} |¶_1(A) - ¶_2(A)| = \frac{1}{2}\int_{\mathbb{R}} |f_1(x) - f_2(x)|dx.$$

(b) Analogous to the proof of (a). Note that $\sup_{A\subseteq\mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$ returns the difference in probabilities for those numbers in the set $A \subset \mathbb{Z}$ where that difference is positive. Suppose without loss of generality that

$$\sup_{A\subseteq\mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \sup_{A\subseteq\mathbb{Z}} \{\mathbb{P}_1(A) - \mathbb{P}_2(A)\}. \tag{1.8}$$

(There is no loss of generality because in the case that $\sup_{A\subseteq\mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \sup_{A\subseteq\mathbb{Z}}\{\mathbb{P}_2(A) - \mathbb{P}_1(A)\}$, we can simply switch the names of $\mathbb{P}_1$ and $\mathbb{P}_2$ to get the desired result.) Then the set A which maximizes the quantity on the right side of (1.8) can be defined as (similarly to part (a)) $A := \{z \in \mathbb{Z} : \mathbb{P}_1(z) > \mathbb{P}_2(z)\}$. That is,

$$\sup_{A\subseteq\mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \sum_{z\in A} \mathbb{P}_1(z) - \mathbb{P}_2(z).$$

Note that

$$\sum_{z\in\mathbb{Z}} |\mathbb{P}_1(z) - \mathbb{P}_2(z)| = \sum_{z\in A} |\mathbb{P}_1(z) - \mathbb{P}_2(z)| + \sum_{z\in\{\mathbb{Z}\setminus A\}} |\mathbb{P}_1(z) - \mathbb{P}_2(z)|$$

$$\iff \sum_{z\in\mathbb{Z}} |\mathbb{P}_1(z) - \mathbb{P}_2(z)| = \sum_{z\in A} \mathbb{P}_1(z) - \mathbb{P}_2(z) + \sum_{z\in\{\mathbb{Z}\setminus A\}} (\mathbb{P}_2(z) - \mathbb{P}_1(z)). \tag{1.9}$$

Since we already have $\sum_{z\in A}\mathbb{P}_1(z)-\mathbb{P}_2(z) = \sup_{A\subseteq\mathbb{Z}}|\mathbb{P}_1(A)-\mathbb{P}_2(A)|$, if it is also true that $\sum_{z\in\{\mathbb{Z}\setminus A\}}(\mathbb{P}_2(z)-\mathbb{P}_1(z)) = \sup_{A\subseteq\mathbb{Z}}|\mathbb{P}_1(A)-\mathbb{P}_2(A)|$ we are done by (1.9), so this is what we will show next. Let $\sum_{z\in A}\mathbb{P}_2(z) = a_2$ and let $\sum_{z\in A}\mathbb{P}_1(z) = a_1$, so that

$$\sup_{A\subseteq\mathbb{Z}}|\mathbb{P}_1(A)-\mathbb{P}_2(A)| = \sum_{z\in A}\mathbb{P}_1(z)-\mathbb{P}_2(z) = a_1-a_2.$$

Note that

$$1 = \sum_{z\in Z}\mathbb{P}_2(z) = \sum_{z\in A}\mathbb{P}_2(z) + \sum_{z\in\{\mathbb{Z}\setminus A\}}\mathbb{P}_2(z) = a_2 + \sum_{z\in\{\mathbb{Z}\setminus A\}}\mathbb{P}_2(z) \iff \sum_{z\in\{\mathbb{Z}\setminus A\}}\mathbb{P}_2(z) = 1-a_2,$$

and similarly $\sum_{z\in\{\mathbb{Z}\setminus A\}}\mathbb{P}_1(z) = 1-a_1$. Therefore

$$\sum_{z\in\{\mathbb{Z}\setminus A\}}(\mathbb{P}_2(z)-\mathbb{P}_1(z)) = \sum_{z\in\{\mathbb{Z}\setminus A\}}\mathbb{P}_2(z) - \sum_{z\in\{\mathbb{Z}\setminus A\}}\mathbb{P}_1(z) = 1-a_2-(1-a_1) = a_1-a_2 = \sup_{A\subseteq\mathbb{Z}}|\mathbb{P}_1(A)-\mathbb{P}_2(A)|.$$

So by (1.9), we have

$$\sum_{z\in\mathbb{Z}}|\mathbb{P}_1(z)-\mathbb{P}_2(z)| = 2\sup_{A\subseteq\mathbb{Z}}|\mathbb{P}_1(A)-\mathbb{P}_2(A)| \iff \sup_{A\subset\mathbb{Z}}|\P_1(A)-\P_2(A)| = \frac{1}{2}\sum_{z\in\mathbb{Z}}|\mathbb{P}_1(z)-\mathbb{P}_2(z)|.$$

□

1.3.3 Ancillary Statistics

Definition 1.3.4 (**Ancillary Statistic**)**.** Suppose $X_1,\dots,X_n$ is a random sample of size n from a distribution f where $f\in\{f_\theta:\theta\in\Theta\}$ is a family of distributions. A statistic $Y = t(X_1,\dots,X_n), t:\mathbb{R}^n\to\mathbb{R}^m$ is **ancillary** for θ if the distribution of Y does not depend on θ.

Example 1.3.3. [**Example from 541B lecture**]

$X_1,\dots,X_n\sim\mathcal{N}(\mu,1)$. Then $V(x) := X_1-\overline{X}$ is ancillary. Proof: let $X_d = \mu + X'_d, X'_d\sim\mathcal{N}(0,1)$. Then

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n}X_j = \mu+\overline{X'}.$$

so

$$V(X_1,\dots,X_n) = \mu+X'_1-\mu-\overline{X'} = X'_1-\overline{X'}$$

Since X' does not depend on μ, $V(X_1,\dots,X_n)$ is ancillary for μ.

Example 1.3.4 (**Example 5.15 from 541A notes**)**.** Let $X_1,\dots,X_n$ be random sample of size n from the location family for the Cauchy distribution:

$$f_\theta(x) := \prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1+(x_i-\theta)^2}, \qquad \forall\ x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \quad \forall \theta \in \mathbb{R}.$$

Then the order statistics $X_{(1)} \leq \ldots \leq X_{(n)}$ are minimal sufficient for θ.

Proof. Sufficiency follows by the Factorization Theorem (Theorem 1.3.1.3, Theorem 5.4 in Math 541A notes) (**usually the easiest way to prove sufficiency**) since if $t(x) := (x_{(1)}, \ldots, x_{(n)})$, then $f_\theta(t(x)) = f_\theta(x)$ (because $t(x)$ is just a permutation so the product won't change).

To get minimal sufficiency, apply Theorem 1.3.2.3 (Theorem 5.8 in 541A notes). Recall we have minimal sufficiency if the following condition holds for every $x, y \in \mathbb{R}^n$:

> There exists $c(x,y) \in \mathbb{R}$ that does not depend on θ such that $f_\theta(x) = c(x,y) f_\theta(y) \quad \forall\ \theta \in \Theta$
>
> if and only if
>
> $$t(x) = t(y).$$

Let's try to show it.

$$\frac{f_\theta(x)}{f_\theta(y)} = \prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1+(x_i-\theta)^2} \Bigg/ \prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1+(y_i-\theta)^2} = \prod_{i=1}^{n} [1+(y_i-\theta)^2] \Bigg/ \prod_{i=1}^{n} [1+(x_i-\theta)^2] \tag{1.10}$$

Keep x, y fixed, $\theta \in \mathbb{R}$ variable. Then the likelihood ratio (1.10) does not depend on θ if and only if the roots in θ on top are equal to the roots on bottom. Roots on top: $\theta = y_i \pm \sqrt{-i}, 1 \leq i \leq n$. Roots on bottom: $\theta = x_i \pm \sqrt{-i}, 1 \leq i \leq n$. So we can see this is true if and only if the vector $(x_1, \ldots, x_n)$ is a permutation of $(y_1, \ldots, y_n)$, which is exactly the case if $t(x) = t(y)$.

□

However, this statistic has ancillary information. Specifically, $X_{(n)} - X_{(1)}$ is ancillary (its distribution does not depend on θ).

Proof. Let $Z_1, \ldots, Z_n$ be independent centered Cauchy random variables; that is, they all have density $\frac{1}{\pi} \frac{1}{1+a^2}, a \in \mathbb{R}$. Then $X_i = Z_i + \theta,\ \forall 1 \leq i \leq n, \forall \theta \in \mathbb{R}$. Also, $X_{(i)} = Z_{(i)} + \theta$. So $X_{(n)} - X_{(1)} = Z_{(n)} - Z_{(1)}$ does not depend on $\theta \in \mathbb{R}$. So, $X_{(n)} - X_{(1)}$ is ancillary for θ. That is, there exists a constant c that does not depend on θ such that $\mathbb{E}_\theta[X_{(n)} - X_{(1)} - c] = 0$ for all $\theta \in \mathbb{R}$.

Let $f : \mathbb{R}^n \to \mathbb{R}$, $f(x_1, \ldots, x_n) = x_n - x_1 - c\ \forall\ (x_1, \ldots, x_n) \in \mathbb{R}^n$. Then $\mathbb{E}_\theta f(Y) = 0,\ \forall \theta \in \Theta, Y = (X_{(1)}, \ldots, X_{(n)})$. Note that $f \neq 0$ (in fact $f(Y) \neq 0$ with probability 1).

□

1.3.4 Complete Statistics

Definition 1.3.5 (**Complete statistic; definition 5.16 in 541A notes**). Suppose $X_1, \dots, X_n$ is a random sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of distributions. A statistic $Y = t(X_1, \dots, X_n), t : \mathbb{R}^n \to \mathbb{R}^k$ is **complete** for θ if the following holds:

$$\text{For any } f : \mathbb{R}^k \to \mathbb{R} \text{ such that } \mathbb{E}_\theta f(Y) = 0 \ \forall \ \theta \in \Theta, \text{ it holds that } f(Y) = 0.$$

Intuition: Y has no "excess information."

Definition 1.3.6 (**541B definition**). T is **complete** for $\{f_\theta, \theta \in \Theta\}$ if

$$\mathbb{E}_\theta \phi(T) = 0 \ \forall \theta \implies \phi(T) = 0 \text{ almost surely.}$$

Remark 13. If a statistic has ancillary information, then it is not complete. Therefore if a statistic is complete, it is not ancillary. Minimal sufficient statistics pretty much always exist, but a complete sufficient statistic might not exist (**see example from homework 4**).

Remark 14. A complete statistic may not be sufficient (example: a constant).

Example 1.3.5 (**Exercise 5.19 in Math 541A notes**). The following is an example of a statistic Y that is complete and nonconstant, but not sufficient. Suppose $X_1, \dots, X_n$ is a random sample of known size n from a Bernoulli distribution with unknown probability parameter $\theta \in (0, 1)$. Let

$$Y = t(X_1, \dots, X_n) = \sum_{i=1}^{n-1} X_i.$$

Then Y is complete for μ because for any $f : \mathbb{R}^m \to \mathbb{R}$, suppose

$$0 = \mathbb{E}_\theta f(Y) = \sum_{j=0}^{n-1} f(j) \Pr(Y = j \mid \theta) = \sum_{j=0}^{n-1} f(j) \binom{n-1}{j} \theta^j (1-\theta)^{n-1-j}, \qquad \forall \theta \in (0, 1).$$

Divide by $(1-\theta)^{n-1}$ and let $\alpha = \theta/(1-\theta)$ for notational ease. (Note that since $\theta \in (0, 1)$, $\alpha > 0$.)

$$0 = \sum_{j=0}^{n-1} f(j) \binom{n-1}{j} \theta^j \frac{(1-\theta)^{n-1-j}}{(1-\theta)^{n-1}} = \sum_{j=0}^{n-1} f(j) \binom{n-1}{j} \left(\frac{\theta}{1-\theta}\right)^j = \sum_{j=0}^{n-1} f(j) \binom{n-1}{j} \alpha^j, \qquad \forall \alpha > 0.$$

The sum on the right side is a polynomial in $\alpha > 0$. That means the sum on the right can only equal 0 if every coefficient on the polynomial equals zero. $\binom{n}{j}$ is of course nonzero for all $j \in 0, \dots, n-1$. Therefore for all $\alpha > 0$ we have that $\mathbb{E}_\theta f(Y) = 0$ only if $f(Y) = 0$, so Y is complete.

However, Y is not sufficient for μ. Using the definition of conditional probability,

$$\Pr\big((X_1,\dots,X_n)=(x_1,\dots,x_n)\mid Y=y\big)=\frac{1}{\Pr(Y=y)}\cdot\Pr\big((X_1,\dots,X_n)=(x_1,\dots,x_n)\cap Y=y\big)$$

$$=\frac{1}{\Pr(Y=y)}\cdot\Pr\big((X_1,\dots,X_n)=(x_1,\dots,x_n)\big)$$

Using independence and the definition of a binomial distribution, we have

$$=\frac{1}{\binom{n-1}{y}\theta^y(1-\theta)^{n-1-y}}\cdot\prod_{i=1}^{n}\Pr\big(X_i=x_i\big)=\frac{1}{\binom{n-1}{y}\theta^y(1-\theta)^{n-1-y}}\cdot\prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i}$$

$$=\frac{1}{\binom{n-1}{y}\theta^y(1-\theta)^{n-1-y}}\cdot\theta^y(1-\theta)^{n-y}=\frac{1-\theta}{\binom{n-1}{y}}.$$

Because $\Pr\big((X_1,\dots,X_n)=(x_1,\dots,x_n)\mid Y=y\big)=(1-\theta)/\binom{n-1}{y}$ depends on θ, Y is not sufficient for θ.

Exercise 2 (**Exercise 5.20 in Math 541A notes**)**.** This exercise shows that a complete sufficient statistic might not exist.

Let $X_1,\dots,X_n$ be a random sample of size n from the uniform distribution on the three points $\{\theta,\theta+1,\theta+2\}$, where $\theta\in\mathbb{Z}$.

(a) Show that the vector $Y:=(X_{(1)},X_{(n)})$ is minimal sufficient for θ.

(b) Show that Y is not complete by considering $X_{(n)}-X_{(1)}$.

(c) Using minimal sufficiency, conclude that any sufficient statistic for θ is not complete.

Proof. (a) First we need to show that Y is sufficient for θ. Informally, it makes sense that this would be the case because there are three possibilities:

(1) If θ and $\theta+2$ appear in the data set, we can identify θ with certainty as the smallest of them. Simply observing $x_{(1)}$ and $x_{(n)}$ would show us the smallest and largest observations, which would be 2 units apart. Then we would known that the these observations are θ and $\theta+2$, and we could identify θ with certainty. (Note that it doesn't matter in this case whether we observe $\theta+1$.)

(2) If only the values $\{\theta,\theta+1\}$ or $\{\theta+1,\theta+2\}$ appear in the data set, we have to guess which one of these pairs we observed. If we guess that we have observed $\theta+1$ and $\theta+2$ and subtract one from the smallest observation to estimate θ, or we can guess that we have observed θ and $\theta+1$ and use the smallest value as our estimate for θ. (Or we can hedge and take the mean of these values.) In any case, simply observing $x_{(1)}$ and $x_{(n)}$ would show us the smallest and largest observations, which would be 1 apart, leaving us in the same position as if we had all the data.

(3) If only one value appears, we have to guess if it is θ, $\theta+1$, or $\theta+2$ in a way similar to if we only observe two values. But if we observe that the smallest and largest values are equal, we are observing the same information and we are in the same position for estimating θ as if we had all of the data.

We can formally show that Y is sufficient using the Factorization Theorem (Theorem 1.3.1.3). Note that because on each trial we observe $\theta, \theta+1$, or $\theta+2$ with equal probability, the mass function for the unordered observations $f_{u,\theta} : \mathbb{Z}^n \to \mathbb{R}$ is the same as that of a multinomial distribution with three outcomes with equal probabilities. That is, if $n_0 = \sum_{i=1}^n \mathbf{1}_{\{x_i=\theta\}}$ (where $\mathbf{1}_{\{x_i=\theta\}}$ is an indicator variable for the ith observation having value θ), $n_1 = \sum_{i=1}^n \mathbf{1}_{\{x_i=\theta=1\}}$, and $n_2 = \sum_{i=1}^n \mathbf{1}_{\{x_i=\theta+2\}}$, we have

$$f_{u,\theta}(x) = \binom{n}{n_0, n_1, n_2}\left(\frac{1}{3}\right)^n = \frac{n!}{n_0!n_1!n_2!}\left(\frac{1}{3}\right)^n.$$

But taking into account the order in which we observe the samples, the probability of observing any one sample of size n ($x = (x_1, \ldots, x_n)$) is simply

$$f_\theta(x) = \begin{cases} \left(\frac{1}{3}\right)^n & x_1 \in \{\theta, \theta+1, \theta+2\}, \ldots, x_n \in \{\theta, \theta+1, \theta+2\} \\ 0 & \text{otherwise.} \end{cases}$$

We have $t : \mathbb{Z}^n \to \mathbb{Z}^2$ is given by

$$t(X_1, \ldots, X_n) = (X_{(1)}, X_{(n)}).$$

Choose

$$h(x) = (1/3)^n, \qquad g_\theta(t(x)) = \begin{cases} 1 & t(x) \in \{\theta, \theta+1, \theta+2\} \times \{\theta, \theta+1, \theta+2\} \\ 0 & \text{otherwise.} \end{cases} \tag{1.11}$$

Then we have $f_\theta(x) = g_\theta(t(x))h(x)$, as desired. Now we will use Theorem 1.3.2.3 (Theorem 5.8 from the lecture notes) to show that Y is not only sufficient, but is also minimal sufficient. Let $x, z \in \mathbb{Z}^n$, and let $y_x = (x_{(1)}, x_{(n)})$, $y_z = (z_{(1)}, z_{(n)})$. We seek to show that for every $x, z \in \mathbb{Z}^n$, the likelihood ratio $\frac{f_\theta(x)}{f_\theta(z)}$ is constant (the constant may depend on x and z) if and only if $y_x = y_z$. (We only need to consider $x, z \in \mathbb{Z}^n$ rather than all of $\mathbb{R}^n$ because since $\theta \in \mathbb{Z}$, $X_i \in \mathbb{Z}$ $\forall i \in \{1, \ldots, n\}$, $\forall \theta \in \Theta$.) Using the expressions in (1.11), we can write the equation in (1.12) as

$$f_\theta(x) = c(x,y)f_\theta(z) \iff g_\theta(t(x))\left(\frac{1}{3}\right)^n = c(x,z)g_\theta(t(z))\left(\frac{1}{3}\right)^n \iff g_\theta(t(x)) = c(x,z)g_\theta(t(z)) \tag{1.12}$$

I will argue that the equality in (1.12) only holds for some $c(x, z) \in \mathbb{R}$ if $t(x) = t(z)$. Suppose we have observed data x and z from a distribution with a specific θ_0. There are three cases to consider:

(1) $t(x) = \{\theta_0, \theta_0 + 2\}$ **(full information):** Then the only θ for which $g_\theta(t(x)) \neq 0$ is $\theta = \theta_0$. (This corresponds to situation (1) above.)

(2) $t(x) = \{\theta_0, \theta_0 + 1\}$ **or** $\{\theta_0, \theta_0 + 1\}$**:** Then $g_\theta(t(x)) \neq 0$ for two values of θ. In the first case, those two values will be $\theta_0 - 1$ and θ_0. In the second case, those two values will be θ_0 and $\theta_0 + 1$. (This corresponds to situation (2) above.)

(3) $t(x) = \{\theta_0, \theta_0\}$**,** $\{\theta_0 + 1, \theta_0 + 1\}$**, or** $\{\theta_0 + 2, \theta_0 + 2\}$**:** Then $g_\theta(t(x)) \neq 0$ for three values of θ. In the first case, those three values will be $\theta_0 - 2$, $\theta_0 - 1$, and θ_0. In the second case, those three values will be $\theta_0 - 1$, θ_0, and $\theta_0 + 1$. In the last case, those three values will be θ_0, $\theta_0 + 1$, and $\theta_0 + 2$. (This corresponds to situation (3) above.)

I have enumerated all possible values of $t(x)$ or $t(z)$ for a given true $\theta = \theta_0$, and note that there is no overlap among any of the possibilities for what values of θ will yield identical values of $g_\theta(t(x))$ and $g_\theta(t(z))$ for all θ. That is, that is true (and (1.12) holds for all $\theta \in \Theta = \mathbb{Z}$) if and only if $t(x) = t(z)$, which is what we were trying to show. So Y is minimal sufficient.

(b) Recall the definition of a complete statistic:

Definition 1.3.7 (Complete statistic; definition 5.16 in 541A notes). Suppose $X_1, \ldots, X_n$ is a random sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of distributions. A statistic $Y = t(X_1, \ldots, X_n), t : \mathbb{R}^n \to \mathbb{R}^k$ is **complete** for θ if the following holds:

$$\text{For any } f : \mathbb{R}^k \to \mathbb{R} \text{ such that } \mathbb{E}_\theta f(Y) = 0 \ \forall \ \theta \in \Theta, \text{ it holds that } f(Y) = 0.$$

We will show that Y is not complete by showing that Y contains ancillary information. Specifically, we will show that $\mathbb{E}_\theta[f(Y)] \neq 0$ where $f(Y) = X_{(n)} - X_{(1)} - c$ for some $c \in \mathbb{Z}$.

Let $Z_1, \ldots, Z_n$ be a random sample of size n from the uniform distribution on the three points $\{0, 1, 2\}$. Then $X_i = Z_i + \theta, \ \forall \ 1 \leq i \leq n, \forall \ \theta \in \mathbb{Z}$. Also, $X_{(i)} = Z_{(i)} + \theta$. So $X_{(n)} - X_{(1)} = Z_{(n)} - Z_{(1)}$ does not depend on $\theta \in \mathbb{Z}$. So, $X_{(n)} - X_{(1)}$ is ancillary for θ. That is, there exists a constant $c \in \mathbb{Z}$ that does not depend on θ such that $\mathbb{E}_\theta[X_{(n)} - X_{(1)} - c] = 0$ for all $\theta \in \mathbb{Z}$.

Define this c and let $f : \mathbb{Z}^2 \to \mathbb{Z}$ be $f(x_1, x_n) := x_n - x_1 - c \ \forall \ (x_1, x_n) \in \mathbb{R}^n$. Then $\mathbb{E}_\theta f(Y) = 0, \ \forall \theta \in \Theta, Y = (X_{(1)}, X_{(n)})$.

(c) Let S be any sufficient statistic for θ. Since Y is minimal sufficient, there exists a function ϕ such that $Y = \phi(\theta)$. Therefore S is not complete because $\mathbb{E}_\theta(f(\phi(S)) = \mathbb{E}_\theta(f(Y)) = 0$ for all $\theta \in \mathbb{Z}$. So any sufficient statistic for θ is not complete.

□

Example 1.3.6 (Discrete RV example, Example 5.21 in Math 541A notes; return to Example 1.3.1.1). Suppose we take a sample of size n from a Bernoulli distribution with parameter $0 < \theta < 1$. We already showed $Y := X_1 + \ldots + X_n$ is sufficient for θ. Now we show Y is complete.

Proof. Let $f : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}_\theta f(Y) = 0 \ \ \forall \theta \in \Theta$. Since Y is binomial,

$$0 = \mathbb{E}_\theta f(Y) = \sum_{j=0}^{n} f(j) \binom{n}{j} \theta^j (1-\theta)^{n-j}, \qquad \forall \theta \in (0, 1)$$

Let $\alpha = \theta/(1-\theta)$ and divide by $(1-\theta)^n$;

$$0 = \sum_{j=0}^{n} f(j) \binom{n}{j} \alpha^j, \qquad \forall \alpha > 0.$$

The sum on the right side is a polynomial in $\alpha > 0$. That means the sum on the right can only equal 0 if every coefficient on the polynomial equals zero. $\binom{n}{j}$ is of course nonzero for all $j \in 0, \ldots, n-1$. Therefore

for all $\alpha > 0$ we have that $\mathbb{E}_\theta f(Y) = 0$ only if we have $f(j) = 0$ for all $0 \leq j \leq n$, so $f(Y) = 0$ so Y is complete.

□

Example 1.3.7 (Continuous RV example; return to Example 1.3.2). For a random sample from a Gaussian distribution with known variance $\sigma^2 > 0$ and unknown $\mu \in \mathbb{R}$, we showed that $Y = (X_1 + \ldots + X_n)/n$ is sufficient for μ. Now we will show it is complete.

Proof. **Found this proof confusing** For simplicity, let $\sigma = 1, n = 1$. Let $f : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}_\mu |f(Y)| < \infty$ for all $\mu \in \mathbb{R}$. Then

$$0 = \mathbb{E}_\mu(f(Y)) = \int_{-\infty}^{\infty} f(y) \exp\left(-\frac{(y-\mu)^2}{2} \right) \frac{1}{\sqrt{2\pi}} dy, \qquad \forall \mu \in \mathbb{R}$$

Multiplying both sides by $e^{\mu^2/2}\sqrt{2\pi}$ yields

$$0 = \int_{-\infty}^{\infty} f(y) e^{-y^2/2} e^{y\mu} dy, \qquad \forall \mu \in \mathbb{R} \tag{1.13}$$

If $f(y) \geq 0$, we are done since (1.13) is the moment generating function of a random variable with density

$$\frac{f(y)e^{-y^2/2}}{\int_{\mathbb{R}} f(x)e^{-x^2/2}dx}$$

Then by Theorem 9.2 from the appendix of the Math 541A notes (uniqueness of moment-generating functions), this describes a unique random variable. But this is a contradiction because we can't divide by 0. **(???)**

In the case that f is positive and negative at different points, we write $f = f_+ - f_-$ where $f_+(x) := \max\{f(x), 0\}$ and $f_-(x) := \max\{-f(x), 0\}$. use (1.13) for any μ and divide by case $\mu = 0$,

$$\int_{-\infty}^{\infty} f_-(y) e^{-y^2/2} e^{y\mu} dy, \qquad \int_{-\infty}^{\infty} f_-(y) e^{-y^2/2} dy$$

which yields

$$\int_{-\infty}^{\infty} f_+(y) e^{-y^2/2} e^{y\mu} dy \Big/ \int_{-\infty}^{\infty} f_-(y) e^{-y^2/2} dy$$

so we are done again by Theorem 9.2. So $f_y = f_-$, $f = f_+ - f_- = 0$.

So basically we started by assuming that the expression equals zero and concluded that f must equal 0; therefore the statistic is complete.

□

Remark 15. Later we will show that complete and sufficient statistics are minimal sufficient.

Exercise 3 (**Conditional expectation exercise relevant to proof of Bahadur's Theorem**)**.** Let $X, Y : \Omega \to \mathbb{R}$ both be discrete or both continuous. For all y in the range of Y, define $g(y) := \mathbb{E}(X \mid Y = y)$. Define the conditional expectation of X given Y, denoted $\mathbb{E}(X \mid Y)$, as the random variable $g(Y)$.

Solution: Theorem **??**

Theorem 1.3.4.1 (**Bahadur's Theorem; Theorem 5.25 in Math 541A notes**)**.** If Y is a complete sufficient statistic for a family $\{f_\theta : \theta \in \Theta\}$ of probability densities or probability mass functions, then Y is a minimal sufficient statistic for θ.

Remark 16. By Remark 5.11 in Math 541A notes, a complete sufficient statistic is unique up to an invertible map. Also by Example 5.15 in Math 541A notes, the converse of Bahadur's Theorem is false.

Proof. By Proposition 1.3.2.2 (Proposition 5.12 in Math 541A notes), there exists a minimal sufficient statistic Z for θ. To show that Y is minimal sufficient, it suffices to find a function r such that $Y = r(Z)$. Define $r(Z) = \mathbb{E}_\theta(Y \mid Z)$. Since Z is minimal sufficient and Y is sufficient by assumption, there exists a function u such that $Z = u(Y)$. By conditioning on Y we have by Exercise 3 (Exercise 5.24 in the Math 541A notes)

$$\mathbb{E}_\theta(r(u(Y)) = \mathbb{E}_\theta(r(Z)) = \mathbb{E}_\theta[\mathbb{E}_\theta(r(Z) \mid Y)] = \mathbb{E}_\theta[\mathbb{E}_\theta(\mathbb{E}_\theta(Y \mid Z) \mid Y)] = \mathbb{E}_\theta[\mathbb{E}_\theta(\mathbb{E}_\theta(Y \mid u(Y)) \mid Y)]$$

$$= \mathbb{E}_\theta[\mathbb{E}_\theta(Y \mid u(Y))] = \mathbb{E}_\theta(Y).$$

That is, $\mathbb{E}_\theta(r(u(Y) - Y) = 0$ for all $\theta \in \Theta$. Since Y is complete, we conclude that $r(u(Y)) = Y$, and since $r(u(Y)) = r(Z)$, we have $r(Z) = Y$, as desired.

□

Basu's theorem tells us that a complete sufficient statistic implies independence from any ancillary statistic. So complete sufficient statistics have no ancillary information, unlike minimal sufficient statistics.

Theorem 1.3.4.2 (**Basu's Theorem, Theorem 5.27 in Math 541A notes**)**.** Let $Y : \Omega \to \mathbb{R}^k$ and $Z : \Omega \to \mathbb{R}^m$ be statistics. If Y is a complete sufficient statistic for $\{f_\theta : \theta \in \Theta\}$ and Z is ancillary for θ, then for all $\theta \in \Theta$, Y and Z are independent with respect to f_θ.

Proof. Let $A \subseteq \mathbb{R}^k$ and $B \subseteq \mathbb{R}^m$. We need to show that

$$\mathbb{P}_\theta(Y \in A, Z \in B) = \mathbb{P}_\theta(Y \in A)\mathbb{P}_\theta(Z \in B), \qquad \forall \theta \in \Theta.$$

Note that

$$\mathbb{P}_\theta(Y \in A, Z \in B) = \mathbb{E}_\theta \mathbf{1}_{\{Y \in A\}} \mathbf{1}_{\{Z \in B\}} = \mathbb{E}_\theta \mathbb{E}_\theta(\mathbf{1}_{\{Y \in A\}} \mathbf{1}_{\{Z \in B\}} \mid Y) = \mathbb{E}_\theta[\mathbf{1}_{\{Y \in A\}} \mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}} \mid Y)].$$

Let $g(Y) := \mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}} \mid Y)$. Then

$$\mathbb{E}_\theta(g(Y)) = \mathbb{E}_\theta(\mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}} \mid Y)) = \mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}}) = \mathbb{P}_\theta(Z \in B). \tag{1.14}$$

Let $c := \mathbb{P}_\theta(Z \in B) = \mathbb{E}_\theta(g(Y)) = \mathbb{E}_\theta[\mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}} \mid Y)]$. Then c does not depend on θ since Z is ancillary by assumption. Then $\mathbb{E}_\theta(g(Y) - c) = 0, \forall\ \theta \in \Theta$ for all $\theta \in \Theta$. Note that $g(Y) := \mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}} \mid Y)$ does not depend on θ since Y is sufficient. Since Y is complete, $g(Y) - c - 0 \iff g(Y) = c$, so Y is constant. Therefore by (1.14)

$$c = \mathbb{E}_\theta(c) = \mathbb{E}_\theta(g(Y) = \mathbb{P}_\theta(Z \in B),$$

so we have

$$\mathbb{P}_\theta(Y \in A, Z \in B) = \mathbb{P}_\theta(Y \in A)g(Y) = \mathbb{P}_\theta(Y \in A)c = \mathbb{P}_\theta(Y \in A)\mathbb{P}_\theta(Z \in B), \qquad \forall \theta \in \Theta.$$

as desired.

□

Example 1.3.8 (Example from 541B). $X_1, \dots, X_n \sim \mathcal{N}(\mu, 1)$, i.i.d. Then $\overline{X}_n$ is complete sufficient for μ. Therefore by Basu's Theorem (Theorem 1.3.4.2) and Example 1.3.3, $\overline{X}_n$ and $X_1 - \overline{X}_n$ are independent.

Example 1.3.9. By Basu's Theorem (Theorem 1.3.4.2), the sample mean and sample variance of a random sample of a Gaussian random variable are independent.

Theorem 1.3.4.3 (Complete statistics in the exponential family; Theorem 6.2.25 in Casella and Berger [2001]). Let $X_1, \dots, X_n$ be i.i.d. observations from an exponential family with pdf or pmf of the form

$$f(x \mid \theta) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} w(\theta_j)t_j(x)\right)$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$, $\boldsymbol{\theta} \in \Theta$. Then the statistic

$$T(X) = \left(\sum_{i=1}^{n} t_1(X_i), \dots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is complete as long as the parameter space Θ contains an open set in $\mathbb{R}^k$.

1.4 Point Estimation

Definition 1.4.1 (Point estimator). Let $X_1, \dots, X_n$ be a random sample of size n from a family of distribution $\{f_\theta : \theta \in \Theta\}$. If Y is a statistic that is used to estimate the parameter θ that fits the data at hand, we then refer to Y as a **point estimator** or **estimator**.

1.4.1 Heuristic Principles for Finding Good Estimators

Definition 1.4.2 (**Likelihood, Definition 6.1 in Math 541A notes**). Let $X_1, \dots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta : \theta \in \Theta\}$. If we have data $x \in \mathbb{R}^n$, then the function $L : \Theta \to [0, \infty)$ defined by $L(\theta) := f_\theta(x)$ is called the **likelihood function**.

- **Likelihood principle:** All data relevant to estimating the parameter θ is contained in the likelihood function.
- **Sufficiency principle:** If $Y = t(X_1, \dots, X_n)$ is a sufficient statistic and if we have two results $x, y \in \mathbb{R}^n$ from an experiment with the same statistics $t(x) = t(y)$, then our estimate of the parameter θ should be the same for either experimental result.
- **Equivariance principle:** If the family of distributions $\{f_\theta : \theta \in \Theta\}$ is invariant under some symmetry, then the estimator of θ should respect the same symmetry. (For example, a location family is invariant under translation, so an estimator for the location parameter should commute with translations.)

1.4.2 Evaluating Estimators

We can enumerate several desirable properties for estimators.

Definition 1.4.3 (**Unbiasedness, Definition 6.2 in Math 541A notes**). Let $X_1, \dots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta : \theta \in \Theta\}$. Let $t : \mathbb{R}^n \to \mathbb{R}^k$ and let $Y := t(X_1, \dots, X_n$ be an estimator for $g(\theta)$. Let $g : \Theta \to \mathbb{R}^k$. We say that Y is **unbiased** for $g(\theta)$ if $\mathbb{E}_\theta Y = g(\theta)$ for all $\theta \in \Theta$.

One common way to check the quality of an estimator is the mean squared error, or squared L_2 norm, of the estimator minus θ, $\mathbb{E}_\theta(Y - g(\theta))^2$. If the estimator is unbiased, this quantity is equal to the variance of Y.

Definition 1.4.4 (**UMVU, sometimes called MVUE (minimum variance unbiased estimator); Definition 6.3 in Math 541A Notes**). Let $X_1, \dots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta : \theta \in \Theta\}$. Let $g : \Theta \to \mathbb{R}$. Let $t : \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X_1, \dots, X_n)$ be an unbiased estimator for $g(\theta)$. We say that Y is **uniformly minimum variance unbiased (UMVU)** if, for any other unbiased estimator Z for $g(\theta)$, we have $\text{Var}_\theta(Y) \leq \text{Var}_\theta(Z)$ for all $\theta \in \Theta$.

Remark 17. The "uniform" property has to do with the fact that this inequality must hold for every $\theta \in \Theta$ (as opposed to for a particular θ, or averaged over all $\theta \in \Theta$, or something like that).

Example 1.4.1 (**Math 541B example**). Suppose $X \sim \text{Bin}(3, \theta)$, $\theta \in [0, 1]$. Does there exist an unbiased estimator of θ^5? No. Assume that $T(X)$ is unbiased for θ^5, so that

$$\mathbb{E}_\theta T(x) = \sum_{j=0}^{3} T(j) \binom{3}{j} \theta^j (1-\theta)^{3-j} = \theta^5 \qquad \forall \theta \in [0, 1].$$

But this cannot be, because the quantity on the left side of the last equality is a third degree polynomial in θ and the quantity on the right side is a 5th degree polynomial. So the only powers of θ that are possible to estimate unbiasedly in this case are powers less than or equal to 3.

Example 1.4.2 (Math 541B example). $X_1, \dots, X_n \sim \mathcal{N}(\mu, 1), \mu \in \mathbb{R}$. Let $g(\mu) = \mathbb{P}_\mu(X_1 \leq x)$. Find a UMVUE of $g(\mu)$.

Solution

Consider $S(X_1, \dots, X_n) = I\{X_1 \leq x\}$ (an indicator function). Then $\mathbb{E}_\mu S(X_1, \dots, X_n) = g(\mu)$. Define

$$\hat{S}(X_1, \dots, X_n) := \mathbb{E}\left[I\{X_1 \leq x\} \mid \overline{X}_n\right].$$

Then by Lehmann-Scheffe (Theorem 1.4.2.3), $\hat{S}$ is UMVU. Note that

$$\hat{S}(X_1, \dots, X_n) = \mathbb{E}\left[I\{X_1 - \overline{X}_n \leq x - \overline{X}_n\} \mid \overline{X}_n\right]$$

By Example 1.3.3, $X_1 - \overline{X}_n$ is independent of $\overline{X}_n$, so we can write

$$= \mathbb{E}\left[I\{X_1 - \overline{X}_n \leq x - \overline{X}_n\}\right] = \mathbb{P}\left[X_1 - \overline{X}_n \leq x - \overline{X}_n\right] = \phi\left(\sqrt{\frac{n}{n+1}}(x - \overline{X}_n)\right)$$

where $\phi(\cdot)$ is the cdf of a standard Gaussian random variable.

More generally, given a family of distributions $\{f_{\tilde{\theta}} : \tilde{\theta} \in \Theta\}$, we could be given a **loss function** $L(\theta, y) : \Theta \times \mathbb{R}^k \to \mathbb{R}$ and be asked to minimize the **risk function** $r(\theta, Y) := \mathbb{E}_{\tilde{\theta}}\big(\ell(\theta, Y)\big)$ over all possible estimators Y. In the case of mean squared error loss, we have $L(\theta, y) := (y - g(\theta))^2$ for all $y, \theta \in \mathbb{R}$.

The Rao-Blackwell Theorem says that if $L(\theta, y)$ is convex in y then we can create an optimal estimator for $g(\theta)$ from a sufficient statistic and any estimator for $g(\theta)$ (we can lower the risk of an estimator Y by conditioning on a sufficient statistic Z.).

Theorem 1.4.2.1 (Rao-Blackwell; Theorem 6.4 in Math 541A notes). Let Z be a sufficient statistic for $\{f_\theta : \theta \in \Theta\}$ and let Y be an estimator for $g(\theta)$. Define $W := \mathbb{E}_\theta(Y \mid Z)$. Let $\theta \in \Theta$. Then

$$\text{Var}_\theta(W) \leq \text{Var}_\theta(Y).$$

Further, let $r(\theta, y) < \infty$ and such that $\ell(\theta, y)$ is convex in y. Then

$$r(\theta, W) \leq r(\theta, Y).$$

Proof. To get the first statement, note that

$$\begin{aligned}
\text{Var}_\theta(W) &= \mathbb{E}_\theta\left[W - g(\theta)\right]^2 \\
&= \mathbb{E}_\theta\left[\mathbb{E}_\theta(Y \mid Z) - g(\theta)\right]^2 \\
&= \mathbb{E}_\theta\left[\mathbb{E}_\theta(Y - g(\theta) \mid Z)\right]^2 \\
&\leq \mathbb{E}_\theta\left[\mathbb{E}_\theta([Y - g(\theta)]^2 \mid Z)\right] \\
&= \mathbb{E}_\theta\left[(Y - g(\theta))^2\right] \\
&= \text{Var}_\theta(Y),
\end{aligned}$$

where the inequality follows that $\text{Var}(W) = \mathbb{E}W^2 - (\mathbb{E}W)^2 \geq 0$ for any random variable W (in this case, $W = Y - g(\theta) \mid Z$, and there is equality only if $\text{Var}(W) = 0$; that is, $Y - g(\theta)$ can take just one value for each value of T, so Y is a deterministic function of T.

To get the second statement, note that since Z is sufficient, W does not depend on θ. By the Conditional Jensen's Inequality (Theorem **??**) and using the convexity of $\ell(\theta, y)$ in y,

$$\ell(\theta, w) = \ell(\theta, \mathbb{E}_{\tilde{\theta}}(Y \mid Z)) \leq \mathbb{E}_{\tilde{\theta}}[\ell(\theta, Y) \mid Z].$$

Take expectations of both sides to get

$$\mathbb{E}_{\tilde{\theta}}\ell(\theta, w) = r(\theta, W) \leq \mathbb{E}_{\tilde{\theta}}\mathbb{E}_{\tilde{\theta}}[\ell(\theta, Y) \mid Z] = \mathbb{E}_{\tilde{\theta}}\ell(\theta, Y) = r(\theta, Y).$$

□

Definition 1.4.5 (**Definition 6.4 in 541A notes; MISSED SOME NOTES TODAY**)**.** We say Y is **uniformly minimum risk unbiased** (UMRU) if for any other unbiased estimator Z for $g(\theta)$,

$$r(\theta, Y) \leq r(\theta, z), \qquad \forall\ \theta \in \Theta$$

Remark 18. Unfortunately, UMRU or UMVU may not exist. More fundamentally, an unbiased estimator for $g(\theta)$ may not exist. For example, let X be a binomial random variable with known n, unknwon $0 < \theta < 1$, and $g(\theta) = \theta/(1-\theta)$. Then no unbiased estimator exists for $g(\theta)$. Why?

$$\mathbb{E}_\theta t(X) = \sum_{j=0}^{n} t(j)\binom{n}{j}\theta^j(1-\theta)^{n-j}, \qquad \forall \theta \in \Theta, \text{ by definition of } X.$$

where the summation is a polynomial of degree at most n in θ. Then it is impossible to have $\mathbb{E}_\theta t(x) = g(\theta)$ when $g(\theta) = \theta/(1-\theta)$.

Recall the definition of strict convexity (Definition **??**).

Theorem 1.4.2.2 (**Rao-Blackwell restated; Theorem 6.7 in Math 541A notes**)**.** Let Z be a sufficient statistic for $\{f_\theta : \theta \in \Theta\}$ and let Y be an estimator for $g(\theta)$. Define $W := \mathbb{E}_\theta(Y \mid Z)$. Let $\theta \in \Theta$ with $r(\theta, y) < \infty$ and such that $\ell(\theta, y)$ is convex in y. Then

$$r(\theta, W) \leq r(\theta, Y).$$

Further, if $\ell(\theta, y)$ is strictly convex in $y \in \mathbb{R}$, then $r(\theta, W) < r(\theta, Y)$ unless $W = Y$ (that is, there is a unique minimizer of the risk).

So Z makes the estimator better. Question: can we construct $\mathbb{E}_\theta(Y \mid Z)$ to be UMRU?

Remark 19 (**Remark 6.9 in Math 541A notes**)**.** $\mathbb{E}_\theta W = \mathbb{E}_\theta \mathbb{E}_\theta(Y \mid Z) = \mathbb{E}_\theta Y$. So if Y is unbiased for $g(\theta)$, then so is W.

Remark 20. What happens if Z is constant in Rao-Blackwell? Then in general Z will not be sufficient, so W might depend on θ which is not allowed. Put another way, if Z has insufficient information, then W gets messed up **(???)**.

Example 1.4.3. Let $X_1, \ldots, X_n$ be a random sample with unknown mean $\mu \in \mathbb{R}$. We want to construct an estimator for μ using Rao-Blackwell. Let $t : \mathbb{R}^n \to \mathbb{R}$ so that $t(x_1, \ldots, x_n) = x_1$ for all $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Let $Y = t(X_1, \ldots, X_n) = X_1$. Note that Y is unbiased. First use of Rao-Blackwell: use $Z = (X_1, \ldots, X_n)$. Then by Exercise 5.24,

$$W := \mathbb{E}_\mu(X_1 \mid (X_1, \ldots, X_n)) = \mathbb{E}(X_1 \mid X_1) = X_1.$$

We can think of this as failing to improve the estimator because we used "too much" information. Second try: use $Z = \sum_{i=1}^n X_i$. Note that in the Gaussian case Z is sufficient for μ and unbiased for $n\mu$. Since $X_1, \ldots, X_n$ are i.i.d. for all $1 \leq k \leq \ell \leq n$ the joint distribution of $(X_k, \sum_{i=1}^n X_i)$ is the same as the joint distribution of $(X_\ell, \sum_{i=1}^n X_i)$. So

$$\mathbb{E}(X_k \mid \sum_{i=1}^n X_i) = \mathbb{E}(X_\ell \mid \sum_{i=1}^n X_i).$$

So we have

$$W := \mathbb{E}_\mu\left(X_1 \mid \sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{j=1}^n \mathbb{E}_\mu\left(X_j \mid \sum_{i=1}^n X_i\right) = \frac{1}{n}\mathbb{E}_\mu\left(\sum_{j=1}^n X_j \mid \sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n X_i.$$

So we started with a trivial estimator X_1 and ended up with the sample mean using Rao-Blackwell.

Theorem 1.4.2.3 (**Lehmann-Scheffe, Theorem 6.13 in Math 541A notes**)**.** Let Z be a complete sufficient statistic for a family of distributions $\{f_\theta : \theta \in \Theta\}$. Let Y be an unbiased estimator for $g(\theta)$. Define $W := \mathbb{E}_\theta(Y \mid Z)$. (Since Z is sufficient, W does not depend on θ.) Then W is UMRU for $g(\theta)$. Further, if $\ell(\theta, y)$ is strictly convex in y for all $\theta \in \Theta$, then W is unique. In particular, W is the unique UMVU for $g(\theta)$.

Proof. W is unbiased by Remark 6.10 in math 541A notes ("$\mathbb{E}_\theta W = \mathbb{E}_\theta \mathbb{E}_\theta(Y \mid Z) = \mathbb{E}_\theta Y$. So if Y is unbiased for $g(\theta)$, then so is W.") We first show W does not depend on Y. Let Y' be an unbiased estimator for $g(\theta)$. We show that $\mathbb{E}_\theta(Y \mid Z) = \mathbb{E}_\theta(Y' \mid Z)$ for all $\theta \in \Theta$. Note that

$$\mathbb{E}_\theta(\mathbb{E}_\theta(Y \mid Z) - \mathbb{E}_\theta(Y' \mid Z)) = \mathbb{E}_\theta(Y - Y') = g(\theta) - g(\theta) = 0, \ \forall \ \theta \in \Theta$$

Note that $\mathbb{E}_\theta(Y \mid Z)$ and $\mathbb{E}_\theta(Y' \mid Z)$ are functions of Z. Therefore since Z is complete, $\mathbb{E}_\theta(Y \mid Z) = \mathbb{E}_\theta(Y' \mid Z)$ for all $\theta \in \Theta$.

Next, by Rao Blackwell,

$$r(\theta, Y') = r(\theta, \mathbb{E}_\theta(Y' \mid Z)) = r(\theta, \mathbb{E}_\theta(Y \mid Z)) = r(\theta, W), \ \forall \theta \in \Theta.$$

□

Remark 21 (**Remark 6.14 in Math 541A notes, Theorem 7.3.23 in Casella and Berger [2001, p. 347]**). Let $Z : \Omega \to \mathbb{R}^k$ be a complete sufficient statistic for $\{f_\theta : \theta \in \Theta\}$ and let $h : \mathbb{R}^k \to \mathbb{R}^m$. Let $g(\theta) := \mathbb{E}_\theta h(Z)$ for all $\theta \in \Theta$. Then $h(Z)$ is unbiased for $g(\theta)$, since $\mathbb{E}_\theta h(Z) = g(\theta) = \mathbb{E}_\theta(g(\theta))$. Applying Theorem 1.4.2.3, we have

$$W := \mathbb{E}_\theta(h(Z) \mid Z) = \mathbb{E}_\theta(\mathbb{E}_\theta[h(Z) \mid h(Z)] \mid Z) = \mathbb{E}_\theta[h(Z) \mid h(Z)] = h(Z).$$

Therefore by Theorem 1.4.2.3, $h(Z)$ is UMVU for $g(\theta)$. That is, any function of a complete sufficient statistic is UMVU for its expected value. So one way to find a UMVU is to come up with a function of a complete sufficient statistic that is unbiased for a given function $g(\theta)$.

Summary of methods for finding UMVU (given a complete sufficient statistic Z, want to estimate $g(\theta)$

(1) **(Condition method/Rao-Blackwell):** Follow Theorem 1.4.2.3: find an unbiased Y and let $W := \mathbb{E}_\theta(Y \mid Z)$. (problem: can be hard to find an unbiased Y.)

(2) Solve for $h : \mathbb{R}^k \to \mathbb{R}$ satisfying

$$\mathbb{E}_\theta h(Z) = g(\theta) \tag{1.15}$$

by the above remark, (1.15) will also give you the UMVU. Then $h(Z)$ is UMVU for $g(\theta)$. By "solve", consider that we have g and Z and somehow solve for the h satisfying (1.15). For example if Z is binomial the left side of (1.15) will be the sum of a bunch of numbers. Find the h values that satisfy (1.15), if possible.

(3) **(Luck method):** Somehow guess the h such that (1.15) is satisfied.

Example 1.4.4. Suppose we are sampling from a Gaussian distribution with unknown mean and variance. By the Factorization Theorem and Exercise 5.23 (**on homework 4**), we know $(\overline{X}, S^2)$ is complete sufficient of (μ, σ^2) (sufficiency follows by the Factorization Theorem (Theorem 1.3.1.3), completeness follows by the exercise). For example, using method (2) above, $\overline{X}$ is UMVU for μ (with finite σ) by method (3) above, since $\overline{X}$ is a function of $(\overline{X}, S^2)$, $h(x, y) := x$, $g(\mu, \sigma^2) := \mu$ (then (1.15) is satisfied). Similarly, S^2 is UMVU for σ^2 be (3) using $h(x, y) := y$, $g(\mu, \sigma^2) := \sigma^2$ (then (1.15) is satisfied).

Suppose we want a UMVU for μ^2. Try guessing $(\overline{X})^2$ as an estimator. Note that

$$\mathbb{E}[(\overline{X})^2] = \mathbb{E}\bigg(\frac{1}{n}\sum_{i=1}^n X_i\bigg)^2 = \frac{1}{n^2}\bigg(\sum_{i=1}^n \mathbb{E}X_i^2 + 2\sum_{1\le i<j\le n}\mathbb{E}X_i\mathbb{E}X_j\bigg) = \ldots = \mu^2 + \sigma^2/n$$

So,

$$\mathbb{E}\bigg(\overline{X}^2 - S^2/n\bigg) = \mu^2$$

which means that $\overline{X}^2 - S^2/n$ is UMVU since it is a function of $(\overline{X}, S^2)$.

Example 1.4.5. Try Method (2). Let X be a binomial random variable with known parameter n and unknown $0 < \theta < 1$. Suppose we want to estimate $g(\theta) := \theta(1-\theta)$. Solve (1.15): find the h satisfying

$$\theta(1-\theta) = \mathbb{E}_\theta h(X)$$

$$\iff \theta(1-\theta) = \sum_{j=0}^n h(j)\binom{n}{j}\theta^j(1-\theta)^{n-j}, \qquad \forall \theta \in (0,1)$$

For convenience, let $a := \theta/(1-\theta)$, so that $a(1-\theta) = \theta \iff \theta = a/(1+a)$, $1-\theta = 1/(1+a)$. Then we have

$$(1-\theta)^{-n} = \sum_{j=0}^n h(j)\binom{n}{j}a^j = \theta(1-\theta)^{1-n} = \frac{a}{1+a}\bigg(\frac{1}{1+a}\bigg)^{1-n} = a(1+a)^{n-2} \tag{1.16}$$

So we want to solve for $h(j)$ so that for all $a > 0$,

$$\sum_{j=0}^n h(j)\binom{n}{j}a^j = a(1+a)^{n-2} = \text{ (RHS of (1.16), by binomial theorem) } a\sum_{j=0}^{n-2}\binom{n-2}{j}a^j = \sum_{j=1}^{n-1}\binom{n-2}{j}-1a^j$$

We need the coefficients to match up one by one. So $h(0) = h(n) = 0$, and then it works if $h(j)\binom{n}{j} = \binom{n-2}{j-1}$ for all $j \in 1, \ldots, n-1$. So

$$h(j) = \frac{\binom{n-2}{j-1}}{\binom{n}{j}} = \frac{(n-2)!}{n!}\frac{(n-j)!j!}{|}(n-j-1)!(j-1)! = \frac{(n-j)j}{n(n-1)}$$

So in fact,

$$h(j) = \frac{(n-j)j}{n(n-1)}, \qquad \forall\ 0 \le j \le n$$

Therefore the UMVU for $\theta(1-\theta)$ is

$$\frac{X(n-X)}{n(n-1)}.$$

Example 1.4.6. Try Method (1). Suppose we have n independent samples $X_1, \dots, X_n$ from a Bernoulli distribution with unknown $\theta \in (0,1)$. From Example **??** (Example 3.15 in Math 541A notes) and Exercise 5.23 in Math 541A notes, a complete sufficient statistic is $Z := \sum_{i=1}^n X_i$ is complete and sufficient for θ. Also, $(1/n)\sum_{i=1}^n X_i$ is unbiased for θ. So, $(1/n)\sum_{i=1}^n X - i$ is UMVU for θ. Suppose we want to estimate θ^2. We need an unbiased estimator Y for θ^2. Let $Y = X_1X_2$. Then $\mathbb{E}(Y) = \mathbb{E}(X_1)\mathbb{E}(X_2) = \theta^2$. By Theorem 1.4.2.3, $W := \mathbb{E}_\theta(Y \mid Z)$ is UMVU for σ^2. Note, $Y = 1$ when $X_1 = X_2 = 1$ and 0 otherwise. So

$$\mathbb{E}_\theta(Y \mid Z = z) = \mathbb{E}_\theta(\mathbf{1}_{\{X_1=X_2=1\}} \mid Z = z) = \mathbb{P}_\theta(X_1 = X_2 = 1 \mid Z = z) = \mathbb{P}_\theta(X_1 = X_2 = 1 \mid \sum_{i=1}^n X_i = z)$$

$$= \frac{1}{\mathbb{P}_\theta\left(\sum_{i=1}^n X_i = z\right)} \cdot \mathbb{P}_\theta\left(X_1 = X_2 = 1 \cap \sum_{i=1}^n X_i = z\right)$$

$$= \frac{1}{\mathbb{P}_\theta\left(\sum_{i=1}^n X_i = z\right)} \cdot \mathbb{P}_\theta\left(X_1 = X_2 = 1 \cap \sum_{i=3}^n X_i = z-2\right)$$

$$= \frac{\theta^2\binom{n-2}{z-2}\theta^{z-2}(1-\theta)^{n-z}}{\binom{n}{z}\theta^z(1-\theta)^{n-z}} = \frac{\binom{n-z}{z-2}}{\binom{n}{z}} = \frac{(n-z)!(n-z)!z!}{n!(n-z)!(z-2)!} = \frac{z(z-1)}{n(n-1)}$$

So we have shown that for all $0 \le Z \le n$,

$$\mathbb{E}_\theta(Y \mid Z = z) = \frac{z(z-1)}{n(n-1)}.$$

So, by Theorem 1.4.2.3,

$$W := \mathbb{E}_\theta(Y \mid Z) = \frac{Z(Z-1)}{n(n-1)}$$

is UMVU for θ^2.

Question: If W_1 is UMVU for $g_1(\theta)$ and W_2 is UMVU for $g_2(\theta)$, is $W_1 + W_2$ UMVU for $g_1(\theta) + g_2(\theta)$? If there is a complete sufficient statistic, then by Lehmann-Scheffe (Theorem 1.4.2.3), $W_1 = \mathbb{E}_\theta(Y_1 \mid Z)$, $W_2 = \mathbb{E}(Y_2 \mid Z)$ where Y_1 is unbiased for g_1, Y_2 is unbiased for g_2. Then

$$W_1 + W_2 = \mathbb{E}_\theta(Y_1 + Y_2 \mid Z).$$

Note: $\mathbb{E}_\theta(Y-1+Y_2) = \mathbb{E}_\theta Y_1 + \mathbb{E}_\theta Y_2 = g_1(\theta) = g_Z 2(\theta)$. So W_1+W_2 is UMVU by Lehmann-Scheffe (Theorem 1.4.2.3) for $g_1(\theta) + g_2(\theta)$. But is it true if we don't have a complete sufficient statistic (and this argument doesn't apply)? **yes, by the theorem below; this condition will clearly hold across sums.**

Theorem 1.4.2.4 (**Alternate Characterization of UMVU; Theorem 6.18 in Math 541A notes**). Let $f \in \{f_\theta : \theta \in \Theta\}$ be a family of distributions and let $g : \Theta \to \mathbb{R}$. Let W be an unbiased estimator for $g(\theta)$ (note that the existence of an unbiased estimator is a nontrival assumption). Let $L_2(\Omega)$ be the set of statistics with finite second moment. Then $W \in L_2(\Omega)$ is UMVU for $g(\theta)$ if and only if for any $\theta \in \Theta$,

$$\mathbb{E}_\theta(WU) = 0, \qquad \forall U \in L_2(\Omega) \text{ that are unbiased estimators of } 0$$

Thinking of this as an inner product, we have to be orthogonal to all such U.

Proof. Assume W is UMVU for $g(\theta)$. Let U be an unbiased estimator of 0. Let $s \in \mathbb{R}$, consider $W + sU$. Note that $W + sU$ is also unbiased for $g(\theta)$. Since W is UMVU,

$$\text{Var}_\theta(W) \leq \text{Var}_\theta(W + sU) = \text{Var}_\theta(W) + s^2\text{Var}_\theta(U) + 2s\text{Cov}_\theta(W, U)$$

$$= \text{Var}_\theta(W) + s^2\text{Var}(U) + 2s\mathbb{E}_\theta[(W - \mathbb{E}_\theta(W))U], \qquad \forall \theta \in \Theta.$$

Note that we have equality when $s = 0$. Also, the derivative of the right side with respect to s must be 0 when $s = 0$ or else the inequality does not hold (the minimum value occurs at $s = 0$ if and only if the derivative of the right side in s is 0 at $s = 0$). Note that the derivative of the right side is

$$0 = 2\mathbb{E}_\theta[(W - \mathbb{E}_\theta W)U] = 2\mathbb{E}_\theta(WU).$$

The converse is also true because this reasoning can be reversed, since if Y is any unbiased estimator for $g(\theta)$, then $U := W - Y$ is an unbiased estimator for 0, and $Y = W + sU$ with $s = 1$. We have

$$\text{Var}_\theta(Y) = \text{Var}_\theta(W - U) = \text{Var}_\theta W + \text{Var}_\theta U + 2\text{Cov}_\theta(W, U) = \text{Var}_\theta W + \text{Var}_\theta U + 2\mathbb{E}_\theta(WU)$$

So $\text{Var}_\theta Y \geq \text{Var}_\theta W$ for all $\theta \in \Theta$.

□

Remark 22. If we have a complete sufficient statistic, better to use the earlier methods in general (unless it is really complicated to work with). If we don't have a complete sufficient statistic, use this theorem.

1.4.3 Efficiency of an Estimator

Another desirable property of an estimator is high efficiency—"good" with a small number of samples. One way to quantify this notion is to define a notion of "information" and try to maximize the information content of the estimator.

Definition 1.4.6 (**Fisher Information, Definition 6.19 in Math 541A notes**). Let $f \in \{f_\theta : \theta \in \Theta\}$ be a family of multivariate probability densities or probability mass functions. Assume $\Theta \subseteq \mathbb{R}$ (this is a one-parameter situation). Let X be a random variable with distribution f_θ. Define the **Fisher information** of the family to be

$$I(\theta) = I_X(\theta) := \mathbb{E}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right)^2, \qquad \forall \theta \in \Theta$$

if this quantity exists and is finite.

(See also Section **??** for connections to generalized linear models.)

Remark 23. Note that if X is continuous,

$$\mathbb{E}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right) = \int_{\mathbb{R}^n} \frac{1}{f_\theta(x)}\frac{\mathrm{d}}{\mathrm{d}\theta} f_\theta(x) \cdot f_\theta(x) dx = \int_{\mathbb{R}^n} \frac{\mathrm{d}}{\mathrm{d}\theta} f_\theta(x) dx = \frac{\mathrm{d}}{\mathrm{d}\theta}\int_{\mathbb{R}^n} f_\theta(x) dx = \frac{\mathrm{d}}{\mathrm{d}\theta} 1 = 0.$$

So we could have equivalently defined the Fisher information as

$$I_X(\theta) = \mathrm{Var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right)$$

Example 1.4.7 (**Example 6.20 in Math 541A notes**). Let $\theta > 0$, let

$$f_\theta(x) := \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right), \qquad \forall \theta \in \mathbb{R} = \Theta,\ \forall x \in \mathbb{R}.$$

Then

$$I(\theta) = \mathrm{Var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta} - \frac{(x-\theta)^2}{2\sigma^2}\right) = \frac{1}{\sigma^4}\mathrm{Var}_\theta(x-\theta) = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

Observe that a small σ means large $I(\theta)$ in this case.

Proposition 1.4.3.1 (**Proposition 6.21 in Math 541A notes**). Let X be a random variable with distribution from $\{f_\theta : \theta \in \Theta\}$ (densities or mass functions). Let Y be a random variable with distribution from $\{g_\theta : \theta \in \Theta\}$ (densities or mass functions. Assume $\Theta \subseteq \mathbb{R}$ (one parameter in θ). If X and Y are independent, then

$$I_{(X,Y)}(\theta) = I_X(\theta) I_Y(\theta).$$

Proof. This proof will just be for the case of densities (continuous random variables; the case for probability mass functions is similar). Since X and Y are independent, (X, Y) has a distribution with density $f_\theta(x)g_\theta(y)$, for all $x, y \in \mathbb{R}$. Also, $\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)$ and $\frac{\mathrm{d}}{\mathrm{d}\theta}\log g_\theta(Y)$ are independent for all $\theta \in \Theta$. So,

$$I_{(X,Y)}(\theta) = \mathrm{Var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log[f_\theta(X)g_\theta(Y)]\right) = \mathrm{Var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log[f_\theta(X)] + \frac{\mathrm{d}}{\mathrm{d}\theta}\log[g_\theta(Y)]\right)$$

By independence we can write

$$= \text{Var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log[f_\theta(X)]\right) + \text{Var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log[g_\theta(Y)]\right) = I_X(\theta) + I_Y(\theta).$$

□

Remark 24. This is consistent with a notion of "information" since if variables are independent, the information is the sum of the information of each variable. This proof also shows the main reason why the logarithm is in the definition of the Fisher information—it brings a product to a sum.

Proposition 1.4.3.2 (**Exercise 6.22 in Math 541A notes**)**.** Let X be a random variable with distribution from $\{f_\theta : \theta \in \Theta\}$ (densities or mass functions). Let Y be a random variable with distribution from $\{g_\theta : \theta \in \Theta\}$ (densities or mass functions). Then

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_{Y|X=x}(\theta), \qquad \forall \theta \in \Theta, x \in \mathbb{R}.$$

Proof. Recall that $Y \mid X$ has density $f_{X,Y}(x,y)/f_X(x)$ for any fixed x. And if X, Y are discrete random variables, recall that $Y \mid X$ has mass function $\mathbb{P}(X = x, Y = y)/\mathbb{P}(Y = y)$.

$$I_{(X,Y)}(\theta) = \text{Var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log[f_{\theta(X,Y)}(x,y)]\right) = \text{Var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log[f_\theta(x) f_{\theta(Y|X=x)}(y)]\right)$$

$$= \text{Var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log[f_\theta(X)] + \frac{\mathrm{d}}{\mathrm{d}\theta}\log[f_{\theta(Y|X=x)}(y)]\right)$$

Note that $Y \mid X = x$ is independent of X (because we have conditioned out the dependence). Therefore we have

$$= \text{Var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log[f_\theta(X)]\right) + \text{Var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log[f_{\theta(Y|X=x)}(y)]\right) = I_X(\theta) + I_{Y|X=x}(\theta).$$

□

Theorem 1.4.3.3 (**Cramer-Rao/Information Inequality, Theorem 6.23 in Math 541A Notes**)**.** Let $X : \Omega \to \mathbb{R}^n$ be a random variable with distribution from a family of multivariable probability densities or probability mass functions $\{f_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$ and Θ an open interval. Let $t : \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X)$ be a statistic. For any $\theta \in \Theta$ let $g(\theta) := \mathbb{E}_\theta Y$. Assume $g'(\theta)$ and $\frac{\partial p_\theta(x)}{\partial \theta}$ exist for all x. Assume the set $A = \{x : p_\theta(x) = 0\}$ does not depend on θ, and assume the Fisher information $I_X(\theta)$ is finite and nonzero. Lastly, assume

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int T(x) p_\theta(x)\, dx = \int T(x)\frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(x)\, dx.$$

Then

$$\text{Var}_\theta(Y) \geq \frac{|g'(\theta)|^2}{I_X(\theta)}, \qquad \forall \theta \in \Theta.$$

In particular, if Y is unbiased for θ, then $g(\theta) = \theta$, so,

$$\text{Var}_\theta(Y) \geq \frac{1}{I_X(\theta)}, \qquad \forall \theta \in \Theta.$$

Equality occurs for some $\theta \in \Theta$ only when $\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(x)$ and $Y - \mathbb{E}_\theta Y$ are multiples of each other.

Remark 25. For a one-parameter family of distributions, the equality case of Theorem 1.4.3.3 gives a new way to find a UMVU that avoids any discussion of complete sufficient statistics. This is another way to find a UMVU ($\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X)$) that sidesteps the need for a complete sufficient statistic. That is, to find a UMVU. we look for affine functions of $\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X)$.

Proof.

$$|g'(\theta)| = \left| \frac{\mathrm{d}}{\mathrm{d}\theta} \int_{\mathbb{R}} f_\theta(x) t(x) dx \right| = \left| \int_{\mathbb{R}} \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(x)\right) t(x) f_\theta(x) dx \right| = \left| \mathbb{E}_\theta \frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X) t(X) \right|$$

Note that $\mathbb{E}_\theta \frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X) = 0$, so this can be written as

$$= \left| \text{Cov}_\theta \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X), t(X) \right) \right|$$

Then by Remark 1.63 in math 541A notes, by the Cauchy-Schwarz inequality,

$$\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] \leq \sqrt{\text{Var}_\theta(X)\text{Var}_\theta(Y)},$$

so we have

$$\left| \text{Cov}_\theta \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X), t(X) \right) \right| \leq \sqrt{\text{Var}_\theta \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X) \right) \text{Var}_\theta(Y)} = \sqrt{I_X(\theta)} \sqrt{\text{Var}_\theta(Y)}$$

$$\iff |g'(\theta)|^2 \leq I_X(\theta) \text{Var}_\theta(Y) \iff \text{Var}_\theta(Y) \geq \frac{|g'(\theta)|^2}{I_X(\theta)}, \qquad \forall \theta \in \Theta.$$

Recall that equality occurs in the Cauchy-Schwarz inequality if and only if $\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(x)$ is a constant multiple of $Y - \mathbb{E}_\theta Y$ with probability 1. (See also Corollary 7.3.15 in Casella and Berger [2001, p. 341].)

□

Example 1.4.8 (Example 6.24). Suppose $f_\theta(x) := \theta x^{\theta-1} \mathbf{1}_{0<x<1}$ for all $x \in \mathbb{R}, \theta > 0$. (This is a beta distribution with $\beta = 1$.) We have

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(x) = \frac{1}{\theta} + \log x, \qquad \forall 0 < x < 1.$$

A vector $X = (X_1, \ldots, X_n)$ of n independent samples from f_θ is distributed according to the product $\prod_{i=1}^n f_\theta(x_i)$, so that

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log\prod_{i=1}^{n} f_\theta(x_i) = \frac{\mathrm{d}}{\mathrm{d}\theta}\sum_{i=1}^{n}\log f_\theta(x_i) = \sum_{i=1}^{n}\left(\frac{1}{\theta}+\log x_i\right) = n\left(\frac{1}{\theta}+\frac{1}{n}\log\prod_{i=1}^{n} x_i\right), \qquad \forall 0 < x_i < 1, 1 \le i \le n.$$

Then by Theorem 1.4.3.3 (Theorem 6.23 in Math 541A notes), any function of $\frac{\mathrm{d}}{\mathrm{d}\theta}\log\prod_{i=1}^{n} f_\theta(X_i)$ (plus a constant) is UMVU of its expectation. So, e.g.

$$Y := -\frac{1}{n}\log\prod_{i=1}^{n} X_i$$

is UMVU of its expectation, and $\mathbb{E}_\theta Y = \theta^{-1}$ since $\mathbb{E}_\theta \frac{\mathrm{d}}{\mathrm{d}\theta}\log\prod_{i=1}^{n} f_\theta(X_i) = \mathbb{E}n\left(\frac{1}{\theta}+\frac{1}{n}\log\prod_{i=1}^{n} x_i\right) = 0$.

Definition 1.4.7 (**Efficiency, Defintion 6.25 in Math 541A notes**)**.** Let $X : \Omega \to \mathbb{R}$ be a random variable with distribution from a family of multivariable probability densities or probability mass functions $\{f_\theta : \theta \in \Theta\}$ with $\Theta \in \mathbb{R}$. Let $t : \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X)$ be a statistic. Define the **efficiency** of Y to be

$$\frac{1}{I_X(\theta)\mathrm{Var}_\theta(Y)}, \qquad \forall\theta \in \Theta$$

if this quantity exists and is finite. If Z is another statistic, we define the **relative efficiency** of Y to Z to be

$$\frac{I_X(\theta)\mathrm{Var}_\theta(Z)}{I_X(\theta)\mathrm{Var}_\theta(Y)} = \frac{\mathrm{Var}_\theta(Z)}{\mathrm{Var}_\theta(Y)}, \qquad \forall\theta \in \Theta.$$

Definition 1.4.8 (**Efficient estimator; Math 541B definition**)**.** We say that $\hat{\theta}_n$ is an **efficient** estimator of θ if and only if its variance achieves the Cramer-Rao lower bound **(and it is unbiased?)**.

Definition 1.4.9 (**Efficient estimator; Math 541B definition**)**.** We say that $\hat{\theta}_n$ is an **efficient** estimator of θ if and only if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right).$$

Superefficiency (Math 541B). Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$. Then $\overline{X}_n$ is efficient, and it is also asymptotically efficient. Now consider **Hodge's Estimator:**

$$T'(X_1, \ldots, X_n) := \begin{cases} \overline{X}_n & |\overline{X}_n| \ge n^{-1/4} \\ 0 & |\overline{X}_n| < n^{-1/4}. \end{cases}$$

Claim: $\sqrt{n}(T' - \theta) \xrightarrow{d} \mathcal{N}(0, 1)$ for $\theta \neq 0$. Moreover, for $r_n(T' - \theta) \xrightarrow{p} 0$ for any sequence r_n when $\mu = 0$.

1.4.4 Bayes Estimation

In Bayes estimation, the parameter $\theta \in \Theta$ is regarded as a random variable Ψ. The distribution of Ψ reflects our prior knowledge about the probable values of Ψ. Then, given that $\Psi = \theta$, the conditional distribution of $X \mid \Psi = \theta$ is assumed to be $\{f_\theta : \theta \in \Theta\}$, where $f_\theta : \mathbb{R}^n \to [0, \infty)$. Suppose $t : \mathbb{R}^n \to \mathbb{R}^k$ and we have a statistic $Y := t(X)$ and a loss function $\ell : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$. Let $g : \Theta \to \mathbb{R}^k$.

Definition 1.4.10 (**Bayes estimator, Defintion 6.26 in Math 541A notes**). A **Bayes estimator** Y for $g(\theta)$ with respect to Ψ is defined such that

$$\mathbb{E}\ell(g(\Psi), Y) \leq \mathbb{E}\ell(g(\Psi), Z)$$

for all estimators Z. Here the expectation is with respect to both Ψ and Y. Note that we have not made any assumptions about bias for Y or Z. To find a Bayes estimator, it is sufficient to minimize the conditional risk.

Remark 26. $t(X)$ can depend on Ψ.

Proposition 1.4.4.1 (**Proposition 6.27 in Math 541A notes**). Suppose there exists $t : \mathbb{R}^k \to \mathbb{R}$ such that for almost every $x \in \mathbb{R}^n$, $Y := t(X)$ minimizes

$$\mathbb{E}(\ell(g(\Psi), Z) \mid X = x)$$

over all estimators Z. Then $t(X)$ is a Bayes estimator for $g(\theta)$ with respect to Ψ.

Proof. By assumption,

$$\mathbb{E}(\ell(g(\Psi), t(X)) \mid X = x) \leq \mathbb{E}(\ell(g(\Psi), Y) \mid X = x)$$

for any estimator Y and for almost every x. Taking expected values of both sides, we get

$$\mathbb{E}\ell(g(\Psi, t(X)) \leq \mathbb{E}\ell(g(\Psi, Y).$$

□

Example 1.4.9 (**Example 6.29 in Math 541A notes**). Suppose $n = 1, g(\theta) = \theta$, and $\ell(\Psi, Y) = (\Psi - Y)^2$. The minimum value of

$$\mathbb{E}[(\Psi - Y(X))^2 \mid X = x] = \mathbb{E}(\Psi^2 - 2\Psi t(X) + (t(X))^2 \mid X = x)$$

$$= \mathbb{E}(\Psi^2 \mid X = x) - 2t(X)\mathbb{E}(\Psi \mid X = x) + (t(X))^2$$

(where we remove expressions with x from the expectation because we take x to be fixed) occurs when $t(x) = \mathbb{E}(\Psi \mid X = x)$. So in this specific case the Bayes estimator is $Y = t(X) = \mathbb{E}(\Psi \mid X)$.

Given that $\Psi = \theta < 0$, suppose X is uniform on the interval $[0, \theta]$. (Suppose X is a single sample $n = 1$ from this distribution.) Also assume that Ψ has the gamma distribution with parameters $\alpha = 2$ and $\beta = 1$, so that Ψ has density $\theta e^{-\theta} \mathbf{1}_{\{\theta > 0\}}$. The joint distribution of X and Ψ is then

$$f_{\Psi,X}(\theta, x) := \frac{1}{\theta} \mathbf{1}_{\{0 < x < \theta\}} \theta e^{-\theta} \mathbf{1}_{\{\theta > 0\}} = \mathbf{1}_{\{0 < x < \theta\}} e^{-\theta}.$$

The marginal distribution of X is then

$$f_X(x) = \mathbf{1}_{\{x > 0\}} \int_{-\infty}^{\infty} f_{\Psi,X}(\theta, x) d\theta = \mathbf{1}_{\{x > 0\}} \int_x^{\infty} e^{-\theta} d\theta = \mathbf{1}_{\{x > 0\}} e^{-x}.$$

So the conditional distribution of Ψ given X is

$$f_{\Psi | X = x}(\theta \mid x) = \frac{f_{\Psi,X}(\theta, x)}{f_X(x)} = \frac{e^{-\theta} \mathbf{1}_{\{0 < x < \theta\}}}{e^{-x} \mathbf{1}_{\{x > 0\}}} = e^{x - \theta} \mathbf{1}_{\{0 < x < \theta\}}.$$

So,

$$\mathbb{E}(\Psi \mid X = x) = \int_{-\infty}^{\infty} \theta f_{\Psi | X = x}(\theta \mid x) d\theta = e^x \int_x^{\infty} \theta e^{-\theta} d\theta = e^x((x + 1)e^{-x}) = x + 1.$$

So the Bayes estimator is $Y = t(X) = \mathbb{E}(\Psi \mid X) = X + 1$. This estimator minimizes $\mathbb{E}(Y - Z)^2$ over all estimators Z. $(\ell(a, b) = (a - b)^2, g(\theta) = \theta, \mathbb{E}\ell(g(\Psi), Z))$.

In contrast, the UMVU for one sample is $2X$ by Theorem 1.4.2.3, since $2X$ is complete sufficient and unbiased for θ and $\mathbb{E}_\theta(2X \mid 2X) = 2X$. ($X$ uniform on $[0, \theta]$, θ unknown, $\mathbb{E}_\theta X = \theta/2, \forall\ \theta < 0$.)

(Not obvious) For n samples, $(1 + n^{-1})X_{(n)}$ is the UMVU for θ. Note that $2X$ is recovered when $n = 1$. Remarks: this estimator seems to be sufficient because you could factorize it as in the Factorization Theorem (Theorem 1.3.1.3).

Exercise 4 (2018 DSO Statistics Group In-Class Screening Exam, Question 3). Suppose that given the vector μ, the random vector X has a normal distribution in $\mathbb{R}^n$ with mean μ and identity covariance matrix. We want to make inference about $\|\mu\|^2$.

(a) Find an unbiased estimate of $\|\mu\|^2$. Call this estimator $\hat{\delta}_{\text{unbiased}}$.

(b) Suppose that a Bayesian has a proper prior distribution for μ that is Gaussian with mean vector 0 and covariance kI, where k is any fixed positive real number and I is the identity matrix. He wants to minimize mean squared error (MSE). The estimator minimizing the MSE is the posterior mean of $\|\mu\|^2$, i.e., $\mathbb{E}(\|\mu\|^2 \mid X)$. Find this estimator. Call this estimator $\hat{\delta}_{\text{proper}}$.

(c) Suppose now the Bayesian uses the uniform prior (which is also called a "flat" or "noninformative" prior) for μ. Report $\mathbb{E}(\|\mu\|^2 \mid X)$ in this case. Call it $\hat{\delta}_{\text{flat}}$. Report $\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}}$.

(d) Now, if the true distribution of μ is indeed Gaussian with mean vector 0 and covariance kI, then show that with respect to the unconditional (i.e. marginal) distribution of X, the Bayes estimator $\hat{\delta}_{\text{proper}}$ is closer in Euclidean distance to $\hat{\delta}_{\text{unbiased}}$ than it is to $\hat{\delta}_{\text{flat}}$ when n is large. That is, show

$$\mathbb{E}\left(\hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}}\right)^2 < \mathbb{E}\left(\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}}\right)^2$$

for large n, where the expectation is over the unconditional distribution of X, which is

$$\int_{\mathbb{R}^n} f(x \mid \mu)\pi(\mu)\, d\mu$$

with $f(x \mid \mu) = \mathcal{N}_n(\mu, I$ and $\pi(\mu) = \mathcal{N}_n(0, kI)$. (Hint: let $\hat{D} = \hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}}$. Compute the mean and variance of $\hat{D}$ under the unconditional distribution of X.)

Solution

(a) We have

$$X \mid \mu \sim \mathcal{N}(\mu, \boldsymbol{I}_n)$$

Let $X = (X_1, \ldots, X_n)^T$ and let $\mu = (\mu_1, \ldots, \mu_n)^T$. Notice that

$$\mathbb{E}(\boldsymbol{X}^T\boldsymbol{X}) = \mathbb{E}[\mathbb{E}(\boldsymbol{X}^T\boldsymbol{X} \mid \mu)] = \mathbb{E}\left[\mathbb{E}\left(X_1^2 + X_2^2 + \ldots + X_n^2 \mid \mu\right)\right] = \mathbb{E}\left[\sum_{i=1}^n \mathbb{E}(X_i^2 \mid \mu)\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^n \text{Var}(X_i \mid \mu) + \mathbb{E}(X_i \mid \mu)^2\right] = \mathbb{E}\left[\sum_{i=1}^n 1 + \mu_i^2\right] = n + \mathbb{E}\|\mu\|_2^2$$

$$\implies \mathbb{E}(\boldsymbol{X}^T\boldsymbol{X} - n) = \mathbb{E}\|\mu\|_2^2$$

Therefore $\boxed{\hat{\delta}_{\text{unbiased}} = \boldsymbol{X}^T\boldsymbol{X} - n}$ is unbiased for $\mathbb{E}\|\mu\|_2^2$ (and given μ it is unbiased for $\|\mu\|_2^2$).

(b) We will begin by finding the posterior distribution of μ. The prior distribution of μ is

$$f(\boldsymbol{\mu}) = (2\pi)^{-n/2}|k\boldsymbol{I}_n|^{-1/2} \cdot \exp\left(-\frac{1}{2}\boldsymbol{\mu}^T(k\boldsymbol{I}_n)^{-1}\boldsymbol{\mu}\right) = \frac{1}{\sqrt{(2\pi k)^n}} \exp\left(-\frac{1}{2k}\boldsymbol{\mu}^T\boldsymbol{\mu}\right).$$

The likelihood is

$$f_{\boldsymbol{X}|\boldsymbol{\mu}}(\boldsymbol{x} \mid \boldsymbol{\mu}) = (2\pi)^{-n/2}|\boldsymbol{I}_n|^{-1/2} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T(\boldsymbol{I}_n)^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)$$

$$= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T(\boldsymbol{x} - \boldsymbol{\mu})\right).$$

So the unconditional distribution of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{\mathbb{R}^n} f_{\boldsymbol{X}|\boldsymbol{\mu}}(\boldsymbol{x} \mid \boldsymbol{\mu}) f(\boldsymbol{\mu}) \; d\boldsymbol{\mu}$$

$$= \int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})\right) \cdot \frac{1}{\sqrt{(2\pi k)^n}} \exp\left(-\frac{1}{2k}\boldsymbol{\mu}^T\boldsymbol{\mu}\right) \; d\boldsymbol{\mu}$$

$$= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\left[\boldsymbol{\mu}^T\boldsymbol{\mu} + \frac{1}{k}\boldsymbol{\mu}^T\boldsymbol{\mu} - 2\boldsymbol{x}^T\boldsymbol{\mu} + \boldsymbol{x}^T\boldsymbol{x}\right]\right) \; d\boldsymbol{\mu}$$

$$= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{k+1}{2k}\left[\boldsymbol{\mu}^T\boldsymbol{\mu} - \frac{2k}{k+1}\boldsymbol{x}^T\boldsymbol{\mu}\right] - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right) \; d\boldsymbol{\mu}$$

$$= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{k+1}{2k}\left[\boldsymbol{\mu}^T\boldsymbol{\mu} - \frac{2k}{k+1}\boldsymbol{x}^T\boldsymbol{\mu} + \left(\frac{k}{k+1}\right)^2 \boldsymbol{x}^T\boldsymbol{x}\right] - \left(-\frac{k+1}{2k}\right)\left(\frac{k}{k+1}\right)^2 \boldsymbol{x}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right) \; d\boldsymbol{\mu}$$

$$= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{k+1}{2k}\left[\left(\boldsymbol{\mu} - \frac{k}{k+1}\boldsymbol{x}\right)^T\left(\boldsymbol{\mu} - \frac{k}{k+1}\boldsymbol{x}\right)\right] - \frac{1}{2}\left[-\left(\frac{k}{k+1}\right)\boldsymbol{x}^T\boldsymbol{x} + \frac{k+1}{k+1}\boldsymbol{x}^T\boldsymbol{x}\right]\right) \; d\boldsymbol{\mu}$$

$$= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{k+1}{2k}\left[\left(\boldsymbol{\mu} - \frac{k}{k+1}\boldsymbol{x}\right)^T\left(\boldsymbol{\mu} - \frac{k}{k+1}\boldsymbol{x}\right)\right]\right) \exp\left(-\frac{1}{2}\frac{1}{k+1}\boldsymbol{x}^T\boldsymbol{x}\right) \; d\boldsymbol{\mu}$$

$$= \frac{1}{\sqrt{(2\pi k)^n}} \exp\left(-\frac{1}{2}\frac{1}{k+1}\boldsymbol{x}^T\boldsymbol{x}\right)\left(\frac{k}{k+1}\right)^{n/2}$$

$$\cdot \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n}} \cdot \left(\frac{k}{k+1}\right)^{-n/2} \exp\left(-\frac{1}{2}\left(\boldsymbol{\mu} - \frac{k}{k+1}\boldsymbol{x}\right)^T\left(\frac{k}{k+1}\boldsymbol{I}_n\right)^{-1}\left(\boldsymbol{\mu} - \frac{k}{k+1}\boldsymbol{x}\right)\right) \; d\boldsymbol{\mu}$$

The second row is the integral over $\mathbb{R}^n$ of an n-dimensional multivariate Gaussian distribution with mean $k/(k+1)\boldsymbol{x}$ and covariance $k/(k+1)\boldsymbol{I}_n$, so it equals 1. Then we are left with

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n}} \frac{1}{k^{n/2}} \exp\left(-\frac{1}{2}\frac{1}{k+1}\boldsymbol{x}^T\boldsymbol{x}\right)\left(\frac{k}{k+1}\right)^{n/2}$$

$$= \frac{1}{\sqrt{(2\pi)^n}} \left[(k+1)^n\right]^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\left((k+1)\boldsymbol{I}_n\right)^{-1}\boldsymbol{x}\right) \tag{1.17}$$

which is the density of an n-dimensional multivariate Gaussian random variable with mean $\boldsymbol{0}$ and covariance $(k+1)\boldsymbol{I}_n$. Therefore the posterior distribution of $\boldsymbol{\mu}$ is

$$f_{\boldsymbol{\mu}|\boldsymbol{X}}(\boldsymbol{\mu} \mid \boldsymbol{x}) = \frac{f_{\boldsymbol{X}|\boldsymbol{\mu}}(\boldsymbol{x} \mid \boldsymbol{\mu}) f(\boldsymbol{\mu})}{f_{\boldsymbol{X}}(\boldsymbol{x})}$$

$$= \left[\frac{1}{(2\pi\sqrt{k})^n} \exp\left(-\frac{1}{2}\left[\boldsymbol{\mu}^T\boldsymbol{\mu} + \frac{1}{k}\boldsymbol{\mu}^T\boldsymbol{\mu} - 2\boldsymbol{x}^T\boldsymbol{\mu} + \boldsymbol{x}^T\boldsymbol{x}\right]\right)\right] \Bigg/ \left[\frac{1}{\sqrt{(2\pi)^n}} \left[(k+1)^n\right]^{-1/2} \exp\left(-\frac{1}{2}\frac{1}{k+1}\boldsymbol{x}^T\boldsymbol{x}\right)\right]$$

$$= \frac{1}{\sqrt{(2\pi)^n}} \left(\frac{k+1}{k}\right)^{n/2} \exp\left(-\frac{k+1}{2k}\left[\left(\boldsymbol{\mu} - \frac{k}{k+1}\boldsymbol{x}\right)^T \left(\boldsymbol{\mu} - \frac{k}{k+1}\boldsymbol{x}\right)\right] - \frac{1}{2}\frac{1}{k+1}\boldsymbol{x}^T\boldsymbol{x} + \frac{1}{2}\frac{1}{k+1}\boldsymbol{x}^T\boldsymbol{x}\right)$$

$$= \frac{1}{\sqrt{(2\pi)^n}} \left(\frac{k+1}{k}\right)^{n/2} \exp\left(-\frac{k+1}{2k}\left(\boldsymbol{\mu} - \frac{k}{k+1}\boldsymbol{x}\right)^T \left(\boldsymbol{\mu} - \frac{k}{k+1}\boldsymbol{x}\right)\right)$$

$$= \frac{1}{\sqrt{(2\pi)^n}} \cdot \left(\frac{k}{k+1}\right)^{-n/2} \exp\left(-\frac{1}{2}\left(\boldsymbol{\mu} - \frac{k}{k+1}\boldsymbol{x}\right)^T \left(\frac{k}{k+1}\boldsymbol{I}_n\right)^{-1} \left(\boldsymbol{\mu} - \frac{k}{k+1}\boldsymbol{x}\right)\right)$$

which is an n-dimensional multivariate Gaussian distribution with mean $k/(k+1)\boldsymbol{x}$ and covariance $k/(k+1)\boldsymbol{I}_n$. That is, conditional on $\boldsymbol{X}$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, where

$$\mu_i \mid \boldsymbol{X} \overset{i.i.d.}{\sim} \mathcal{N}\left(\frac{k}{k+1}X_i, \frac{k}{k+1}\right).$$

$$\iff \left(\mu_i - \frac{k}{k+1}X_i\right)\frac{k+1}{k} \mid \boldsymbol{X} \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \iff \frac{k+1}{k}\mu_i - X_i \mid \boldsymbol{X} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

$$\implies \mathbb{E}\left(\left[\frac{k+1}{k}\mu_i - X_i\right]^2 \mid \boldsymbol{X}\right) = 1 \iff \mathbb{E}\left(\left[\frac{k+1}{k}\right]^2 \mu_i^2 + X_i^2 - 2\frac{k+1}{k}\mu_i X_i \mid \boldsymbol{X}\right) = 1 \tag{1.18}$$

So,

$$\hat{\delta}_{\text{proper}} = \mathbb{E}\left(\|\boldsymbol{\mu}\|_2^2 \mid \boldsymbol{X}\right) = \mathbb{E}\left(\sum_{i=1}^n \mu_i^2 \mid \boldsymbol{X}\right) = \left[\frac{k}{k+1}\right]^2 \mathbb{E}\left(\sum_{i=1}^n \left[\frac{k+1}{k}\right]^2 \mu_i^2 \mid \boldsymbol{X}\right)$$

$$= \left[\frac{k}{k+1}\right]^2 \mathbb{E}\left(\sum_{i=1}^n \left[\frac{k+1}{k}\right]^2 \mu_i^2 + X_i^2 - 2\frac{k+1}{k}\mu_i X_i \mid \boldsymbol{X}\right) - \left[\frac{k}{k+1}\right]^2 \mathbb{E}\left(\sum_{i=1}^n X_i^2 - 2\frac{k+1}{k}\mu_i X_i \mid \boldsymbol{X}\right)$$

$$= \left[\frac{k}{k+1}\right]^2 - \left[\frac{k}{k+1}\right]^2 \sum_{i=1}^n X_i^2 + 2\frac{k+1}{k}\left[\frac{k}{k+1}\right]^2 \sum_{i=1}^n X_i \mathbb{E}\left(\mu_i \mid \boldsymbol{X}\right)$$

$$= \left[\frac{k}{k+1}\right]^2 - \left[\frac{k}{k+1}\right]^2 \sum_{i=1}^n X_i^2 + 2\frac{k}{k+1}\sum_{i=1}^n X_i \cdot \frac{k}{k+1}X_i$$

$$= \left[\frac{k}{k+1}\right]^2 + 2\left[\frac{k}{k+1}\right]^2 \sum_{i=1}^n X_i^2 - \left[\frac{k}{k+1}\right]^2 \sum_{i=1}^n X_i^2 = \left[\frac{k}{k+1}\right]^2 \left(1 + \boldsymbol{X}^T\boldsymbol{X}\right).$$

(c) We will again begin by finding the posterior distribution of μ. The (improper) prior distribution of μ is constant; that is, for some $c \in \mathbb{R}$,

$$f(\boldsymbol{\mu}) = c, \qquad \forall \boldsymbol{\mu} \in \mathbb{R}^n.$$

The likelihood is

$$f_{\boldsymbol{X}|\boldsymbol{\mu}}(\boldsymbol{x} \mid \boldsymbol{\mu}) = (2\pi)^{-n/2}|\boldsymbol{I}_n|^{-1/2} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{I}_n)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

$$= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

So the unconditional distribution of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{\mathbb{R}^n} f_{\boldsymbol{X}|\boldsymbol{\mu}}(\boldsymbol{x} \mid \boldsymbol{\mu}) f(\boldsymbol{\mu})\, d\boldsymbol{\mu} = c\int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})\right)\, d\boldsymbol{\mu}$$

$$= c\int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{x})^T(\boldsymbol{\mu}-\boldsymbol{x})\right)\, d\boldsymbol{\mu}$$

The expression inside the integral is the density of a Gaussian random variable with mean $\boldsymbol{x}$ and covariance $\boldsymbol{I}_n$, so the integral evaluates to 1. Therefore the unconditional distribution of $\boldsymbol{X}$ is also flat. Therefore the posterior distribution of $\boldsymbol{\mu}$ is the same as the likelihood:

$$f_{\boldsymbol{\mu}|\boldsymbol{X}}(\boldsymbol{\mu} \mid \boldsymbol{x}) = \frac{f_{\boldsymbol{X}|\boldsymbol{\mu}}(\boldsymbol{x} \mid \boldsymbol{\mu}) f(\boldsymbol{\mu})}{f_{\boldsymbol{X}}(\boldsymbol{x})} = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{x})^T(\boldsymbol{\mu}-\boldsymbol{x})\right);$$

that is, conditional on $\boldsymbol{X}$, $\boldsymbol{\mu}$ is normally distributed with mean $\boldsymbol{X}$ and covariance $\boldsymbol{I}_n$. So conditional on $\boldsymbol{X}$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, where

$$\mu_i \mid \boldsymbol{X} \overset{i.i.d.}{\sim} \mathcal{N}(X_i, 1).$$

$$\iff \mu_i - X_i \mid \boldsymbol{X} \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \implies \mathbb{E}\left(\left[\mu_i - X_i\right]^2 \mid \boldsymbol{X}\right) = 1 \iff \mathbb{E}\left(\mu_i^2 + X_i^2 - 2\mu_i X_i \mid \boldsymbol{X}\right) = 1 \tag{1.19}$$

So,

$$\hat{\delta}_{\text{flat}} = \mathbb{E}\left(\|\boldsymbol{\mu}\|_2^2 \mid \boldsymbol{X}\right) = \mathbb{E}\left(\sum_{i=1}^n \mu_i^2 \mid \boldsymbol{X}\right) = \mathbb{E}\left(\sum_{i=1}^n \mu_i^2 + X_i^2 - 2\mu_i X_i \mid \boldsymbol{X}\right) - \mathbb{E}\left(\sum_{i=1}^n X_i^2 - 2\mu_i X_i \mid \boldsymbol{X}\right)$$

$$= 1 - \sum_{i=1}^n X_i^2 + 2\sum_{i=1}^n X_i \mathbb{E}\left(\mu_i \mid \boldsymbol{X}\right) = 1 - \sum_{i=1}^n X_i^2 + 2\sum_{i=1}^n X_i^2 = 1 + \boldsymbol{X}^T\boldsymbol{X}.$$

So,

$$\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}} = 1 + \boldsymbol{X}^T\boldsymbol{X} - \left(\boldsymbol{X}^T\boldsymbol{X} - n\right) = 1 + n. \tag{1.20}$$

(d) If the true distribution of $\boldsymbol{\mu}$ is the prior from part (b), then the marginal distribution of $\boldsymbol{X}$ is (1.17):

$$= \frac{1}{\sqrt{(2\pi)^n}} \left[(k+1)^n\right]^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\left((k+1)\boldsymbol{I}_n\right)^{-1}\boldsymbol{x}\right);$$

that is, $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, (k+1)\boldsymbol{I}_n)$. (Note that this means $(k+1)^{-1}X_i$ is standard Gaussian and i.i.d. for all $i \in \{1, \ldots, n\}$.) Per the suggestion, let

$$\hat{D} = \hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}} = \left[\frac{k}{k+1}\right]^2 \left(1 + \boldsymbol{X}^T\boldsymbol{X}\right) - \left(\boldsymbol{X}^T\boldsymbol{X} - n\right) = \frac{k^2 - (k^2+2k+1)}{(k+1)^2}\boldsymbol{X}^T\boldsymbol{X} + \left[\frac{k}{k+1}\right]^2 + n$$

$$= -\frac{2k+1}{(k+1)^2}\boldsymbol{X}^T\boldsymbol{X} + \left[\frac{k}{k+1}\right]^2 + n \tag{1.21}$$

Since from (1.20) we have

$$\mathbb{E}\left(\hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}}\right)^2 = n^2 + 2n + 1,$$

we seek

$$\mathbb{E}\left(\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}}\right)^2 = \mathbb{E}(\hat{D}^2) = \text{Var}(\hat{D}) + \left[\mathbb{E}(\hat{D})\right]^2.$$

Note that since $(k+1)^{-1}X_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ for all $i \in \{1,\dots,n\}$,

$$\sum_{i=1}^{n}\left((k+1)^{-1}X_i\right)^2 \sim \chi^2_n,$$

so

$$\mathbb{E}\left[\sum_{i=1}^{n}\left(\frac{1}{k+1}X_i\right)^2\right] = n \iff \frac{1}{(k+1)^2}\mathbb{E}\left[\sum_{i=1}^{n}X_i^2\right] = n \iff \mathbb{E}\left[\sum_{i=1}^{n}X_i^2\right] = n(k+1)^2,$$

and

$$\text{Var}\left[\sum_{i=1}^{n}\left(\frac{1}{k+1}X_i\right)^2\right] = 2n \iff \frac{1}{(k+1)^4}\text{Var}\left[\sum_{i=1}^{n}X_i^2\right] = 2n \iff \text{Var}\left[\sum_{i=1}^{n}X_i^2\right] = 2n(k+1)^4.$$

Under the unconditional distribution of $\boldsymbol{X}$,

$$\mathbb{E}\left(\hat{D}\right) = \mathbb{E}\left(-\frac{2k+1}{(k+1)^2}\boldsymbol{X}^T\boldsymbol{X} + \left[\frac{k}{k+1}\right]^2 + n\right) = -\frac{2k+1}{(k+1)^2}\mathbb{E}\left(\boldsymbol{X}^T\boldsymbol{X}\right) + \left[\frac{k}{k+1}\right]^2 + n$$

$$= -\frac{2k+1}{(k+1)^2}\mathbb{E}\left(\sum_{i=1}^{n}X_i^2\right) + \left[\frac{k}{k+1}\right]^2 + n = -\frac{2k+1}{(k+1)^2}n(k+1)^2 + \left[\frac{k}{k+1}\right]^2 + n$$

$$= -n\frac{(2k+1)(k^2+2k+1)+k^2}{(k+1)^2} + n = n\left[\frac{k^2+2k+1}{(k+1)^2} - \frac{2k^3+4k^2+2k+k^2+2k+1+k^2}{(k+1)^2}\right]$$

$$= n\left[\frac{-2k^3-5k^2-2k}{(k+1)^2}\right] = -n\left[\frac{2k^3+5k^2+2k}{(k+1)^2}\right]$$

Next,

$$\text{Var}(\hat{D}) = \text{Var}\left(-\frac{2k+1}{(k+1)^2}\boldsymbol{X}^T\boldsymbol{X} + \left[\frac{k}{k+1}\right]^2 + n\right) = \frac{(2k+1)^2}{(k+1)^4}\text{Var}\left(\sum_{i=1}^{n}X_i^2\right)$$

$$= \frac{(2k+1)^2}{(k+1)^4} 2n(k+1)^4 = 2n(2k+1)^2$$

So

$$\mathbb{E}\left(\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}}\right)^2 = 2n(2k+1)^2 + \left(-n\left[\frac{2k^3+5k^2+2k}{(k+1)^2}\right]\right)^2 = n^2\left[\frac{2k^3+5k^2+2k}{(k+1)^2}\right]^2 + n \cdot 2(2k+1)^2$$

$$\approx 4k^2n^2 + 8k^2n,$$

so in general when $k \geq 1$ and n is large,, $\mathbb{E}\left(\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}}\right)^2 > \mathbb{E}\left(\hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}}\right)^2$; that is, the flat prior Bayes estimator is further from the unbiased estimator than the proper prior Bayes estimator.

1.4.5 Method of Moments

Definition 1.4.11 (**Consistency, Definition 6.30 in Math 541A notes**). Let $\{f_\theta : \theta \in \Theta\}$ be a family of distributions. Let $Y_1, Y_2, \ldots$ be a sequence of estimators of $g(\theta)$. We say that $Y_1, Y_2, \ldots$ is **consistent** for $g(\theta)$ if for any $\theta \in \Theta$, $Y_1, Y_2, \ldots$ converges in probability to the constant value $g(\theta)$ with respect to the probability distribution f_θ. That is, $Y_n \xrightarrow{p} g(\theta)$. (Typically we will take Y_n to be a function of a random sample of size n for all $n \geq 1$.)

Example 1.4.10 (**Example 6.31 in Math 541A notes**). Let $X_1, \ldots, X_n$ be a random sample of size n with distribution f_θ. The Weak Law of Large Numbers (Theorem **??**) says that the sample mean is consistent when $\mathbb{E}_\theta|X_1|M\infty$ for all $\theta \in \Theta$. More generally, if $j \geq 1$ is a positive integer such that $\mathbb{E}_\theta|X_1|^j < \infty$ for all $\theta \in \Theta$, then the jth sample moment

$$M_j = M_{j,n}(\theta) := \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

is a consistent estimator for $\mu_j(\theta) := \mathbb{E}X_1^j$.

Definition 1.4.12 (**Method of Moments, Definition 6.32 in Math 541A notes**). Let $g : \Theta \to \mathbb{R}^k$. Suppose we want to estimate $g(\theta)$ for any $\theta \in \Theta$. Suppose there exists $h : \mathbb{R}^j \to \mathbb{R}^k$ such that $g(\theta) = h(\mu_1, \ldots, \mu_j)$. Then the estimator

$$h(M_1, \ldots, M_j)$$

is a **method of moments** estimator for $g(\theta)$, where M_j is the jth sample moment

$$M_j = M_{j,n}(\theta) := \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

Example 1.4.11 (**Example 6.33 in Math 541A notes**)**.** Recall that the standard deviation is

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

To estimate the standard deviation, we can use $\Theta = \mathbb{R} \times (0, \infty) = \{(\mu_1, \mu_2) : \mu_1 \in \mathbb{R}, \mu_2 > 0\}, j = 2$, and $h(\mu_1, \mu_2) = \sqrt{\mu_2 - \mu_1^2}$, so that the method of moments estimator of the standard deviation is $\sqrt{M_2 - M_1^2}$.

Remark 27. The method of moments estimator is not necessarily unbiased.

1.4.6 Maximum likelihood estimator

Definition 1.4.13 (**Maximum Likelihood Estimator (Math 541B Definition)**)**.** $P = \{P_\theta, \theta \in \Theta\}$, $X_1, \ldots, X_n$ i.i.d. $P_{\theta_0}, \theta_0 \in \Theta$. The likelihood function

$$L_n(\theta \mid X_1, \ldots, X_n) := \prod_{j=1}^{n} p_\theta(x_j)$$

$$\ell_n(\theta \mid X_1, \ldots, X_n) = \log\left(L_n(\theta \mid X_1, \ldots, X_n)\right) = \sum_{i=1}^{n} \log p_\theta(X_j)$$

Then the MLE $\hat{\theta}_n$ is defined as

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta}\{L_n(\theta)\} = \arg\max_{\theta \in \Theta}\{\ell_n(\theta)\}.$$

Remark 28. Under some reasonable assumptions, the maximum likelihood estimator is consistent. However, the MLE does not always exist and might not be unique. See Keener for details.

Remark 29 (**Math 541 B remarks**)**.** Equivalently, $\hat{\theta}_n$ maximizes

$$\prod_{j=1}^{n} \frac{p_\theta(X_j)}{p_{\theta_0}(X_j)}$$

$$\iff \hat{\theta}_n = \arg\max_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log\left[\frac{p_\theta(X_j)}{p_{\theta_0}(X_j)}\right] \right\}$$

As $n \to \infty$, for every $\theta \in \Theta$, by the Law of Large Numbers

$$\frac{1}{n} \sum_{i=1}^{n} \log\left[\frac{p_\theta(X_j)}{p_{\theta_0}(X_j)}\right] \xrightarrow{a.s.} \mathbb{E}_{\theta_0} \log\left(\frac{p_\theta(X)}{p_{\theta_0}(X)}\right)$$

$$\iff \hat{\theta}_n \approx \arg\min_{\theta \in \Theta} \mathbb{E}_{\theta_0} \log\left(\frac{p_{\theta_0}(X)}{p_\theta(X)}\right) =: KL(p_{\theta_0} || p_\theta)$$

where $KL(p_{\theta_0} || p_\theta)$ is the **Kullback-Leibler Divergence**. Facts:

1.
$$KL(p_{\theta_0}||p_\theta) = 0 \iff p_{\theta_0} = p_\theta \iff \theta = \theta_0 \text{ (identifiability)}$$

Proof.

$$KL(p_{\theta_0}||p_\theta) = \mathbb{E}_{\theta_0} - \log\left(\frac{p_\theta(X)}{p_{\theta_0}(x)}\right) \geq \text{ (by Jensen's Inequality) } - \log\left(\mathbb{E}_{\theta_0}[] \log\left(\frac{p_\theta(X)}{p_{\theta_0}(x)}\right)\right) = 0$$

for Jensen's Inequality, equality holds if and only if

$$\mathbb{P}\left(\frac{p_\theta(X)}{p_{\theta_0}(x)} = 1\right) = 1$$

□

2. $KL(p||q)$ is not a distance: $KL(p||q) \neq KL(q||p)$, and the KL divergence does not satisfy the Triangle Inequality.

3. **total variational distance (Definition 13.1.2 in Lehmann and Romano [2005]):** (in principal, existence of density is not required, but typically we assume a density exists)

$$TV(P,Q) := \sup_A |P(A) - Q(A)| = \frac{1}{2}\int |p - q| \, d\mu$$

Pinsker's Inequality:

$$TV(P,Q) \leq \sqrt{\frac{KL(P||Q)}{2}}.$$

Hellinger's Distance (Definition 13.1.3 in Lehmann and Romano [2005]): (need to assume existence of density)

$$H^2(P,Q) := \int (\sqrt{p} - \sqrt{q})^2 \, d\mu$$

(ℓ_2 distance between square roots of densities. See also Definition 1.8.4.)

Le Cann's Inequality:

$$TV(P,Q) \geq \frac{1}{2}H^2(P,Q)$$

Using Pinsker's Inequality and Le Cann's Inequality, you can relate KL divergence to Hellinger's Distance.

For more information on KL Divergence, see Sections **??** and **??**).

Proposition 1.4.6.1 (Math 541A Proposition 6.40). For all $i \in 1, \ldots, n$, if $\Theta \to f_\theta(x_i)$ is strictly log-concave, then $\ell(\theta)$ has at most one maximum value.

Remark 30. One example of a function whose log likelihood has no maximum value is $\exp(-e^{-\theta})$ (as in an extreme value distribution).

Remark 31. Intuition of Lemma 6.50 in Math 541A: if distribution follows θ then it is more likely to match distribution function of f_θ than f_ω.

Note on Theorem 6.53: exponential family is an example of a family that satisfies condition (a).

Note from proof:

$$\sqrt{n}\ell_n(\theta) = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} \log f_\theta(x_i) - \mathbb{E}(\log f_\theta(x_i))\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I_{X_1}(\theta)}\right)$$

by the Central Limit Theorem (and Assumption 3 for the variance) since $\mathbb{E}(\log f_\theta(x_i)) = 0$. Also,

$$\ell_n''(\theta') = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}^2}{\mathrm{d}[\theta']^2} \log f_{\theta'}(x_i)$$

which explains the applicability of the Weak Law of Large Numbers.

Proposition 1.4.6.2 (**Stats 100B homework problem**)**.** Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Bernoulli(p) distribution. Let $X = \sum_{i=1}^{n} X_i$. Then

(a) The maximum likelihood estimator of p is $\hat{p} = X/n$.

(b) The maximum likelihood estimator attains the Cramer-Rao lower bound.

(c) The maximum likelihood estimator is a consistent estimator for p.

(d) $\frac{\hat{p}(1-\hat{p})}{n-1}$ is an unbiased estimator for $\mathrm{Var}(\hat{p}) = p(1-p)/n$.

Proof. a. Bernoulli random variable:

$$P(X_i = x) = p^x(1-p)^{1-x}$$

Assuming independent samples,

$$L = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^{n} X_i}(1-p)^{\sum_{i=1}^{n} 1-X_i}$$

$$\log(L) = \sum_{i=1}^{n} X_i \log(p) + \left(\sum_{i=1}^{n} 1 - X_i\right)\log(1-p)$$

$$\frac{\mathrm{d}\log(L)}{\mathrm{d}p} = \frac{1}{p}\sum_{i=1}^{n} X_i - \frac{1}{1-p}\sum_{i=1}^{n}(1-X_i) = 0$$

$$\frac{1}{\hat{p}}\sum_{i=1}^{n} X_i = \frac{1}{1-\hat{p}}\sum_{i=1}^{n}(1-X_i)$$

$$(1-\hat{p})\sum_{i=1}^{n} X_i = \hat{p}\sum_{i=1}^{n}(1-X_i)$$

$$\sum_{i=1}^{n} X_i = \hat{p}\sum_{i=1}^{n}(X_i + 1 - X_i)$$

$$\sum_{i=1}^{n} X_i = n\hat{p}$$

$$\boxed{\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

b.

$$\text{Var}(\hat{p}) = \frac{1}{n^2}\text{Var}\bigg(\sum_{i=1}^{n} X_i\bigg)$$

Since X_i are independent, we can write this as

$$\frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i)$$

And since X_i is Bernoulli, $\text{Var}(X_i) = p(1-p)$.

$$= \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{1}{n^2}np(1-p) = \boxed{\frac{p(1-p)}{n}}$$

Cramer-Rao lower bound:

$$\text{Var}(\hat{\theta}) \geq 1/\bigg(-n\mathbb{E}\bigg[\frac{\partial^2 \log(f(X;\theta))}{\partial\theta^2}\bigg]\bigg)$$

$$\frac{\partial}{\partial p}\log(p^x(1-p)^{1-x}) = \frac{\partial}{\partial p}(x\log(p) + (1-x)\log(1-p)) = \frac{x}{p} - \frac{1-x}{1-p}$$

$$\frac{\partial^2 \log(f(X;\theta))}{\partial\theta^2} = \frac{\partial}{\partial p}(\frac{x}{p} - \frac{1-x}{1-p}) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

$$\mathbb{E}\bigg[\frac{\partial^2 \log(f(X;\theta^2))}{\partial\theta}\bigg] = \mathbb{E}(-\frac{x}{p^2} - \frac{1-x}{(1-p)^2}) = -\frac{1}{p^2}\mathbb{E}(x) - \frac{1}{(1-p)^2}\mathbb{E}(1-x) = -\frac{1}{p^2}p - \frac{1}{(1-p)^2}(1-p)$$

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

$$\implies \text{Var}(\hat{p}) \geq 1/\bigg(-n(\frac{-1}{p(1-p)})\bigg) = \frac{p(1-p)}{n} = \text{Var}(\hat{p})$$

c. (1) **Unbiased:**

$$E\bigg(\frac{X}{n}\bigg) = \frac{np}{p} = p$$

(2) $\text{Var}(\hat{\theta}) \to 0$ **as** $n \to \infty$ **:**

$$\lim_{n\to\infty}\text{Var}\bigg(\frac{X}{n}\bigg) = \lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \lim_{n\to\infty}\frac{1}{n^2}\cdot np(1-p) = \lim_{n\to\infty}\frac{p(1-p)}{n} = \boxed{0}$$

Therefore $\frac{X}{n}$ is a consistent estimator of p.

d.

$$\mathbb{E}(\hat{\sigma^2}) = \mathbb{E}\left[\frac{1}{n}\left(\frac{X}{n}\big(1-\frac{X}{n}\big)\right)\right] = \mathbb{E}\left[\frac{X(n-X)}{n^3}\right] = \frac{1}{n^3}\mathbb{E}[nX-(X)^2] = \frac{1}{n^2}\mathbb{E}(X) - \frac{1}{n^3}\mathbb{E}(X^2)$$

$$= \frac{1}{n^2}\cdot np - \frac{1}{n^3}\big(\text{Var}(X) + (E(X))^2\big) = \frac{p}{n} - \frac{np(1-p)}{n^3} - \frac{p^2n^2}{n^3} = \frac{pn-p+p^2-p^2n}{n^2}$$

$$= \frac{p(n-1+p-pn)}{n^2} = \frac{p(n-1)(1-p)}{n^2}$$

This is a biased estimator since $\text{Var}(X) = \frac{p(1-p)}{n}$ (since X is bionomial).

$$c\cdot\frac{p(n-1)(1-p)}{n^2} = \frac{p(1-p)}{n} \implies \boxed{c = \frac{n}{n-1}}$$

□

Proposition 1.4.6.3 (**Stats 100B homework problem**)**.** Suppose that X follows a geometric distribution and we take an i.i.d. sample of size n. Then the maximum likelihood estimator of p is

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

Proof. Since sample is i.i.d.:

$$L = \prod_{i=1}^n p(1-p)^{X_i-1} = p^n(1-p)^{-n+\sum_{i=1}^n X_i}$$

$$\log(L) = n\log(p) + \left(-n+\sum_{i=1}^n X_i\right)\log(1-p)$$

$$\frac{\mathrm{d}\log(L)}{\mathrm{d}p} = \frac{n}{p} - \frac{1}{1-p}\left(-n+\sum_{i=1}^n X_i\right) = 0$$

$$\frac{n}{\hat{p}} = \frac{1}{1-\hat{p}}\left(-n+\sum_{i=1}^n X_i\right)$$

$$(1-\hat{p})\hat{p} = -n\hat{p} + \hat{p}\sum_{i=1}^n X_i$$

$$n = \hat{p}\sum_{i=1}^n X_i$$

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

□

Proposition 1.4.6.4 (**Stats 100B homework problem**). Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Poisson(λ) distribution. Then

(a) The maximum likelihood estimator of λ is

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

(b) The variance of the maximum likelihood estimator is

$$\text{Var}(\hat{\lambda}) = \frac{\lambda}{n}$$

(c) The maximum likelihood estimator is a minimum variance unbiased estimator.

(d) The maximum likelihood estimator is consistent.

Proof. (a)

$$f(X_i; \lambda) = \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

Assuming the samples are independent,

$$L = \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} = \left(e^{-n\lambda}\lambda^{\sum_{i=1}^{n} X_i}\right) / \prod_{i=1}^{n} X_i!$$

$$\log(L) = -n\lambda + \left(\sum_{i=1}^{n} X_i\right)\log(\lambda) - \sum_{i=1}^{n}\log(X_i!)$$

$$\frac{\mathrm{d}\log(L)}{\mathrm{d}\lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i = 0$$

$$\implies \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{x}$$

(b)

$$\text{Var}(\hat{\lambda}) = \text{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\text{Var}(\sum_{i=1}^{n} X_i)$$

Since X_i are i.i.d. this can be written as

$$= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\lambda = \frac{\lambda}{n}$$

(c) Cramer-Rao lower bound:

$$\text{Var}(\hat{\lambda}) \geq 1 / \left(-n\mathbb{E}\left[\frac{\partial^2 \log(f(X;\lambda))}{\partial\lambda^2}\right]\right)$$

$$\log(f(X;\lambda)) = \log\left(\frac{\lambda^{X_i} e^{-\lambda}}{X_i!}\right) = X_i\log(\lambda) - \lambda - \log(X_i!)$$

$$\frac{\partial}{\partial\lambda}\log(f(X;\lambda)) = \frac{1}{\lambda}X_i - 1$$

$$\frac{\partial^2\log(f(X;\lambda))}{\partial\lambda^2} = -\frac{1}{\lambda^2}X_i$$

$$\mathbb{E}\left[\frac{\partial^2\log(f(X;\lambda^2))}{\partial\lambda}\right] = -\frac{1}{\lambda^2}\mathbb{E}(X_i) = -\frac{1}{\lambda^2}\lambda = -\frac{1}{\lambda}$$

$$\implies \text{Var}(\hat{\lambda}) \geq 1/\left(-n\mathbb{E}\left[\frac{\partial^2\log(f(X;\lambda))}{\partial\lambda^2}\right]\right) = \frac{1}{n/\lambda} = \boxed{\frac{\lambda}{n} = \text{Var}(\hat{\lambda})}$$

Since $\text{Var}(\hat{\lambda})$ equals the Cramer-Rao lower bound, $\hat{\lambda}$ is a MVUE.

(d) We already know the MLE is unbiased. To show consistency, we show $\text{Var}(\hat{\theta}) \to 0$ as $n \to \infty$.

$$\lim_{n\to\infty}\text{Var}(\hat{\lambda}) = \lim_{n\to\infty}\frac{\lambda}{n} = \boxed{0}$$

Therefore $\hat{\lambda}$ is a consistent estimator of λ.

□

Proposition 1.4.6.5 (**Stats 100B homework problem, similar to Math 541A example 6.47**). Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Exponential(λ) distribution. Then the maximum likelihood estimator of λ is

$$\hat{\lambda} = n/\sum_{i=1}^{n} X_i = \frac{1}{\bar{X}}.$$

Proof.

$$f(X_i;\lambda) = \lambda e^{-\lambda X_i}$$

Assuming the samples are independent,

$$L = \prod_{i=1}^{n}\lambda e^{-\lambda X_i} = \lambda^n\exp(-\lambda\sum_{i=1}^{n}X_i)$$

$$\iff \log(L) = n\log(\lambda) + -\lambda\sum_{i=1}^{n}X_i$$

$$\iff \frac{\mathrm{d}\log(L)}{\mathrm{d}\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n}X_i = 0$$

$$\implies \hat{\lambda} = n/\sum_{i=1}^{n}X_i = \frac{1}{\bar{X}}$$

□

Proposition 1.4.6.6 (**Stats 100B homework problem; similar to Math 541A Example 6.45**). Let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample from a normal population with mean zero and unknown variance σ^2. Then

(a) The maximum likelihood estimator of σ^2 is

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

(b) The maximum likelihood estimator of σ^2 is biased, but asymptotically unbiased.

(c) The maximum likelihood estimator of σ^2 has variance

$$\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) = \frac{2\sigma^4}{n}$$

and is consistent.

(d) The variance of the maximum likelihood estimator of σ^2 reaches the Cramer-Rao lower bound.

(e) The maximum likelihood estimator of μ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and it is unbiased and UMVU/MVUE.

Proof. a. Since sample is i.i.d. $\mathcal{N}(0, \sigma^2)$:

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{X_i - \mu}{\sigma}\right]^2\right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right)$$

$$\log(L) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - (\sigma^2)^{-1}\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$\frac{\partial \log(L)}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + (\sigma^2)^{-2}\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2 = 0$$

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{(\hat{\sigma^2})^2} = \frac{n}{\hat{\sigma^2}}$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

b. **something wrong with this part**

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i^2\right)$$

Since the sample is i.i.d., this can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2)$$

Since $X_i \sim \mathcal{N}(0,\sigma^2)$, $X_i^2/\sigma^2 \sim \chi_1^2$. So we have

$$\mathbb{E}\left(\frac{X_i^2}{\sigma^2}\right) = 1$$

$$\frac{1}{\sigma^2}\mathbb{E}(X_i^2) = 1$$

$$\mathbb{E}(X_i^2) = \sigma^2$$

Therefore

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right) = \frac{1}{n}\sum_{i=1}^{n}\ \sigma^2 = \frac{1}{n}n\sigma^2 = \boxed{\sigma^2}$$

However it is asymptotically biased: that is,

$$\lim_{n\to\infty}\frac{\mathbb{E}(\hat{\sigma}^2)}{\sigma^2} = 1$$

c.

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n}X_i^2\right)$$

Since X_i is i.i.d. this can be written as

$$\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i^2)$$

Again, , $X_i^2/\sigma^2 \sim \chi_1^2$, so we have

$$\mathrm{Var}\left(\frac{X_i^2}{\sigma^2}\right) = 2$$

$$\frac{1}{\sigma^4}\mathrm{Var}(X_i^2) = 2$$

$$\mathrm{Var}(X_i^2) = 2\sigma^4$$

Therefore

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right) = \frac{1}{n^2}\sum_{i=1}^{n}2\sigma^4 = \frac{2n\sigma^4}{n^2} = \boxed{\frac{2\sigma^4}{n}}$$

Test for consistency (already known that estimate is unbiased):

$$\lim_{n\to\infty}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right) = \lim_{n\to\infty}\frac{2\sigma^4}{n} = \boxed{0}$$

So this is a consistent estimator of σ^2.

d. Cramer-Rao lower bound:

$$\text{Var}(\hat{\theta}) \geq 1/\bigg(- n\mathbb{E}\bigg[\frac{\partial^2 \log(f(X;\theta))}{\partial\theta^2}\bigg]\bigg)$$

$$\log(f(X;\theta)) = \log\bigg[\frac{1}{\sigma\sqrt{2\pi}}\exp\bigg(- \frac{1}{2}\bigg[\frac{X_i}{\sigma}\bigg]^2\bigg)\bigg] = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}X_i^2(\sigma^2)^{-1}$$

$$\frac{\partial}{\partial\sigma^2}\bigg(- \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}X_i^2(\sigma^2)^{-1}\bigg) = -\frac{1}{2\sigma^2} + \frac{1}{2}(X_i)^2(\sigma^2)^{-2}$$

$$\frac{\partial^2 \log(f(X;\theta))}{\partial(\sigma^2)^2} = \frac{1}{2}(\sigma^2)^{-2} - X_i^2(\sigma^2)^{-3}$$

$$\mathbb{E}\bigg[\frac{\partial^2 \log(f(X;\theta^2))}{\partial\theta}\bigg] = \mathbb{E}\bigg[\frac{1}{2}(\sigma^2)^{-2} - X_i^2(\sigma^2)^{-3}\bigg] = \frac{1}{2\theta^4} - \frac{1}{\theta^6}\mathbb{E}(X_i^2) = \frac{1}{2\theta^4} - \frac{\theta^2}{\theta^6} = -\frac{1}{2\theta^4}$$

$$\implies \text{Var}(\hat{\sigma^2}) \geq 1/\bigg(- n\mathbb{E}\bigg[\frac{\partial^2 \log(f(X;\theta))}{\partial\theta^2}\bigg]\bigg) = \frac{1}{n/(2\theta^4)} = \boxed{\frac{2\theta^4}{n}}$$

Therefore the variance of this estimator is equal to the Cramer-Rao lower bound.

Alternative solution (Math 541A):

$$I_X(\sigma) = I_{(X_1,\ldots,X_n)}(\sigma) = \text{(by Proposition 1.4.3.1 (Proposition 6.21 in Math 541A notes))}\ nI_{X_1}(\sigma)$$

$$= n\text{Var}_\sigma\bigg(\frac{\mathrm{d}}{\mathrm{d}\sigma}\log\bigg(\frac{1}{\sigma\sqrt{2\pi}}\exp\bigg(- \frac{(X_1-\mu)^2}{2\sigma^2}\bigg)\ \text{(by definition of Fisher information)}$$

$$= n\text{Var}_\sigma\bigg[- \frac{1}{\sigma} - \frac{\mathrm{d}}{\mathrm{d}\sigma}\bigg(\frac{(-X_1-\mu)^2}{2\sigma^2}\bigg)\bigg] = n\sigma^{-6}\text{Var}_\sigma\big((X_1-\mu)^2\big)$$

$$= n\sigma^{-6}\big(\mathbb{E}(X_1-\mu)^4 - \big[\mathbb{E}(X_1-\mu)^2\big]^2\big) = n\sigma^{-6}\sigma^4(3-1) = \frac{2n}{\sigma^2}.$$

By Cramer-Rao,

$$\text{Var}_\sigma(z) \geq \frac{|g'(0)|^2}{I_X(0)}$$

Note that $g(\sigma) = \mathbb{E}Y = \frac{n-1}{n}\sigma^2$ and that $\mathbb{E}\sum_{j=1}^n(X_j - \overline{X})^2 = \sigma^2(n-1)$. If Z is unbiased for σ^2,

$$g'(\sigma) = \frac{2\sigma(n-1)}{n}|g'(\sigma)|^2 = \frac{4\sigma^2(n-1)^2}{n^2}$$

So,

$$\text{Var}_\sigma(z) \geq \frac{4\sigma^2(n-1)^2}{n^2 2n\sigma^{-2}} = \frac{2(n-1)^2\sigma^4}{n^3}.$$

Note that

$$fracn-1\sigma^2\sum_{j=1}^{n}(X_j-\overline{X})^2\sim\chi^2_{n-1}$$

$$\vdots$$

Note that

$$\frac{(n-1)S_n^2}{\sigma^2}\sim\chi^2_n\implies\text{Var}(\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2)=(n-1)\cdot 2$$

$$\implies\text{Var}(\hat{\sigma}^2)=\text{Var}_\sigma(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2)=\frac{1}{n^2}\sigma^4\text{Var}_\sigma(\sigma^{-2}\sum_{i=1}^{n}(X_i-\overline{X})^2)=\frac{2\sigma^4(n-1)}{n^2}$$

e.

$$\log(L)=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma^2)-(\sigma^2)^{-1}\frac{1}{2}\sum_{i=1}^{n}X_i^2$$

$$\frac{\partial\log(L)}{\partial\mu}=\sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2}=0\iff n\mu=\sum_{i=1}^{n}x_i\iff\mu=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\implies\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

Also note that

$$\mathbb{E}(\hat{\mu})=\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{1}{n}\cdot n\mu=\mu.$$

□

Example 1.4.12 (**Example 6.46, Math 541A notes**). Alternative solution at https://math.stackexchange.com/questions/49543/maximum-estimator-method-more-known-as-mle-of-a-uniform-distribution

Proposition 1.4.6.7 (**Functional Equivariance of the MLE; Proposition 6.49 from Lecture Notes, Theorem 7.2.10 in Casella and Berger [2001],**). Let $g:\Theta\to\Theta'$ be a bijection. Suppose Y is the MLE of θ. Then $g(Y)$ is the MLE of $g(\theta)$.

Proof (case when g is invertible). Note that if $\ell(\theta)$ is the likelihood function for θ, then the likelihood function for $g(\theta)$ can be expressed as

$$\ell(g(\theta))=\prod_{i=1}^{n}f_\theta(x_i\mid g^{-1}(g(\theta)))=\ell(g^{-1}(g(\theta)))=\ell(g^{-1}(\theta'))$$

where $\theta'=g(\theta)$. By definition of the MLE, $Y=t(X_1,\dots,X_n)$ achieves the maximum value of $\theta\mapsto\ell(\theta)$. Therefore we can equivalently say $g(Y)=g(t(X_1,\dots,X_n))$ achieves the maximum value of $\theta'\mapsto\ell(g^{-1}(\theta'))$.

(For a proof when g is not invertible, see Theorem 7.2.10 in Casella and Berger [2001][p.320].)

□

Proposition 1.4.6.8 (Math 541A Homework Problem). Let $X_1, \dots, X_n$ be a random sample of size n, so that X_1 has the Laplace density $\frac{1}{2}e^{-|x-\theta|}$ for all $x \in \mathbb{R}$, where $\theta \in \mathbb{R}$ is unknown. Then the MLE of θ is the mdiean.

Proof.

$$f(x_i;\theta) = \frac{1}{2}e^{-|x_i-\theta|}$$

$$\implies L = \prod_{i=1}^{n} \frac{1}{2}e^{-|x_i-\theta|} = 2^{-n}\exp\left(-\sum_{i=1}^{n}|x_i-\theta|\right)$$

$$\implies \log(L) = -n\log(2) - \sum_{i=1}^{n}|x_i-\theta| \implies \frac{\mathrm{d}\log(L)}{\mathrm{d}\theta} = -\sum_{i=1}^{n}\frac{\mathrm{d}}{\mathrm{d}\theta}|x_i-\theta| = -\sum_{i=1}^{n}\mathrm{sgn}(x_i-\theta)$$

since $\frac{\mathrm{d}|x|}{\mathrm{d}x} = \mathrm{sgn}(x)$. Next set this equal to 0 and solve:

$$-\sum_{i=1}^{n}\mathrm{sgn}(x_i-\hat{\theta}_{MLE}) = 0$$

Notice that if n is even then the median set as $\hat{\theta}_{MLE}$ satisfies the above equation. If n is odd, the median is still the best we can do. So the MLE is the median.

□

1.4.7 Bayes estimator

1.4.8 EM Algorithm

Remark 32 (Correction to Remark 6.57). If Y is constant, the algorithm just outputs θ_0 in one step by the Likelihood Inequality (Lemma 6.50 in lecture notes):

$$\mathbb{E}_\theta \log\left(\frac{f_\theta(X)}{f_\omega(X)}\right) \geq 0 \iff \mathbb{E}_\theta \log f_\theta(X) - \mathbb{E}_\theta \log f_\omega(X) \geq 0$$

has equality only when $\omega = \theta$ (if $\mathbb{P}_\theta \neq \mathbb{P}_\omega \ \forall \theta \neq \omega$). So

$$\mathbb{E}_\theta \log f_\theta(X) \geq \mathbb{E}_\theta \log f_\omega(X).$$

Remark 33 (note on proof of Lemma 6.58).

$$Y = t(X), \qquad f_{X,Y}(x,.y) = f_X(x)\mathbf{1}_{y=t(x)}.$$

1.4.9 Comparison of estimators

1.5 Resampling and Bias Reduction

1.5.1 Jackknife Resampling

$$Z_n := Y_n + (n-1)\left(Y_n - \frac{1}{n}\sum_{i=1}^{n} t_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)\right)$$

1.5.2 Bootstrapping

Suppose $X_1, \ldots, X_n$ i.i.d. from $\mathbb{P}_\theta$. $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is an estimator for θ. Question: what is the distribution of $\hat{\theta} - \theta$. or

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}} \tag{1.22}$$

(where $\hat{\sigma}$ is an estimator of standard deviation)? Another scenario: T_n is a test statistic. What is the distribution of T_n under the null hypothesis?

Definition 1.5.1 (**Parametric bootstrap**)**.** Let $\hat{X}_1, \ldots, \hat{X}_n$ be an i.i.d. sample from $\mathbb{P}_{\hat{\theta}}$. Consider

$$\hat{T}(\hat{X}_1, \ldots, \hat{X}_n, \hat{\theta}) = \frac{\hat{\theta}(\hat{X}_1, \ldots, \hat{X}_n) - \hat{\theta}(X_1, \ldots, X_n)}{\hat{\sigma}(\hat{X}_1, \ldots, \hat{X}_n}. \tag{1.23}$$

(Unlike in (1.22), note that all of the quantities in (1.23) are known—data dependent, can compute.)

Example 1.5.1. $\mathcal{N}(\theta, \sigma^2)$. $\hat{\theta}(X_1, \ldots, X_n) = n^{-1}\sum_{i=1}^n X_i$, $\hat{\sigma}^2(X_1, \ldots, X_n) = (n-1)^{-1}\sum_{i=1}^n (X_i - \overline{X}_i)^2$.

Question: when are the distributions of T and $\hat{T}$ "close"? Intuition: if $T(\theta, X_1, \ldots, X_n)$ is continuous with respect to all variables, then bootstrap should work.

Definition 1.5.2 (**Non-parametric bootstrap**)**.** Estimate distribution of $\mathbb{P}$ via empirical distribution:

$$\hat{\mathbb{P}}_n := \hat{\mathbb{P}}_n(\mathcal{A}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in \mathcal{A}\}$$

In other words, $\hat{\mathbb{P}}_n$ is a uniform distribution on $\{X_1, \ldots, X_n\}$. Samping from $\hat{\mathbb{P}}_n$ is easy—corresponds to a multinomial distribution.

Treats the distribution itself as a parameter. Used when nothing is known about the distribution of the data, or when sampling from $\mathbb{P}_\theta$ is difficult. Note that $\mathbb{E}_{\mathbb{P}}\hat{\mathbb{P}}_n(\mathcal{A}) = \mathbb{P}(\mathcal{A})$.

Example 1.5.2. $X_1, \ldots, X_n$ i.i.d. $\mathbb{P}$, $Y_1, \ldots, Y_m$ i.i.d. $\mathbb{Q}$. Let μ_P, μ_Q be the means of $\mathbb{P}$ and $\mathbb{Q}$ respectively. $H_0 : \mu_P = \mu_Q$, $H_a : \mu_P \neq \mu_Q$. Consider the test statistic

$$T = \frac{\overline{X} - \overline{Y}}{S_{n+m}}$$

where

$$S_{n+m} := \frac{n-1}{m+n-2} S_X^2 + \frac{m-1}{m+n-2} S_Y^2$$

is the **pooled variance**. What is the distribution of T under H_0? Let

$$\overline{Z} := \frac{1}{n+m} \left(\sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i \right),$$

and let

$$\tilde{X}_i = X_i - \overline{X}_n + \overline{Z}, \qquad i \in [n], \qquad \tilde{Y}_i = Y_i - \overline{Y}_m + \overline{Z}, \qquad i \in [m].$$

Then $\tilde{X}_1, \dots, \tilde{X}_n$ and $\tilde{Y}_1, \dots, \tilde{Y}_m$ can be used to model the distribution under H_0. For $i \in [B]$, repeat the following:

1. Sample $\hat{X}_1^{(i)}, \dots, \hat{X}_n^{(i)}$ from $\hat{\mathbb{P}}_n(\tilde{X}_1, \dots, \tilde{X}_n)$.
2. Sample $\hat{Y}_1^{(i)}, \dots, \hat{Y}_n^{(i)}$ from $\hat{\mathbb{P}}_m(\tilde{Y}_1, \dots, \tilde{Y}_n)$.
3. Compute $\hat{T}^{(i)} = T(\hat{X}_1^{(i)}, \dots \hat{X}_n^{(i)}, \hat{Y}_1^{(i)}, \dots, \hat{Y}_n^{(i)})$.

For each C, compute

$$\hat{\mathbb{P}}(\hat{T} \geq C) = \frac{1}{B} \sum_{i=1}^{B} \mathbb{1} \left\{ \hat{T}^{(i)} \geq C \right\}.$$

To obtain a test of size α, select C such that

$$\hat{\mathbb{P}}(\hat{T} \geq C) = \alpha.$$

Example 1.5.3 (**(Non-parametric) Bootstrap failure**). Let $X_1, \dots, X_n$ be i.i.d. Uniform$[0, \theta]$. Let $T_n = n(\theta - \hat{\theta}_n)$, where $\hat{\theta}_n = \max\{X_1, \dots, X_n\} = X_{(n)}$, the maximum likelihood estimator of θ. Let

$$\hat{T}_n = n(\hat{\theta}_n(X_1, \dots, X_n) - \hat{\theta}_n(\hat{X}_1, \dots, \hat{X}_n)) = n(\max(X_1, \dots, X_n) - \max(\hat{X}_1, \dots, \hat{X}_n)) = n(X_{(n)} - \hat{X}_{(n)}).$$

Are the distributions of T_n and $\hat{T}_n$ close? Turns out no. Note that $\mathbb{P}(\hat{X}_1 = X_1) = n^{-1}$ for $i \in [n]$, and note that $\hat{X}_{(n)} \leq X_{(n)}$, so $\hat{T}_n \geq 0$. Fix $t \geq 0$.

$$\begin{aligned}\mathbb{P}(\hat{T}_n \le t \mid X_1,\dots,X_n) &\ge \mathbb{P}(\hat{T}_n = 0 \mid X_1,\dots,X_n)\\ &= \mathbb{P}(\text{at least one among } \hat{X}_1,\dots,\hat{X}_n \text{ is equal to } X_{(n)} \mid X_1,\dots,X_n)\\ &= 1 - \mathbb{P}(\text{none of } \hat{X}_1,\dots,\hat{X}_n \text{ are equal to } X_{(n)} \mid X_1,\dots,X_n) = 1 - \left(\frac{n-1}{n}\right)^n \xrightarrow{n\to\infty} 1 - \frac{1}{e}. \quad (1.24)\end{aligned}$$

At the same time,

$$\begin{aligned}\mathbb{P}_\theta(T_n \le t) &= \mathbb{P}\left(n(\theta - X_{(n)}) \le t\right) = \mathbb{P}\left(X_{(n)} \ge \theta - \frac{t}{n}\right) = 1 - \mathbb{P}\left(X_{(n)} \le \theta - \frac{t}{n}\right)\\ &= 1 - \mathbb{P}\left(\bigcap_{i=1}^n \left\{X_i \le \theta - \frac{t}{n}\right\}\right) = 1 - \left[\mathbb{P}\left(X_1 \le \theta - \frac{t}{n}\right)\right]^n = 1 - \left[1 + \frac{-t/\theta}{n}\right]^n \quad (1.25)\end{aligned}$$

$$\implies \lim_{n\to\infty} \mathbb{P}_\theta(T_n \le t) = \lim_{n\to\infty} \mathbb{P}\left(n(\theta - X_{(n)}) \le t\right) = 1 - \lim_{n\to\infty}\left[1 + \frac{-t/\theta}{n}\right]^n = 1 - \exp\left(-\frac{t}{\theta}\right).$$

Since the lower bound in (1.24) does not depend on t, bootstrap fails. Exercise: does parametric bootstrap work here?

Theorem 1.5.2.1. Assume that $\mathbb{E}X_1 = \mu$, $\mathrm{Var}(X_1) = \sigma^2$, $\mathbb{E}|X_1 - \mu|^3 < \infty$. Let $\hat{\mu}_n = \overline{X}_n = n^{-1}\sum_{i=1}^n X_i$, $F_n(t) = \mathbb{P}(\sqrt{n}(\hat{\mu}_n - \mu) \le t)$. Let $\hat{X}_1,\dots,\hat{X}_n$ be i.i.d. from $\hat{\mathbb{P}}_n$, and let

$$\hat{\mu}^* = \frac{1}{n}\sum_{i=1}^n \hat{X}_i, \qquad \hat{F}_n(t) = \mathbb{P}\left(\sqrt{n}(\hat{\mu}_n^* - \hat{\mu}_n) \le t(X-1,\dots,X_n)\right)$$

Then

$$\sup_{t\in\mathbb{R}} \left|\hat{F}_n(t) - F_n(t)\right| = o_{\mathbb{P}}(1) \text{ as } n \to \infty.$$

Proof. Let $\phi_\sigma(t)$ be the CDF of $\mathcal{N}(0,\sigma^2)$. We have

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\mu}_n)^2 = \mathrm{Var}(\hat{X}_n), \qquad \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_j = \mathbb{E}_{\hat{P}_n}\hat{X}_1.$$

Note that

$$\begin{aligned}\sup_{t\in\mathbb{R}} \left|F_n(t) - \hat{F}_n(t)\right| &= \sup_{t\in\mathbb{R}} \left|F_n(t) - \phi_\sigma(t) + \phi_\sigma(t) - \phi_{\hat{\sigma}}(t) + \phi_{\hat{\sigma}}(t) - \hat{F}_n(t)\right|\\ &\le \sup_{t\in\mathbb{R}} \underbrace{|F_n(t) - \phi_\sigma(t)|}_{\text{error in normal approximation; close by Berry-Esseen}} + \sup_{t\in\mathbb{R}} \underbrace{|\phi_\sigma(t) - \phi_{\hat{\sigma}}(t)|}_{\text{close since } \hat{\sigma}^2 \text{ is consistent}} + \sup_{t\in\mathbb{R}} \underbrace{\left|\phi_{\hat{\sigma}}(t) - \hat{F}_n(t)\right|}_{\text{close by Berry-Esseen}}\end{aligned}$$

It remains to estimate all three terms. By Berry-Esseen,

$$\sup_{t\in\mathbb{R}} |F_n(t) - \phi_\sigma(t)| \leq \frac{3\mathbb{E}|X_1 - \mu|^3}{\sigma^3\sqrt{n}}.$$

By applying Berry-Esseen conditionally on $X_1, \ldots, X_n$,

$$\sup_{t\in\mathbb{R}} \left|\hat{F}_n(t) - \phi_{\hat{\sigma}}(t)\right| \leq 3\frac{\hat{\gamma}}{\hat{\sigma}^3\sqrt{n}}$$

where

$$\hat{\gamma} := \frac{1}{n}\sum_{i=1}^{n} |X_i - \overline{X}_n|^3 = \mathbb{E}_{\hat{\mathbb{P}}_n}\left|X - \mathbb{E}_{\hat{\mathbb{P}}_n} X\right|^3.$$

By the Strong Law of Large Numbers, $\overline{X}_n \xrightarrow{a.s.} \mu, \hat{\gamma} \xrightarrow{a.s.} \mathbb{E}|X_1\mu|^3, \hat{\sigma}^2 \xrightarrow{a.s.} \sigma^2$. Hence

$$\frac{3\hat{\gamma}}{(\hat{\sigma}^2)^{3/2}\sqrt{n}} = o_{\mathbb{P}}(1).$$

Finally, let $\sigma_1, \sigma_2 > 0$ be fixed. Then

$$\sup_{t\in\mathbb{R}} |\phi_{\sigma_1}(t) - \phi_{\sigma_2}(t)| = \sup_t \left| \int_{-\infty}^{t} \left(\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{y^2}{2\sigma_1^2}\right\} - \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{y^2}{2\sigma_2^2}\right\} \right) dy \right|$$

Note that

$$\begin{aligned} &\frac{1}{\sigma_1}\exp\left\{-\frac{y^2}{2\sigma_1^2}\right\} + \frac{1}{\sigma_1}\exp\left\{-\frac{y^2}{2\sigma_2^2}\right\} - \frac{1}{\sigma_1}\exp\left\{-\frac{y^2}{2\sigma_2^2}\right\} - \frac{1}{\sigma_2}\exp\left\{-\frac{y^2}{2\sigma_2^2}\right\} \\ &\qquad = \underbrace{\frac{1}{\sigma_1}\left(\exp\left\{-\frac{y^2}{2\sigma_1^2}\right\} - \exp\left\{-\frac{y^2}{2\sigma_2^2}\right\}\right)}_{A_1} + \underbrace{\exp\left\{-\frac{y^2}{2\sigma_2^2}\right\}\left(\frac{1}{\sigma_1} - \frac{1}{\sigma_2}\right)}_{A_2} \end{aligned}$$

Then

$$A_1 = \frac{1}{\sigma_1}\exp\left\{-\frac{y^2}{2\sigma_1^2}\right\}\left(1 - \exp\left\{-\frac{y^2}{2}\underbrace{\left[\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right]}_{>0}\right\}\right)$$

For $x > 0$, $1 - e^{-x} \leq x$. Hence

$$A_1 \leq \frac{1}{\sigma_1}\exp\left\{-\frac{y^2}{2\sigma_1^2}\right\} \cdot \frac{y^2}{2}\left[\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right] = \frac{1}{\sigma_1}\frac{y^2}{2}\exp\left\{-\frac{y^2}{2\sigma_1^2}\right\}\frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2\sigma_2^2}.$$

On the other hand,

$$A_2 = \exp\left\{-\frac{y^2}{2\sigma_2^2}\right\}\left(\frac{1}{\sigma_1} - \frac{1}{\sigma_2}\right) = \exp\left\{-\frac{y^2}{2\sigma_2^2}\right\}\left(\frac{\sigma_2 - \sigma_1}{\sigma_1}\right)$$

so

$$\sup_{t\in\mathbb{R}} |\phi_\sigma(t) - \phi_{\sigma_2}(t)| \leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{y^2}{2\sigma_1} \exp\left\{-\frac{y^2}{2\sigma_1^2}\right\} dy \cdot \left|\frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2\sigma_2^2}\right| + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sigma_2} \exp\left\{-\frac{y^2}{2\sigma_2^2}\right\} dy \frac{|\sigma_2 - \sigma_1|}{\sigma_1}$$

Hence $\sup_{t\in\mathbb{R}} |\phi_\sigma(t) - \phi_{\sigma_2}(t)| \to 0$ as $\sigma_1 \to \sigma_2$. Apply this with $\sigma_1^2 = \sigma_2$, $\sigma_2^2 = \hat{\sigma}^2$ to yield the result.

□

Bootstrap and Bias Corrections. Suppose that X is a random variable with $\mathbb{E}X = \mu$, $\text{Var}(X) = \sigma^2$, $\mathbb{E}|X|^3 < \infty$. Assume that we are interested in estimating $g(\mu)$ where g is a smooth function. The **plug-in estimator** of $g(\mu)$ is $g(\hat{\mu}_n)$ where $\hat{\mu}_n$ is an estimator of μ. In particular, if nothing is known about X, $\hat{\mu}_n$ will be the sample mean $\overline{X}_n$. In general, $g(\overline{X}_n)$ is a biased estimator of $g(\mu)$. The bias is

$$\text{bias}(g(\overline{X}_n)) := \mathbb{E}g(\overline{X}_n) - g(\mu).$$

The goal is to estimate $\widehat{\text{bias}}(g(\overline{X}_n))$ using the bootstrap and replace $g(\overline{X}_n)$ with the bias-corrected estimator $g(\overline{X}_n) - \widehat{\text{bias}}(g(\overline{X}_n))$. Estimating the bias: For $i \in [B]$,

1. Generate the bootstrap sample $\hat{X}_1^{(i)}, \ldots, \hat{X}_n^{(i)}$.
2. $\hat{\Theta}^{(i)} = g\left(\widehat{\overline{X}}_n^{(i)}\right)$.
3. Let

$$\widehat{\text{bias}}(g(\overline{X}_n)) := \frac{1}{B}\sum_{i=1}^{B} g\left(\widehat{\overline{X}}_n^{(i)}\right) - g(\overline{X}_n).$$

Question: what is the bias of $g(\overline{X}_n) - \widehat{\text{bias}}(g(\overline{X}_n))$? Observe that

$$\begin{aligned} g(\overline{X}_n) - g(\mu) &= g(\mu + \overline{X}_n - \mu) - g(\mu) \\ &= g'(\mu)(\overline{X}_n - \mu) + \frac{1}{2}g''(\mu)(\overline{X}_n - \mu)^2 + \underbrace{o((\overline{X}_n - \mu)^2)}_{o_{\mathbb{P}}((n^{-1/2})^2)} + \frac{g'''(\tau)}{6}(\overline{X}_n - \mu)^3 \\ &\implies \mathbb{E}g(\overline{X}_n) - g(\mu) = \frac{1}{2}g''(\mu)\frac{\sigma^2}{n} + o(n^{-1}). \end{aligned}$$

Similarly,

$$g(\widehat{\overline{X}}_n) - g(\overline{X}_n) = g'(\overline{X}_n)(\widehat{\overline{X}}_n^{(i)} - \overline{X}_n) + \frac{1}{2}g''(\overline{X}_n)(\widehat{\overline{X}}_n^{(i)} - \overline{X}_n)^2 + o_{\mathbb{P}}(1/n).$$

Hence

$$\mathbb{E}_{\hat{\mathbb{P}}} g\left(\widehat{\overline{X}}_n^{(i)}\right) - g(\overline{X}_n) = \frac{1}{2}g''(\overline{X}_n)\frac{S_n^2}{n} + o(n^{-1})$$

where

$$S_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2.$$

Look at how close estimate is to actual bias:

$$\begin{aligned}\text{bias}(g(\overline{X}_n)) - \widehat{\text{bias}}(g(\overline{X}_n)) &= \frac{1}{2}g''(\mu)\frac{\sigma^2}{n} - \frac{1}{2}g''(\overline{X}_n)\frac{S_n^2}{n} + o_{\mathbb{P}}(n^{-1}) \\ &= \frac{1}{2}g''(\mu)\frac{\sigma^2}{n} - \frac{1}{2}g''(\overline{X}_n)\frac{S_n^2}{n} + \frac{1}{2}g''(\mu)\frac{S^2}{n} - \frac{1}{2}g''(\mu)\frac{S^2}{n} + o_{\mathbb{P}}(n^{-1}) \\ = \frac{1}{2}g''(\mu)\left(\frac{\sigma^2 - S_n^2}{n}\right) + \frac{S^2}{n}\frac{1}{2}\left[g''(\mu) - g''(\overline{X}_n)\right] + o_{\mathbb{P}}(n^{-1}) &= o_{\mathbb{P}}(n^{-1}) + o_{\mathbb{P}}(n^{-1}) = o_{\mathbb{P}}(n^{-1}) = o_{\mathbb{P}}(n^{-1}).\end{aligned}$$

Example 1.5.4 (**Trimmed mean**). $X_1, \ldots, X_n$ i.i.d. Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics. Pick $\alpha \in (0, 1/2)$. Define

$$\hat{\mu}_\alpha := \frac{1}{|i \in \{\lfloor n\alpha \rfloor, \ldots, n - \lfloor n\alpha \rfloor\}|} \sum_{i \in \{\lfloor n\alpha \rfloor, \ldots, n - \lfloor n\alpha \rfloor\}} X_{(i)}.$$

Exercise: what is the bias of this estimator? (Hint: prove that $\mathbb{E}X = \int_0^1 F^{-1}(t)\, dt$ where $F(t)$ is the cdf of X.)

Exercise 5. look carefully at this problem—could be on final.

X_1, X_2, X_3 i.i.d. with cdf F. Let $X_{(i)}$ be the order statistics, $i \in [3]$. Let $\hat{X}_1, \hat{X}_2, \hat{X}_3$ be the bootstrap sample, and $\hat{X}_{(1)}, \hat{X}_{(2)}, \hat{X}_{(3)}$ be the bootstrap order statistics.

(a) Find the distribution of $\hat{X}_{(2)}$.

(b) Find the bootstrap estimate of the bias of the sample median.

(c) Find the bootstrap estimate of the variance of the sample median.

Solution

(a)

$$\mathbb{P}(\hat{X}_{(2)} \le X_{(i)}) = \mathbb{P}(\hat{X}_{(1)} \le X_{(i)}, \hat{X}_{(2)} \le X_{(i)}) = \sum_{j=2}^{3}\binom{3}{j}\left(\frac{i}{3}\right)^j\left(1 - \frac{1}{3}\right)^{3-j}.$$

because the probability that one of the bootstrap sampled variables is less than or equal to the ith smallest sample is equal to $i/3$, so the probability that the two smallest bootstrap sampled variables are less than or equal to the ith smallest sample is equal to the probability that a binomial random variable with $n = 3, p = i/3$ equals 2 or 3. For $i = 1$, this equals 7/27; for $i = 2$, this equals 20/27; for $i = 3$, this equals 1. So then

$$\mathbb{P}(\hat{X}_{(2)} = X_{(1)}) = 7/27; \qquad \mathbb{P}(\hat{X}_{(2)} = X_{(2)}) = 13/27; \qquad \mathbb{P}(\hat{X}_{(2)} = X_{(3)}) = 7/27.$$

(b)

$$\widehat{\text{bias}} = \mathbb{E}_{\hat{\mathbb{P}}_3}(\hat{X}_2) - X_{(2)} = \frac{7}{27}X_{(1)} + \frac{13}{27}X_{(2)} + \frac{7}{27}X_{(3)} - X_{(2)} = \frac{7}{27}X_{(1)} - \frac{14}{27}X_{(2)} - \frac{7}{22}X_{(3)}.$$

(c) Note that

$$\mathbb{E}_{\hat{\mathbb{P}}_3}(\hat{X}_2 \mid X_1, X_2, X_3) = \frac{7}{27}X_{(1)} + \frac{13}{27}X_{(2)} + \frac{7}{27}X_{(3)}, \qquad \mathbb{E}_{\hat{\mathbb{P}}_3}\left(\hat{X}_2^2 \mid X_1, X_2, X_3\right) = \frac{7}{27}X_{(1)}^2 + \frac{13}{27}X_{(2)}^2 + \frac{7}{27}X_{(3)}^2,$$

so the bootstrap estimator of the variance is

$$\begin{aligned}
\widehat{\text{Var}}\left(\hat{X}_{(2)} \mid X_1, X_2, X_3\right) &= \mathbb{E}_{\hat{\mathbb{P}}_3}\left[\hat{X}_2^2 \mid X_1, X_2, X_3\right] - \mathbb{E}_{\hat{\mathbb{P}}_3}\left[\hat{X}_2 \mid X_1, X_2, X_3\right]^2 \\
&= \frac{7}{27}X_{(1)}^2 + \frac{13}{27}X_{(2)}^2 + \frac{7}{27}X_{(3)}^2 - \left(\frac{1}{27^2}\right)\left(7X_{(1)} + 13X_{(2)} + 7X_{(3)}\right)^2 \\
= \frac{7}{27}X_{(1)}^2 + \frac{13}{27}X_{(2)}^2 + \frac{7}{27}X_{(3)}^2 - \frac{1}{27^2}&\left(49X_{(1)}^2 + 2\cdot 91X_{(1)}X_{(2)} + 2\cdot 49X_{(1)}X_{(3)} + 169X_{(2)}^2 + 2\cdot 91X_{(2)}X_{(3)} + 49X_{(3)}^2\right) \\
= \left(\frac{7}{27} - \frac{49}{27^2}\right)X_{(1)}^2 + \left(\frac{13}{27} - \frac{169}{27^2}\right)X_{(2)}^2 + \left(\frac{7}{27} - \frac{49}{27^2}\right)X_{(3)}^2 - \frac{1}{27^2}&\left(2\cdot 91X_{(1)}X_{(2)} + 2\cdot 49X_{(1)}X_{(3)} + 2\cdot 91X_{(2)}X_{(3)}\right) \\
&= \boxed{\frac{140}{729}X_{(1)}^2 + \frac{182}{729}X_{(2)}^2 + \frac{140}{729}X_{(3)}^2 - \frac{182}{729}X_{(1)}X_{(2)} - \frac{98}{729}X_{(1)}X_{(3)} - \frac{182}{729}X_{(2)}X_{(3)}.}
\end{aligned}$$

1.6 Some Concentration of Measure

1.6.1 Concentration for Independent Sums

Can generate similar results for other random variables—just need different bound on moment-generating function (generally is fine as long as values of random variable are bounded between two real numbers, but more work to prove). However, doesn't work when values aren't bounded (e.g. for Gaussian random variables).

- What about other unbounded random variables?
- What about dependent random variables?

General question: **how far is a random variable from its mean?** Will first address functions of independent Gaussian random variables.

Definition 1.6.1 (**Lipschitz functions**)**.** A real-valued function $f : \mathbb{R}^n \to \mathbb{R}$ is called **Lipschitz continuous** or L**-Liptschitz** if there exists a positive real constant L such that for all $x_1, x_2 \in \mathbb{R}^n$,

$$|f(x_1) - f(x_2)| \leq L\|x_1 - x_2\|_2.$$

Theorem 1.6.1.1 (**Theorem 8.5 in 541A notes**)**.** Note from proof: use the fact that since f is 1-Lipschitz, $\|\frac{\mathrm{d}f(X)}{\mathrm{d}x_i}\|_2^2 \leq 1$.

Theorem 1.6.1.2 (**Theorem 8.6 in 541A notes**)**.** Notes from proof: change bounds on integral to $8a/\pi$. Then since $u \geq 8a/\pi \iff -a \geq -\pi u/8$, we have

$$\|\Pi(X)\| > u \implies \|\Pi(X)\| - a > u - a \geq u - \pi/8u > u/2.$$

Therefore

$$\Pr(\|\Pi(X)\| > u) = \Pr(\|\Pi(X)\| > u) \geq \Pr(|\|\Pi(x)\| - a| > u/2).$$

$$\vdots$$

Loose bound: $a^2 e^{-a^2} \leq 10$ for all $a > 0$. $(k > 1)$.

$$\vdots$$

$$\mathbb{E}\|\Pi(X)\|^4 \leq 2^{12}a^4 + 10^6 k^2$$

then by Jensen's Inequality,

$$a^4 = (\mathbb{E}\|\Pi(X)\|)^4 \leq (\mathbb{E}\|\Pi(X)\|^2)^2$$

So $2^{12}a^4 \leq 2^{12}(\mathbb{E}\|\Pi(X)\|^2)^2$. Then we chose 10^{10} to make things easy and say

$$2^{12}a^4 + 10^6 k^2 \leq 10^{10}(\mathbb{E}\|\Pi(X)\|^2)^2.$$

Summary:

$$Z := \|\Pi(X)\|^2, \qquad \mathbb{E}(Z^2) = k, \qquad \mathbb{E}(Z^4) \leq 10^{10}(\mathbb{E}(Z^2))^2.$$

$$\vdots$$

Union bound (Boole's Inequality)

1.7 Math 541B

Definition 1.7.1 (**Statistical model**). A **statistical model** is a family of probability distributions $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$. Assume that $X_1, \ldots, X_n$ are i.i.d. from $P_{\theta_0}, \theta_0 \in \Theta$.

Definition 1.7.2 (**Identifiability**).

$$\theta_1 \neq \theta_1 \implies P_{\theta_1} \neq P_{\theta_2}.$$

Definition 1.7.3 (**Estimator**). An **estimator** $\hat{\theta}(X_1, \ldots, X_n)$ is any measurable function of the data $(X_1, \ldots, X_n)$.

Definition 1.7.4 (**Bias**). The **bias** of an estimator is defined via

$$B(\hat{\theta}) := \mathbb{E}_{\theta_0}\hat{\theta}(X_1, \ldots, X_n) - \theta_0.$$

Example 1.7.1. Alice wants to send a message to Bob. A message is a sequence of symbols from $\{a, b, \ldots, z\} = \mathcal{A}$. **prefix-free codebook:** $\mathcal{A}$: sequence of 0, 1.

Theorem 1.7.0.1 (**Shannon's Theorem**). Assume that X has values in $\mathcal{A}$ and has distribution D. To minimize $\mathbb{E}_\rho$(number of bits), the codebook has to use about $\log_2(1/P(\beta)$ digits to encode $\rho \in \mathcal{A}$. Then

$$\mathbb{E}_\rho[\text{number of bits}] = \sum_{\beta \in \mathcal{A}} D(\beta)\log_2(1/p(\rho))$$

this is the entropy of D.

Assume that the codebook is created assuming that the true distribution of symbols is P but actually the true distribution is Q. In this case, the expected number of bits is

$$\mathbb{E}_Q[\text{number of bits}] = \sum_{\beta \in \mathcal{A}} Q(\rho)\log_2(1/p(\beta)).$$

we can also write this as the entropy of Q etc.

$$= \underbrace{\sum_{\beta \in \mathcal{A}} Q(\beta)\log_2(1/Q(\beta))}_{\text{entropy of } Q,\ \text{Ent}(Q)} + \underbrace{\sum_{\beta} Q(\beta)\log(Q(\beta)/p(\beta))}_{KL(Q||P)}$$

so the roots of MLE are in information theory.

1.8 Hypothesis Testing

Why is the likelihood ratio useful for hypothesis testing? It is a sufficient statistic. Suppose we have H_0 : $X \sim P$, $H_1 : X \sim Q$, and $\theta = 1 \implies \mathbb{P}_\theta = Q$, $\theta = 0 \implies \mathbb{P}_\theta = P$. Then the likelihood function is

$$L(\theta \mid X) = q(x)^{I\{\theta=1\}}p(x)^{1-I\{\theta=1\}Z} = \left(\frac{q(x)}{p(x)}\right)^{I\{\theta=1\}} \cdot p(x),$$

so the likelihood ratio is a sufficient statistic by the Factorization Theorem.

Bayes testing:

Prior probability π on H_0 (that distribution is on P), $1 - pi$ on H_a (distribution on Q). The if $\alpha_\phi := \mathbb{E}_P(\phi(X))$ is the probability of a type I error and $1\beta_\phi := 1 - \mathbb{E}_Q(\phi(X))$ is the probability of a type II error, we want to minimize the Bayes risk:

$$R_\pi(\phi) := \pi\alpha_\phi + (1-\pi)(1-\beta_\phi).$$

The Bayes test is

$$\phi^*_\pi(X) = \begin{cases} 1 & L(X) < (1-\pi)/\pi \\ \gamma & L(X) = (1-\pi)/\pi \\ 0 & L(X) > (1-\pi)/\pi \end{cases}$$

where $L(X) = P(X)/Q(X)$ is the likelihood ratio.

Mimimax: minimize worst case:

$$\min\left\{\max\left\{\alpha_\phi, 1-\beta_\phi\right\}\right\}.$$

Neymann-Pearson: acceptable Type I error rate α. So we have the constraint $\mathbb{E}_P(\phi(X)) \leq \alpha$. Maximize power subject to this constraint.

All give roughly similar tests.

Theorem 1.8.0.1. Assume that for some $\pi_* \in [0,1]$ that the Bayes test $\phi^*_{\pi_*}$ satisfies

$$\alpha^*_{\phi_{\pi_*}} = 1 - \beta_{\phi^*_{\pi_*}}. \tag{1.26}$$

Then $\phi^*_{\pi_*}$ is minimax. Moreover, a π_* with this desired property always exists.

Proof. Assume that $\phi^*_{\pi_*}$ is a Bayes test satisfying (1.26), but is not minimax. Then there must exist $\tilde{\phi}$ such that

$$\min\left\{\max\left\{\alpha_{\tilde{\phi}}, 1-\beta_{\tilde{\phi}}\right\}\right\} < \min\left\{\max\left\{\alpha^*_{\phi_{\pi_*}}, 1-\beta^*_{\phi_{\pi_*}}\right\}\right\}$$

and note that

$$\max\left\{\alpha_{\tilde{\phi}}, 1-\beta_{\tilde{\phi}}\right\} < \max\left\{\alpha^*_{\phi_{\pi_*}}, 1-\beta^*_{\phi_{\pi_*}}\right\} = \phi^*_{\pi_*} \text{ (by (1.26))}$$

We have

$$R_{\pi_*}(\tilde{\phi}) = \pi_* \alpha_{\tilde{\phi}} + (1-\pi_*)(1-\beta_{\tilde{\phi}}) < \pi_* \alpha^*_{\phi_{\pi_*}} + (1-\pi_*)(1-\beta^*_{\phi_{\pi_*}}) = R_{\pi_*}(\phi^*_{\pi_*})$$

But this contradicts the fact that the Bayes test minimizes this ...

To prove existence, consider for $\pi \in [0,1]$ the non-randomized test

$$\phi_\pi = \begin{cases} 1 & L(X) \leq (1-\pi)/\pi \\ 0 & L(x) > (1-\pi)/\pi \end{cases}$$

Since the range of π varies the Type I and Type II error rate, if we could show this function is continuous in π then we would be done by the Intermediate Value Theorem (the point where both error rates are equal would exist). We can accommodate the possibility of a jump discontinuity right at what would have been the best value by using a randomized test that takes an appropriate convex combination

□

Theorem 1.8.0.2 (same theorem, re-stated on 09/13). *$X : \Omega \to S$. $H_0 : X \sim P$. $X_a : X \sim Q$. Assume μ is a probability measure on Ω. And P has density p with respect to μ, Q has density q with respect to μ. Let $\phi : S \to [0,1]$ be a measurable function (randomized statistical test; the value it takes is interpreted as a probability). Let $\alpha_\phi := \mathbb{E}_P \phi(X)$ be the probability of a Type I error, and let $\beta_\phi := \mathbb{E}_Q \phi(X)$.*

Assume $\pi_* \in [0,1]$. Then the Bayes optimal test $\phi^*_{\pi_*}$ is such that $\alpha_{\phi^*_{\pi_*}} = 1 - \beta_{\phi^*_{\pi_*}}$. Then $\phi^*_{\pi_*}$ is minimax optimal. Moreover, π_* with required properties holds.

Proof of second claim. We now show that π_* and $\phi^*_{\pi_*}$ exist. Let $\pi \in [0,1]$. Consider the non-randomized

$$\phi_\pi(x) = \begin{cases} 1 & \text{if } \frac{p(x)}{q(x)} \leq \frac{1-\pi}{\pi} \\ 0 & \text{if } \frac{p(x)}{q(x)} > \frac{1-\pi}{\pi} \end{cases}$$

Let

$$L(x) := \frac{p(x)}{q(x)}.$$

Consider

$$G(\pi) := \underbrace{Q(L(x) > (1-\pi)/\pi)}_{\mathbb{P} \text{ type II errors of } \pi_\pi} - \underbrace{P(L(x) \leq (1-\pi)/\pi)}_{\mathbb{P} \text{ type I error of } \phi_\pi}$$

$$= Q(L(x) > (1-\pi)/\pi) + P(L(x) > (1-\pi)/\pi) - 1$$

Observe:

(a) $\lim_{\pi \to 0^+} G(\pi) = -1$, since the probability of being over infinity is 0.

(b) $G(1) = Q(L(x) > 0) + P(L(x) > 0) - 1 = Q(L(x) > 0) \geq 0$.

(c) G is nondecreasing, continuous from the left (because a CDF is right continuous, but the arguments are decreasing as π increases, so instead it is left continuous.) So, G will either intersect the π axis or it will jump over it. That is, there exists so $\pi_* \in (0, 1]$ such that either (i) $G(\pi_*) = 0$ or (ii) G jumps over the π axis; $G(\pi_*) < 0$ and $\lim_{\pi \to \pi_*^+} G(\pi) > 0$.

Case (i). In this case we are done. $\phi^*_{\pi_*}$ is minimax optimal by the first part of the theorem.

Case (ii). Let $\gamma := G(\pi_*^+)/(G(\pi_*^+) - G(\pi_*)) \in (0, 1)$. Let

$$\phi'_{\pi_*} = \begin{cases} 1 & L(x) < (1 - \pi_*)/\pi_* \\ \gamma & L(x) = (1 - \pi_*)/\pi_* \\ 0 & L(x) > (1 - \pi_*)/\pi_* \end{cases}$$

ϕ'_{π_*} is (randomized) Bayes optimal for the prior π_*. Note: The difference of type Ii and type II errors of ϕ'_{π_*} is

$$\underbrace{Q(L(x) > (1 - \pi_*)/\pi_*) + (1 - \gamma)Q(L(x) = (1 - \pi_*)/\pi_*)}_{\text{Type II error}} - \underbrace{P(L(x) < (1 - \pi_*)/\pi_*) - \gamma P(L(x) = (1 - \pi_*)/\pi_*)}_{\text{Type I error}}$$

Using $P(L(x) < (1 - \pi_*)/\pi_*) = 1 - P(L(x) > (1 - \pi_*)/\pi_*) - P(L(x) = (1 - \pi_*)/\pi_*)$, we can write this as

$$= Q(L(x) > (1-\pi_*)/\pi_*) + P(L(x) > (1-\pi_*)/\pi_*) - 1 + (1-\gamma)\left[Q(L(x) = (1 - \pi_*)/\pi_*) + P(L(x) = (1 - \pi_*)/\pi_*)\right] \tag{1.27}$$

For the γ we defined, (1.27) is 0. Why?

$$1 - \gamma = -G(\pi_*)/(G(\pi_*^+) - G(\pi_*))$$

Recall that a jump in a cdf occurs when the corresponding random variable takes on a fixed value with nonzero probability. Compute

$$G(\pi_*^+) - G(\pi_*) = Q(L(x) = (1 - \pi)/\pi) - P(L(X) = (1 - \pi)/\pi).$$

so we see everything cancels;

$$-(1-\gamma)\left[Q(L(x) = (1 - \pi_*)/\pi_*) + P(L(x) = (1 - \pi_*)/\pi_*)\right] = Q(L(x) > (1-\pi_*)/\pi_*) + P(L(x) > (1-\pi_*)/\pi_*) - 1$$

and (1.27) equals 0. That is, the difference is zero, so ϕ'_{π_*} is minimax optimal by the first part of the theorem.

□

1.8.1 Neyman-Pearson Tests

Definition 1.8.1 (**Power function; definition 8.3.1 in Casella and Berger [2001]**). The **power function** of a hypothesis test with rejection region R is the function of θ defined by

$$\beta(\theta) := \mathbb{P}_\theta(x \in R).$$

Motivation: some errors are worse than others. Goal: control Type I error, keeping it small for any sample size.

Define the **test function** $\phi(x)$ to be an indicator of x being in the rejection region for a test:

$$\phi(x) = \begin{cases} 1 & x \in R \\ 0 & x \notin R \end{cases}.$$

Let Φ be a set of such tests.

Definition 1.8.2 (**UMP; Definition 8.3.11 in Casella and Berger [2001],**). ϕ^* is **uniformly most powerful** (UMP) of size $\alpha \in [0,1]$ if it achieves the maximum value of B_ϕ subject to $\alpha_\phi \leq \alpha$ over all randomized tests $\phi \in \Phi$. That is, it is uniformly most powerful if and only if it solves

$$\phi^* = \begin{array}{lll} \arg\max_{\phi \in \Phi} & \beta_\phi(\theta_a) & \\ \text{subject to} & \beta_\phi(\theta_0) \leq \alpha \quad \forall \theta_0 \in \Theta_0 & \forall \theta_a \in \Theta_a. \end{array}$$

or (given in class, definitely less formal notation)

$$\beta_\phi = \mathbb{E}_Q \phi(x) \to \max_{\phi \in \Phi} \text{ s.t. } \alpha_\phi = \mathbb{E}_p \phi(x) \leq \alpha.$$

Lemma 1.8.1.1 (**Neyman-Pearson; Theorem 8.3.12, p. 388 in Casella and Berger [2001], Theorem 3.2.1 in Lehmann and Romano [2005]**). Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where the pdf or pmf corresponding to θ_i is $f(x \mid \theta_i)$, $i \in \{0,1\}$. Suppose we have a test ϕ^* such that

$$\phi^* = \begin{cases} 1, & f(x \mid \theta_1) > cf(x \mid \theta_0) \qquad (q(x) > cp(x)) \\ 0, & f(x \mid \theta_1) < cf(x \mid \theta_0) \qquad (q(x) < cp(x)) \end{cases} \tag{1.28}$$

(recall that $\phi^*(x) = 1$ when $x \in R$ and 0 otherwise) for some $c \geq 0$ and

$$\alpha_{\phi^*} = \mathbb{P}_{\theta_0}(x \in R) = \beta^*(\theta_0) = \alpha \tag{1.29}$$

(where β^* is the power test for ϕ^*). Then

1. Any test ϕ^* that satisfies (1.28) and (1.29) is a UMP level α test.
2. There exist $c \geq 0$ and ϕ^* that satisfy (1.28) and (1.29).
3. If $\tilde{\phi}$ is UMP of size α, then it satisfies (1.28) almost surely. Moreover, $\alpha_{\tilde{\phi}} = \alpha$ unless $p_{\tilde{\phi}} = 1$. (In other words, the only complication/exception is if you have a test that has power 1 and Type I error rate less than α; then obviously you choose the test with power 1 and minimal Type I error rate, which will be less than α.)

   (Claim from Casella and Berger [2001]) If there exists a test satisfying (1.28) and (1.29), then every UMP level α test is a size α test (satisfies (1.29)) and every UMP level α test satisfies (1.28) except perhaps on a set A satisfying $\mathbb{P}_{\theta_0}(X \in A) = \mathbb{P}_{\theta_1}(x \in A) = 0$.

Proof (in-class, 541B). 1. Suppose that ϕ is the test function for another level α test; that is, if β is the power function for ϕ,

$$\sup_{\theta \in \{\theta_0\}} \beta(\theta) = \beta(\theta_0) \leq \alpha. \tag{1.30}$$

Note that

$$\int_S (\phi^* - \phi) \left[f(x \mid \theta_1) - c f(x \mid \theta_0) \right] \, d\mu \geq 0 \tag{1.31}$$

for the following reason: by (1.28), when $x \in R$, $\phi^* = 1$, so $\phi^* - \phi \in [0, 1]$. Also, in this region, $f(x \mid \theta_1) > c f(x \mid \theta_0)$, so the second term is positive. On the other hand, when $x \in R^C$, $\phi^* = 0$, so $\phi^* - \phi \in [-1, 0]$, and $f(x \mid \theta_1) < c f(x \mid \theta_0)$ so the second term is negative.

On the other hand, we can rewrite (1.31) as follows:

$$0 \leq \int_S \phi^* f(x \mid \theta_1) \, d\mu - \int_S \phi f(x \mid \theta_1) \, d\mu - c \int_S \phi^* f(x \mid \theta_0) \, d\mu + c \int_S \phi f(x \mid \theta_0) \, d\mu$$

$$\iff 0 \leq \beta^*(\theta_1) - \beta(\theta_1) - c \left[\beta^*(\theta_0) - \beta(\theta_0) \right] \tag{1.32}$$

(where β^* is the power function for ϕ^* and β is the power function for ϕ). Note that $c \left[\beta^*(\theta_0) - \beta(\theta_0) \right] \geq 0$ since $c \geq 0$ by assumption, $\beta^*(\theta_0) = \alpha$ by (1.29), and $\beta(\theta_0) \leq \alpha$ by (1.30). Therefore by (1.32),

$$0 \leq \beta^*(\theta_1) - \beta(\theta_1) - c \left[\beta^*(\theta_0) - \beta(\theta_0) \right] \leq \beta^*(\theta_1) - \beta(\theta_1);$$

that is, β^* is UMP over all level α tests.

2. We will use the randomization possible when $q(x) = cp(x)$ to make the Type I error rate equal α exactly. Consider the test

$$\phi^*(x) = \begin{cases} 1, & q(x) > cp(x) \\ \gamma, & q(x) = cp(x) \\ 0, & q(x) < cp(x) \end{cases} \tag{1.33}$$

for some $\gamma \in [0, 1]$ and some $c \geq 0$. The size of ϕ^* is

$$\alpha_{\phi^*} = \mathbb{P}_P(x : q(x) > cp(x)) + \gamma \mathbb{P}_P(x : q(x) = cp(x))$$

$$= \int_S \phi^*(x)p(x)\ d\mu$$

so it remains to show that for any $\alpha \in [0,1]$, there exists c, γ such that $\alpha_{\phi^*} = \alpha$. Let $H(c) := \mathbb{P}_P(x : q(x) \leq cp(x))$. Then

(a) $H(c)$ is non-decreasing.

(b) $H(c)$ is right-continuous (like a cdf, because of the $\leq$).

Then $\alpha_{\phi^*} = (1 - H(c)) + \gamma[H(c) - \lim_{y \to c^+} H(y)]$, and we want this to equal α.(Notice that this equation is very similar to the ones we encountered when studying minimax tests.) We will consider two cases.

(a) If there exists a c such that $1 - H(c) = \alpha$, simply take $\gamma = 0$.

(b) If there does not exist a c such that $1 - H(c) = \alpha$; that is, $1 - H(c) < \alpha$, $\lim_{y \to c^+} 1 - H(y) > \alpha$, we can set γ appropriately. (Exercise: find the γ.)

More notes:

$$H(c) = \mathbb{P}(x : q(x) \leq cp(x)) = \mathbb{P}(x : q(x)/p(x) \leq x).$$

We can divide by $p(x)$ because the probability that $p(x) = 0$ is 0. Now H is kind of like a distribution function. Consider:

$c < 0 \implies H(c) = 0$. Also, as $c \to \infty$, $H \to 1$. So we can always solve for $c \geq 0$, $\gamma \in [0,1]$ such that

$$1 - H(c) + \gamma[H(c) - H(c^-)] = \alpha.$$

We will show that for any $0 \leq \alpha \leq 1$ there exist $c(\alpha)$ and $\gamma(\alpha)$ such that $\mathbb{E}_{\mathbb{P}}\Phi^*(X) = \alpha$ (meaning the test Φ^* has size α). We have

$$\phi^*(x) = \begin{cases} 1, & q(x) > c(\alpha)p(x) \\ \gamma(\alpha), & q(x) = c(\alpha)p(x) \\ 0, & q(x) < c(\alpha)p(x) \end{cases}$$

We seek $c(\alpha)$ and $\gamma(\alpha)$ satisfying

$$\alpha = \mathbb{E}_P\phi^*(X) = \mathbb{P}_P(\{x : q(x) > cp(x)\}) + \gamma(\alpha)\mathbb{P}_P(\{x : q(x) = cp(x)\}) \tag{1.34}$$

First, simply let

$$\boxed{c(\alpha) := \inf_{c \in \mathbb{R}} \{c : \mathbb{P}_P(\{x : q(x) > cp(x)\}) \leq \alpha\} = \inf_{c \in \mathbb{R}} \left\{c : \mathbb{P}_P\left(\left\{x : \frac{q(x)}{p(x)} > c\right\}\right) \leq \alpha\right\}} \tag{1.35}$$

(where we can divide by $p(x)$ since we are evaluating this probability over P so we only need to consider x such that $p(x) > 0$). Suppose that

$$\mathbb{P}_P(q(x)/p(x) = c) = 0. \tag{1.36}$$

Then we do not need to find $\gamma(\alpha)$ because ϕ^* is already fully defined by this $c(\alpha)$: that is, it holds that the minimum of the set in (1.35) exists and is the infimum, so

$$\mathbb{P}_P\left(\left\{x : \frac{q(x)}{p(x)} > c(\alpha)\right\}\right) = \alpha$$

and we are done with an arbitrary $\gamma(\alpha)$, say $\gamma(\alpha) = 0$. The case where (1.36) does not hold is more complicated. Let

$$a(k) := \mathbb{P}_P(\{x : q(x) > kp(x)\}) = \mathbb{P}_P\left(\left\{x : \frac{q(x)}{p(x)} > k\right\}\right)$$

(again, the last step is permissible since we are calculating this probability under P). Then since $a(k)$ is the probability that the random variable $q(X)/p(X)$ exceeds k under P, $1 - a(k)$ can be considered as the cdf of the random variable $q(x)/p(x)$ as a function of k, with $\lim_{k\to-\infty}(1 - a(k)) = 0$ and $\lim_{c\to\infty}(1 - a(k)) = 1$. This means that $1 - a(k)$ is nondecreasing right-continuous, so $a(k))$ is nonincreasing and right-continuous. As a consequence,

$$\mathbb{P}_P(q(x)/p(x) = c(\alpha)) = [1-a(c(\alpha))] - \lim_{y\to c(\alpha)^-}[1-a(y)] = \lim_{y\to c(\alpha)^-} a(y) - a(c(\alpha)) =: a(c(\alpha)^-) - a(c(\alpha)),$$

(where the last step simply introduces a simpler notation) and we have

$$\mathbb{P}_P(q(x)/p(x) = c(\alpha)) \neq 0 \iff a(c(\alpha)^-) - a(c(\alpha)) \neq 0. \tag{1.37}$$

If (1.37) holds, $\mathbb{P}_P\left(\left\{x : \frac{q(x)}{p(x)} > c(\alpha)\right\}\right) < \alpha$; in particular,

$$\mathbb{P}_P\left(\frac{q(x)}{p(x)} > c(\alpha)\right) = a(c(\alpha)) < \alpha.$$

Therefore in general (without any assumptions about whether $\mathbb{P}_P(\{x : q(x) = c(\alpha)p(x)\}) = 0$) the size of ϕ as defined in (1.34) is

$$\begin{aligned}\mathbb{E}_P(\phi^*(X)) &= \mathbb{P}_P(\{x : q(x)/p(x) > c(\alpha)\}) + \gamma(\alpha)\mathbb{P}_P(\{x : q(x)/p(x) = c(\alpha)\}) \\ &= a(c(\alpha)) + \gamma(\alpha)\left[a(c(\alpha)^-) - a(c(\alpha))\right]\end{aligned}$$

Because of this, it is clear that defining

$$\begin{aligned}\gamma(\alpha) &:= \begin{cases}\frac{\alpha - a(c(\alpha))}{a(c(\alpha)^-) - a(c(\alpha))}, & \mathbb{P}_P(q(x)/p(x) = c(\alpha)) \neq 0 \\ 0, & \mathbb{P}_P(q(x)/p(x) = c(\alpha)) = 0\end{cases} = \begin{cases}\frac{\alpha - a(c(\alpha))}{a(c(\alpha)^-) - a(c(\alpha))}, & a(c(\alpha)^-) - a(c(\alpha)) \neq 0 \\ 0, & a(c(\alpha)^-) - a(c(\alpha)) = 0\end{cases} \\ &= \begin{cases}\frac{\alpha - \mathbb{P}_P(\{x:q(x)>c(\alpha)p(x)\})}{\mathbb{P}_P(\{x:q(x)>c(\alpha)^-p(x)\}) - \mathbb{P}_P(\{x:q(x)>c(\alpha)p(x)\})}, & \mathbb{P}_P(\{x : q(x) > c(\alpha)^-p(x)\}) - \mathbb{P}_P(\{x : q(x) > c(\alpha)p(x)\}) \neq 0 \\ 0, & \mathbb{P}_P(\{x : q(x) > c(\alpha)^-p(x)\}) - \mathbb{P}_P(\{x : q(x) > c(\alpha)p(x)\}) = 0\end{cases}\end{aligned}$$

leads to (1.34) being satisfied.

3. Consider again (1.31), except now let ϕ be any UMP level α test. Since by part (1) ϕ^* is also UMP level α, we have $\beta^*(\theta_1) = \beta(\theta_1)$, so from (1.32) we have

$$0 \leq \beta^*(\theta_1) - \beta(\theta_1) - c\left[\beta^*(\theta_0) - \beta(\theta_0)\right] = c\left[\beta(\theta_0) - \beta^*(\theta_0)\right] \implies \beta(\theta_0) - \beta^*(\theta_0) \geq 0$$

Recall that $\beta^*(\theta_0) = \alpha$ by (1.29), so $\beta(\theta_0) \geq \alpha$. But $\beta(\theta_0) \leq \alpha$ since it is UMP level α. Therefore we must have $\beta(\theta_0) = \alpha$, so ϕ satisfies (1.29) (unless there is a set A such that $\int_S f(x \mid \theta_1)\, dx = \int_S f(x \mid \theta_0)\, dx = 0$).

(Notes from class) Consider

$$A = \int_S (\phi^* - \tilde{\phi}) \underbrace{(q - cp)}_{\text{positive on } c_+} \, d\mu \tag{1.38}$$

where ϕ^* is the UMP test (1.33) from part (2). Then (1.38) is greater than or equal to 0 ($A \geq 0$) by part 1 of the lemma. On the other hand,

$$A = \underbrace{\beta_{\phi^*} - \beta_{\tilde{\phi}}}_{=0 \text{ if they are both UMP}} - \underbrace{c}_{\geq 0} \left(\underbrace{\alpha_{\phi^*}}_{=\alpha} - \underbrace{\alpha_{\tilde{\phi}}}_{\leq \alpha} \right).$$

They must both be UMP tests, so the above notes follow. So either $c = 0$ or $\alpha_{\phi^*} = \alpha_{\tilde{\phi}}$. If $c = 0$,

$$\beta_{\phi^*} = \int \phi^*(x) q(x)\, d\mu = 1$$

$$\tilde{\phi} = \begin{cases} 1, & q(x) > 0 \\ 0, & q(x) < 0 \end{cases}$$

(but of course $q(x)$ is never negative). Then $\beta_{\tilde{\phi}} = \int \tilde{\phi}(x) q(x)\, d\mu = 1$.

Note that we don't need to know the γ, because

The other possibility is $\alpha_{\phi^*} = \alpha_{\tilde{\phi}} = \alpha$. Consider the set $c_+ = \{x : q(x) > cp(x)\} \implies \tilde{\phi}(x) = 1$ on c_+. And $c_- = \{x : q(x) < cp(x)\} \implies \tilde{\phi}(x) = 0$ on c_1. Finally, since $\alpha_{\tilde{\phi}} = \alpha_{\phi^*} = \alpha$, $\tilde{\phi}(x) = \gamma$ when $q(x) = cp(x)$, then γ is the same for ϕ^*, so the tests must be equal.

□

Geometric illustration (Figure 1.1): consider the mapping $T : \Phi \to [0,1] \times [0,1]$ where Φ is the set of all possible tests where $T(\phi) = (\alpha_\phi, \beta_\phi)$. Observation: $(\alpha, \beta) \in \text{im}(T) \implies (1-\alpha, 1-\beta) \in \text{im}(T)$ because if $T(\phi) = (\alpha, \beta)$, then $T(1-\phi) = (1-\alpha, 1-\beta)$. So we have symmetry around the point $(1/2, 1/2)$. This picture shows why there is an exception if $\beta = 1$; then you want the test with minimum α given maximum power. But in practical terms this case is not really important; this is more of a technical detail.

Exercise 6. 1. Let ϕ^* be the MP test of size α. Then $\beta_{\phi^*} \geq \alpha$. Moreover, β_{ϕ^*} is strictly greater than α unless $P = Q$ (Use N-P lemma, part 1).

2. Let $X_1, \ldots, X_n$ be i.i.d. uniform. $H_0 : X_1, \ldots, X_n \sim \text{U}(0,1)$. $H_a : X_1, \ldots, X_n \sim U(1/3, 2/3)$. Derive a UMP test of size α for all values of α.

Solution

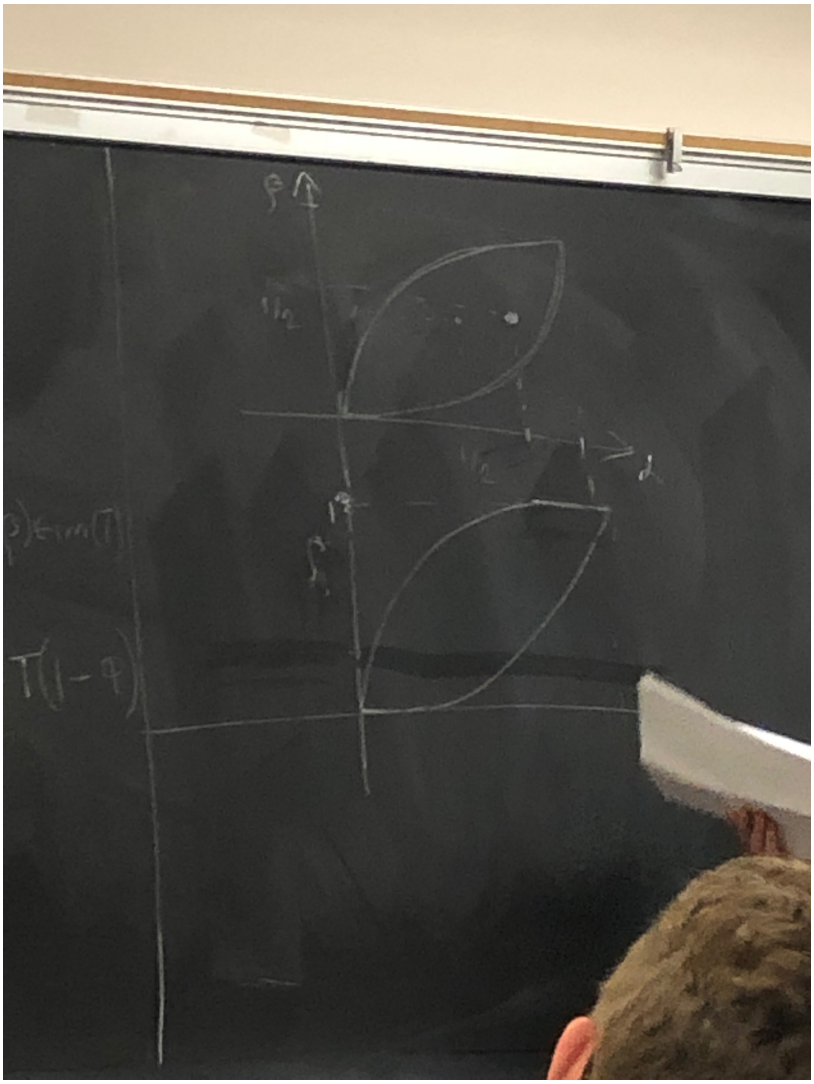

Figure 1.1: Sets of possible Neyman-Pearson tests; similar to Figure 3.1 in Lehmann and Romano [2005].

1.

2.
$$q(x_1, \dots, x_n) = e^n \prod_{i=1}^{n} I\{x_i \in [1/3, 2/3]]\} = e^n I\{x_1, \dots, x_n \in [1/3, 2/3]\}$$

Similarly,

$$p(x_1, \dots, x_n) = I\{x_1, \dots, x_n \in [0, 1]\}.$$

Note that the ratio q/p only takes on three values. By the Neyman-Pearson Lemma (Lemma 1.8.1.1), the most powerful test ϕ^* is either

$$\phi_1^*(x_1, \dots, x_n) = \begin{cases} 1, & q(x_1, \dots, x_n)/p(x_1, \dots, x_n) = 3^n \\ \gamma, & q(x_1, \dots, x_n)/p(x_1, \dots, x_n) = 0 \end{cases}$$

or

$$\phi_2^*(x_1, \dots, x_n) = \begin{cases} \gamma, & q(x_1, \dots, x_n)/p(x_1, \dots, x_n) = 3^n \\ 0, & q(x_1, \dots, x_n)/p(x_1, \dots, x_n) = 0 \end{cases}$$

Now

$$\alpha_{\phi_1^*} = \mathbb{E}_P \phi_1^*(X_1, \dots, X_n) = 1 \cdot \mathbb{P}(X_1, \dots, X_n \in [1/3, 2/3]) + \gamma \mathbb{P}(\text{at least one } X_j \text{ is outside } [1/3, 2/3])$$

$$= 3^{-n} + \gamma(1 - 3^{-n}) = \alpha$$

which has a solution for $\gamma(\alpha)$ for $\alpha \geq 3^{-n}$. On the other hand,

$$\alpha_{\phi_2^*} = \mathbb{E}_P \phi_2^*(X_1, \dots, X_n) = \gamma \mathbb{P}(X_1, \dots, X_n \in [1/3, 2/3]) + 0 \cdot \mathbb{P}(\text{at least one } X_j \text{ is outside } [1/3, 2/3])$$

$$= \gamma 3^{-n} = \alpha$$

which has a unique solution $\gamma(\alpha)$ for $\alpha \leq 3^{-n}$. So for a test with a very small α, you need to consider a strictly randomized test like $\alpha_{\phi_2^*}$.

Theorem 1.8.1.2 (Went over last time). We have a sequence $X_1, \dots, X_n$ distributed i.i.d. Last time we showed that the following hypothesis testing problem:

$$H_0 : X_1, \dots, X_n \sim P, \qquad H_A : X_1, \dots, X_n \sim Q$$

with

$$\frac{dP}{d\mu} = p, \qquad \frac{dQ}{d\mu} = q$$

using the test

$$\phi^*(x_1,\dots,x_n) = \begin{cases} 1, & \prod_{i=1}^n q(x_i) > c\prod_{i=1}^n p(x_i) \\ \gamma, & \prod_{i=1}^n q(x_i) = c\prod_{i=1}^n p(x_i) \\ 0, & \prod_{i=1}^n q(x_i) < c\prod_{i=1}^n p(x_i) \end{cases} \tag{1.39}$$

is UMP of size $\alpha = \mathbb{E}_P(\phi^*)(X_1,\dots,X_n)$ (given the correct choice of γ).

Remark 34. Intuition: if the likelihood function for Q exceeds the likelihood function of P by a reasonable amount (determined by c), choose Q to be more likely than P.

Also, ϕ^* of this kind is referred to as a *Neyman-Pearson test.*

Proposition 1.8.1.3 (**Math 541B Midterm Problem 2**)**.** The uniformly least powerful test of size $\alpha > 0$ is

Proof. Consider again the mapping $T : \Phi \to [0,1]\times[0,1]$ where Φ is the set of all possible tests where $T(\phi) = (\alpha_\phi, \beta_\phi)$ (See Figure 1.1). Observation: $(\alpha,\beta) \in \text{im}(T) \implies (1-\alpha, 1-\beta) \in \text{im}(T)$ because if $T(\phi) = (\alpha,\beta)$, then $T(1-\phi) = (1-\alpha, 1-\beta)$. So we have symmetry around the point $(1/2, 1/2)$. Because of this, we can see that if we find the UMP test of size $1-\alpha$ and reverse the rejection and acceptance regions, we obtain the uniformly least powerful test of size α. By the Neyman-Pearson Lemma, the uniformly most powerful test of size α is

$$\phi^*(x) = \begin{cases} 1, & q(x) > cp(x) \\ 0, & q(x) \le cp(x) \end{cases}$$

for c chosen so that the size of the test is $1-\alpha$. Then the uniformly least powerful test of size α is

$$\tilde{\phi}^*(x) = \begin{cases} 1, & q(x) \le cp(x) \\ 0, & q(x) > cp(x). \end{cases}$$

□

1.8.2 Consistency of Neyman-Pearson Tests

We want the test to be consistent—as we collect more data and $n \to \infty$, errors of both types go to 0. We hope to show the consistency of Neyman-Pearson tests.

Definition 1.8.3 (**Consistency of a hypothesis test**)**.** Consider the framework of the Neyman-Pearson lemma (Lemma 1.8.1.1, described above). A sequence of tests $\{\phi_n\}_{n\ge1}$, where $\phi_n = \phi_n(x_1,\dots,x_n)$ is *consistent* if and only if

$$\alpha_{\phi_n} = \mathbb{E}_P\phi_n(X_1,\dots,X_n) \to 0, \qquad \text{and } \beta_{\phi_n} = \mathbb{E}_Q\phi_n(X_1,\dots,X_n) \to 1 \qquad \text{as } n\to\infty.$$

It turns out that c grows with n, and the growth condition of the size of this c_n is related to a measure of distance between P and Q (Hellinger distance).

Recall the following definition:

Definition 1.8.4 (**Hellinger Distance and Affinity (Definition 13.1.3 in Lehmann and Romano [2005])**). The *Hellinger distance* between probability laws P and Q is defined as

$$H^2(P,Q) := \int_S (\sqrt{p} - \sqrt{q})^2 \, d\mu = 2\int_S (1 - \sqrt{pq}) d\mu = 2 - 2\int_S \sqrt{pq} \, d\mu.$$

(See also Section 1.4.6.) Further,

$$A(P,Q) := \int \sqrt{pq} \, d\mu$$

is known as the *Hellinger affinity.*

Remark 35. This is a true distance—it satisfies the Triangle Inequality, $H(P,Q) = 0$ if and only if $P = Q$ almost everywhere.

Lemma 1.8.2.1. Let $X \sim P$ and $Y \sim Q$. Let $X_1, \ldots, X_n$ be i.i.d. copies of X and likewise for $Y_1, \ldots, Y_n$, so $(X_1, \ldots, X_n) \sim P^n$ and $(Y_1, \ldots, Y_n) \sim Q^n$. Then

$$A(P^n, Q^n) = [A(P,Q)]^n .$$

Proof.

$$A(P^n, Q^n) = \int \sqrt{p(x_1, \ldots, x_n)}\sqrt{q(x_1, \ldots, x_n)} \, d\mu^n = \prod_{j=1}^{n} \int \sqrt{p(x_j)q(x_j)} \, d\mu_j = [A(P,Q)]^n .$$

□

Theorem 1.8.2.2 (**Consistency of Neyman-Pearson tests**). Assume that $\rho := A(P,Q) < 1$. Then take any sequence $\{c_n\} \subset \mathbb{R}_+$ such that

$$\rho^{2n} \ll c_n \ll \rho^{-2n}. \tag{1.40}$$

Then any sequence $\{\phi_n\}_{n\geq 1}$ of Neyman-Pearson tests corresponding to $\{c_n\}_{n\geq 1}$ is consistent.

Proof. First, we will show that $\alpha_{\phi_n} \to 0$ under these conditions. Recall the definition of ϕ from (1.39).

$$\begin{aligned}\alpha_{\phi_n} &= \int \phi(x_1,\dots,x_n)p(x_1,\dots,x_n)\ d\mu^n \\ &\le \int I\left\{\frac{\prod_{i=1}^n q(x_i)}{\prod_{i=1}^n p(x_i)} \ge c_n\right\} p(x_1)\cdots p(x_n)\ d\mu^n = \int I\left\{\frac{\prod_{i=1}^n q(x_i)}{c_n\prod_{i=1}^n p(x_i)} \ge 1\right\} p(x_1)\cdots p(x_n)\ d\mu^n \\ &\le \int \sqrt{\frac{q(x_1)\cdots q(x_n)}{c_n p(x_1)\cdots p(x_n)}}p(x_1)\cdots p(x_n)\ d\mu^n \\ &= \frac{1}{\sqrt{c_n}}\int \sqrt{q(x_1)\cdots q(x_n)}\sqrt{p(x_1)\cdots p(x_n)}d\mu^n = \rho^n c_n^{-1/2} \to 0 \text{ as } n\to\infty.\end{aligned}$$

(The second inequality follows because when the indicator equals 0 the square root term is greater than or equal to 0 but when the indicator equals 1 the square root term is greater than or equal to 1. The last step follows by assumption of the lemma.)

Now consider the Type II error $1-\beta_{\phi_n}$.

$$\begin{aligned}1-\beta_{\phi_n} &= \int [1-\phi(x_1,\dots,x_n)]q(x_1)\cdots q(x_n)\ d\mu^n \\ &\le \int I\left\{q(x_1)\cdots q(x_n) \le cp(x_1)\cdots p(x_n)\right\} q(x_1)\cdots q(x_n)\ d\mu^n \\ &\le \int \sqrt{\frac{p(x_1)\cdots p(x_n)}{q(x_1)\cdots q(x_n)}\cdot c_n}\cdot q(x_1)\cdots q(x_n)\ d\mu^n \\ &= \sqrt{c_n}A(P^n,Q^n) = c - n^{1/2}\rho^n \to 0 \text{ as } n\to\infty\end{aligned}$$

□

Remark 36. Intuition: $A(P,Q)<1 \iff H(P,Q)>0 \iff P\neq Q$. Also, $a_n \ll b_n \iff a_n = o(b_n)$, and $a_n \gg b_n \iff b_n \ll a_n$. Almost all reasonable sequences satisfy (1.40)—this range is very large.

Basically, this theorem says the Type I and Type II errors converge to 0 geometrically fast (as ρ^n) if P and Q are not too close together.

Proposition 1.8.2.3 (Stats 100B Homework problem). Let $Y_1, Y_2, \dots, Y_n$ be the outcomes of n independent Bernoulli trials. Then by the Neyman-Pearson lemma (Lemma 1.8.1.1), the best critical region for testing

$$H_0 : p = p_0 \qquad H_a : p > p_0$$

is

$$\frac{y}{n} = \frac{1}{n}\sum Y_i > \frac{\log(K) + n\log\left(\frac{1-p_a}{1-p_0}\right)}{n\log\left(\frac{p_0(1-p_a)}{p_a(1-p_0)}\right)}.$$

Proof.

$$\Pr(\sum Y_i = y) = \binom{n}{y} p^y (1-p)^{n-y}$$

Using the Neyman-Pearson lemma (let p_a be some particular value of $p > p_0$):

$$\frac{L(p_0)}{L(p_a)} = \frac{\binom{n}{y} p_0^y (1-p_0)^{n-y}}{\binom{n}{y} p_a^y (1-p_a)^{n-y}} < K$$

$$\left(\frac{p_0}{p_a}\right)^y \left(\frac{1-p_0}{1-p_a}\right)^n \left(\frac{1-p_0}{1-p_a}\right)^{-y} < K$$

$$\left(\frac{p_0(1-p_a)}{p_a(1-p_0)}\right)^y < K\left(\frac{1-p_a}{1-p_0}\right)^n$$

$$y \log\left(\frac{p_0(1-p_a)}{p_a(1-p_0)}\right) < \log(K) + n\log\left(\frac{1-p_a}{1-p_0}\right)$$

Aside:

$$\frac{p_0(1-p_a)}{p_a(1-p_0)} = \frac{p_0 - p_0 p_a}{p_a - p_0 p_a} < 1$$

since by assumption $p_a > p_0$. Therefore $\log\left(\frac{p_0(1-p_a)}{p_a(1-p_0)}\right) < 0$. So we have

$$\frac{y}{n} = \frac{1}{n}\sum Y_i > \frac{\log(K) + n\log\left(\frac{1-p_a}{1-p_0}\right)}{n\log\left(\frac{p_0(1-p_a)}{p_a(1-p_0)}\right)}$$

as the form for our critical region.

□

1.8.3 Composite Hypothesis Testing

We will start with one-dimensional family (one parameter), then expand to multiple parameters using tricks based on conditioning and using the principles of sufficiency and completeness.

Definition 1.8.5 (**Composite Hypothesis Test**)**.** Assume we have a statistical model $\{P_\theta, \theta \in \Theta\}$. Let $\Theta_0, \Theta_1 \subset \Theta$ be such that $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$. Assume $X \sim P_\theta$ for some $\theta \in \Theta$ (where X may be multivariate). We would like to test the hypotheses

$$H_0 : \theta \in \Theta_0, \qquad H_a : \theta \in \Theta_1.$$

Such a test is a *composite hypothesis test.*

The Neyman-Pearson Lemma (Lemma 1.8.1.1) does not directly apply to composite hypothesis tests, but it turns out we can use the Neyman-Pearson lemma to prove some results about some composite hypothesis tests under some circumstances.

Definition 1.8.6 (**Power function; definition 8.3.1 in Casella and Berger [2001]**). The *power function* of a test ϕ is defined as

$$\beta_\phi(\theta) := \mathbb{E}_\theta \phi(X).$$

Remark 37. For $\theta \in \Theta_0$, $\beta_\phi(\theta)$ is the probability of a Type 1 error; for $\theta \in \Theta_1$, $\beta_\phi(\theta)$ is the power of ϕ.

Definition 1.8.7 (**Test size; definition 8.3.6 in Casella and Berger [2001]**). The test ϕ is of *size* α if and only if

$$\sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \alpha.$$

Definition 1.8.8 (**Uniformly most powerful test; definition 8.3.11 in Casella and Berger [2001]**). ϕ^* is the *uniformly most powerful test* of size α if ϕ^* is of size α and $\beta_{\phi^*}(\theta) \geq \beta_\phi(\theta)$ for any other test ϕ of size α for all $\theta \in \Theta_1$.

We will be interested in families with *monotone likelihood ratios*. (Note: for the rest of the class today, $\Theta = [a, b] \subseteq \mathbb{R}$ with $a, b \in [-\infty, \infty]$).

Definition 1.8.9 (**Monotone likelihood ratio; definition 8.3.16 in Casella and Berger [2001]**). Let $T(X)$ be a statistic. Consider $\Theta \subseteq \mathbb{R}$ as a segment on the real line. We say that the family of pdfs $\{p_\theta, \theta \in \Theta\}$ has a *monotone likelihood ratio (MLR)* with respect to $T(x)$ if and only if for any $\theta' > \theta''$,

$$\frac{p_{\theta'}(x)}{p_{\theta''}(x)} = \psi_{\theta', \theta''}(T(x))$$

where $\psi_{\theta', \theta''}(\cdot)$ is non-decreasing. (That it is non-decreasing is without loss of generality; in the case that ψ is non-increasing, simply replace $T(x)$ with $-T(x)$ to achieve a non-decreasing ψ).

Example 1.8.1. Let $X \sim \text{Poisson}(\lambda)$, so

$$p_\lambda(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x \in \mathbb{Z}_+.$$

Then $\{\text{Poisson}(\lambda), \lambda \in \mathbb{R}_+\}$ has a monotone likelihood ratio with respect to $T(x) = x$ by the following argument. Take $\lambda' > \lambda''$. Then

$$\frac{e^{-\lambda'}(\lambda')^x x!}{x! e^{-\lambda''}(\lambda'')^x} = e^{-(\lambda' - \lambda'')} \left(\frac{\lambda'}{\lambda''}\right)^x$$

and

$$\psi_{\lambda', \lambda''}(x) = e^{-(\lambda' - \lambda'')} \left(\frac{\lambda'}{\lambda''}\right)^x$$

is non-decreasing.

Theorem 1.8.3.1 (Karlin-Rubin; similar to Theorem 8.3.17 in Casella and Berger [2001], Theorem 3.4.1 in Lehmann and Romano [2005]). Assume $\{p_\theta, \theta \in [a,b]\}$ has a monotone likelihood ratio with respect to some sufficient statistic for $\theta \in \mathbb{R}$, $T(x)$. Suppose we want to test

$$H_0 : \theta \leq \theta_0, \qquad H_a : \theta > \theta_0.$$

Then there exists $c \geq 0$ and $\gamma \in [0,1]$ such that

$$\phi^*(x) = \begin{cases} 1, & T(x) > c \\ \gamma, & T(x) = c \\ 0, & T(x) < c \end{cases}$$

is the uniformly most powerful test of size α. (In other words, the UMP test will always be of this form.)

Proof. We will show that this is the uniformly most powerful test over the larger family of tests that satisfy $\mathbb{E}_{\theta_0}\phi(x) = \alpha$ ($\beta_\phi(\theta_0) \leq \alpha$), which contains tests of size α. So if we can find the UMP test for this larger family (and it is a test of size α), then that is also the UMP test for test of size α (satisfying $\sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \alpha$).

First note that we can always find c and γ such that

$$\mathbb{E}_{\theta_0}(\phi^*(x) = \alpha.$$

Define $F(c) := \mathbb{P}_{\theta_0}(\{x : T(x) \leq x\})$. Note that

$$\mathbb{E}_{\theta_0}\psi^*(x) = F(c) + \gamma[F(c) - F(c^+)]$$

where $F(c) - F(c^+)$ is the size of the jump. (The proof of this claim is an exercise.)

Next, take $\theta' > \theta_0$ and consider the simple hypothesis test

$$H_0' : \theta = \theta_0, \qquad H_a' : \theta = \theta'.$$

By the Neyman-Pearson Lemma (Lemma 1.8.1.1), the most powerful test for this is

$$\phi'(x) = \begin{cases} 1, & p_{\theta'}(x) > c' p_{\theta_0}(x) \\ \gamma', & p_{\theta'}(x) = c' p_{\theta_0}(x) \\ 0, & p_{\theta'}(x) < c' p_{\theta_0}(x) \end{cases} \iff \phi'(x) = \begin{cases} 1, & p_{\theta'}(x)/p_{\theta_0}(x) > c' \\ \gamma', & p_{\theta'}(x)/p_{\theta_0}(x) = c' \\ 0, & p_{\theta'}(x)/p_{\theta_0}(x) < c' \end{cases}$$

But since $p_{\theta'}(x)/p_{\theta_0}(x) = \psi(T(x))$, by the assumption of the monotone likelihood ratio property,

$$\frac{p_{\theta'}(x)}{p_{\theta_0}(x)} > c' \iff T(x) > \psi^{-1}(c').$$

Set $\psi^{-1}(c') = c$ and $\psi^{-1}(\gamma') = \gamma$ (using the values of γ and c from earlier in the proof). Then we have

$$\phi'(x) = \begin{cases} 1, & T(x) > c \\ \gamma', & T(x) = c \\ 0, & T(x) < c \end{cases}$$

as desired. (Moreover, c and γ are uniquely determined by $\mathbb{E}_{\theta_0}(\phi^*) = \alpha$. Note that θ' does not appear anywhere in this test. Since θ' was arbitrary, the proof is almost complete.

The final argument is to show that this test has size α. Namely, we must show

$$\sup_{\theta \leq \theta_0} \beta_{\phi^*}(\theta) = \sup_{\theta \leq \theta_0} \mathbb{E}_\theta \phi^*(x) \leq \alpha.$$

It is sufficient to show that $\beta_{\phi^*}(\theta)$ is monotone. Take $\theta_1 < \theta_2$. Then by the Neyman-Pearson Lemma (Lemma 1.8.1.1), we know that ϕ^* is the UMP test for testing

$$H_0'' : \theta = \theta_1 \qquad H_a'' : \theta = \theta_2$$

of size $\beta_{\phi^*}(\theta)$. Since it is the most powerful test, its power $\mathbb{E}_{\theta_2}\phi^*(x)$ is at least as powerful than the test which is identically equal to $\beta_{\phi^*}(\theta)$; that is, $\mathbb{E}_{\theta_2}(\phi^*(x) \geq \mathbb{E}_{\theta_1}(\phi^* * (x)$.

□

Example 1.8.2. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$. Test $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$. Find the UMP test.

Solution

Take $\mu_1 > \mu_2$. Then

$$\frac{p_{\mu_1}(x_1, \ldots, x_n)}{p_{\mu_2}(x_1, \ldots, x_n)} = \exp\left\{ -\frac{1}{2} \left[\sum_{i=1}^n (x_i - \mu_1)^2 - \sum_{i=1}^n (x_i - \mu_2)^2 \right] \right\}$$
$$= \exp\left\{ -\frac{1}{2} \left[n(\mu_1^2 - \mu_2^2) + (\mu_1 - \mu_2) \sum_{i=1}^n x_i \right] \right\}$$

This is an increasing function with respect to $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$. Therefore by Theorem 1.8.3.1, we have that the UMP test is

$$\phi(x_1, \ldots, x_n) = \begin{cases} 1, & \sum_{i=1}^n x_i \geq c_\alpha \\ 0, & \sum_{i=1}^n x_i < c_\alpha \end{cases} = \begin{cases} 1, & n^{-1/2} \sum_{i=1}^n x_i \geq c'_\alpha \\ 0, & n^{-1/2} \sum_{i=1}^n x_i < c'_\alpha \end{cases} \tag{1.41}$$

Note that

$$\mathbb{E}_{\mu=0}\phi(X_1,\ldots,X_n)=\alpha \iff \mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i\geq c'_\alpha\right)=\alpha$$

and $n^{-1/2}\sum_{i=1}^{n}X_i\sim\mathcal{N}(0,1)$. Therefore $c'_\alpha=z_{1-\alpha}$ (the $1-\alpha$ quantile of a standard Gaussian distribution). Substituting this into (1.41) defines the UMP test.

Remark 38. The Karlin-Rubin Theorem (Theorem 1.8.3.1) still applies if the inequalities in the hypothesis test are reversed.

Example 1.8.3. $X_1,\ldots,X_n\sim\mathcal{N}(\mu,1)$. Test $H_0:\mu\in[0,1]$ against $H_1:\mu<0\cup\mu>1$. Does there exist a UMP test?

Solution

No. Consider $H'_0:\mu\in[0,1], H'_a:\mu>1$ and $H''_0:\mu\in[0,1], H''_a:\mu<1$. If the UMP test for our test exists, it must be identical to the UMP tests for both of these tests (since it would have to be UMP over ever θ). By Theorem 1.8.3.1, the UMP test in the first case is

$$\phi'(x_1,\ldots,x_n)=\begin{cases}1, & \sum_{i=1}^{n}x_i\geq c\\ 0, & \sum_{i=1}^{n}x_i<c.\end{cases}$$

The UMP test in the second case is

$$\phi'(x_1,\ldots,x_n)=\begin{cases}1, & \sum_{i=1}^{n}x_i\leq c\\ 0, & \sum_{i=1}^{n}x_i>c.\end{cases}$$

Since these tests are not identical, there is no UMP test for our test.

Theorem 1.8.3.2 (Generalized Neyman-Pearson Lemma, Theorem 3.6.1 in Lehmann and Romano [2005]). Assume that $f_0,f_1,\ldots,f_n:\mathbb{R}^d\to\mathbb{R}$ such that

$$\int|f_j|\,d\mu<\infty,\qquad j\in\{0,\ldots,N\}.$$

1. We seek to maximize the $\int\phi f_0\,d\mu$ over all tests ϕ subject to $\int\phi f_j\,d\mu=\alpha_j, j\in\{1,\ldots,N\}$.
2. We seek to maximize the $\int\phi f_0\,d\mu$ over all tests ϕ subject to $\int\phi f_j\,d\mu\leq\alpha_j, j\in\{1,\ldots,N\}$.

Suppose there exists $k_1,\ldots,k_N$ such that

$$\phi^*=\begin{cases}1, & f_0(x)>k_1f_1(x)+\ldots+k_Nf_N(x)\\ 0, & f_0(x)<k_1f_1(x)+\ldots+k_Nf_N(x)\end{cases}$$

satisfies $\int\phi^*f_j\,d\mu=\alpha_j, j\in\{1,\ldots,N\}$. Then ϕ^* solves the first problem. Moreover, if $k_1,\ldots,k_N\geq 0$, then ϕ^* also solves the second problem.

Example 1.8.4 (Like Theorem 3.7.1 in Lehmann and Romano [2005]). $\{P_\theta, \theta \in \Theta\}$. $H_0 : \theta \in [\theta_1, \theta_2]$. $H_a : \theta \notin [\theta_1, \theta_2]$. We have

$$p_\theta(x) = \frac{1}{c(\theta)} e^{\theta T(x)}$$

(one-parameter exponential family). Find the UMP unbiased test.

Solution

For the test to be unbiased, the Type I error rate has to be less than or equal to α (for all $\theta \in [\theta_1, \theta_2]$) and the power has to be greater than or equal to α (for all θ in the rejection region). We will use the Generalized Neyman-Pearson Lemma (Theorem 1.8.3.2) to show the solution is of the form

$$\phi^*(x) = \begin{cases} 1, & T(x) < c_1 \text{ or } T(x) > c_2 \\ \gamma_1, & T(x) = c_1 \\ \gamma_2, & T(x) = c_2 \\ 0, & T(x) \in (c_1, c_2) \end{cases}$$

such that $\mathbb{E}_{\theta_1} \phi^*(X) = \mathbb{E}_{\theta_2} \phi^*(X) = \alpha$. By the Generalized Neyman-Pearson Lemma, we look for the solution in the form

$$\begin{aligned} \tilde{\phi}(x) = \begin{cases} 1, & p_{\theta'}(x) > k_1 p_{\theta_1}(x) + k_2 p_{\theta_2}(x) \\ 0, & p_{\theta'}(x) < k_1 p_{\theta_1}(x) + k_2 p_{\theta_2}(x). \end{cases} &= \begin{cases} 1, & \frac{1}{c(\theta')} e^{\theta' T(x)} > k_1 \frac{1}{c(\theta_1)} e^{\theta_1 T(x)} + k_2 \frac{1}{c(\theta_2)} e^{\theta_1 T(x)} \\ 0, & \frac{1}{c(\theta')} e^{\theta' T(x)} < k_1 \frac{1}{c(\theta_1)} e^{\theta_1 T(x)} + k_2 \frac{1}{c(\theta_2)} e^{\theta_1 T(x)}. \end{cases} \\ &= \begin{cases} 1, & 1 > \tilde{k}_1 e^{(\theta_1 - \theta')T(x)} + \tilde{k}_2 e^{(\theta_2 - \theta')T(x)} \\ 0, & 1 < \tilde{k}_1 e^{(\theta_1 - \theta')T(x)} + \tilde{k}_2 e^{(\theta_2 - \theta')T(x)}. \end{cases} \end{aligned} \tag{1.42}$$

Let's analyze the inequality $g(t) < 1$ where

$$g(t) = \tilde{k}_1 e^{a_1 t} + \tilde{k}_2 e^{a_2 t}$$

with $a_1 = \theta_1 - \theta' > 0, a_2 = \theta_2 - \theta' > 0, a_1 < a_2$. Consider the second case.

(a) $\tilde{k}_1 > 0, \tilde{k}_2 > 0 \implies g(t)$ is increasing. $g(t) < 1 \iff t < c$ for some $c \in \mathbb{R}$. From the Karlin-Rubin Theorem, in this case $\beta_{\tilde{\phi}}(\theta)$ is monotone. But this is impossible, as we require $\beta_{\tilde{\phi}}(\theta_1) = \beta_{\tilde{\phi}}(\theta_2)$.

(b) $\tilde{k}_1 < 0, \tilde{k}_2 < 0 \implies g(t) < 1$. Examining (1.42) shows that in this case $\tilde{\phi}(x) = 1$ for all x (so $\tilde{\phi}$ always rejects), so we can't have controlled Type I error.

(c) $\tilde{k}_1 < 0, \tilde{k}_2 > 0$. In this case, $g'(t) = \tilde{k}_1 a_1 e^{a_1 t} + \tilde{k}_2 a_2 e^{a_2 t}$. Then $g'(t) = 0$ has a unique solution (see Figure 1.2), and $g(t) < 1 \iff t \in (c_1, c_2)$ (rejection only happens in this bounded interval). But this can't be a UMP unbiased test because the power function of this test is less than a constant test. We should have high power as our test statistic goes off to infinity or negative infinity, but examining (1.42), we see that $\tilde{\phi}(x) = 0$ as $T(x) \to \infty$ or $-\infty$.

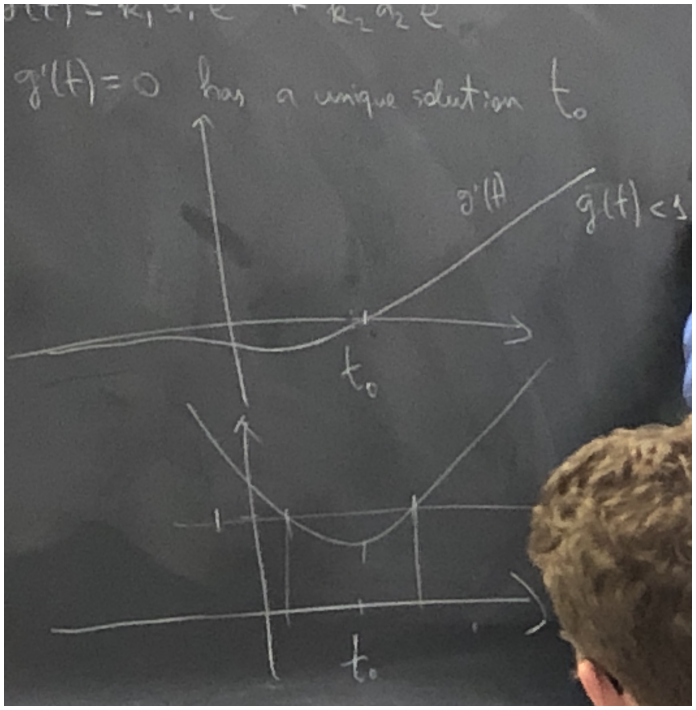

Figure 1.2: Figure for Example 1.8.4, case (c). This can't be the UMP unbiased test because our power is approaches 0 the further away we get from this narrow interval (we accept the null hypothesis when this function is more than 1).

(d) $\tilde{k}_1 > 0, \tilde{k}_2 < 0$. Then $g'(t)$ has a unique zero (see Figure 1.3), and $g(t) < 1 \iff t \in (c_1, c_2)$ which gives us the form of the test that we were looking for (we always accept the null hypothesis when this function is more than 1 and reject it when it is less than 1).

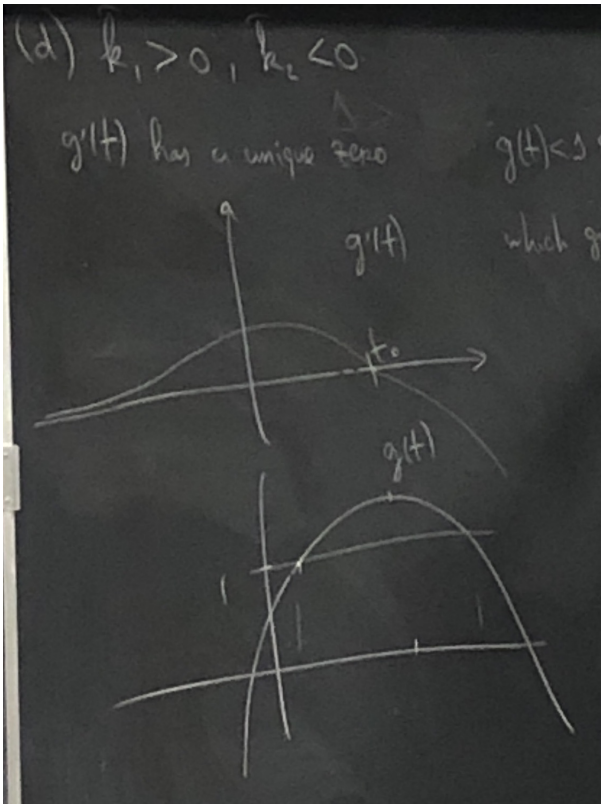

Figure 1.3: Second figure for Example 1.8.4, case (d). This is the form of the UMP unbiased test.

It remains to show that $\beta_{\phi^*}(\theta) \leq \alpha$ for all $\theta \in [\theta_1, \theta_2]$ (the power function appears as in Figure 1.4). We can show this in one of two ways.

(a) $\beta'_{\phi^*}(\theta) = 0$ has a unique solution (then by a similar argument to that used in parts (c) and (d) of the previous part, the power function will have the form we want).

(b) Lemma: the test ϕ^* minimizes the power at any $\theta' \in (\theta_1, \theta_2)$ among all SOB tests. (proof: exercise, similar mechanics to this).

Remark 39. SOB means "similar on the boundary." Recall that the definition of unbiasedness for a hypothesis test is that for a null hypothesis region Ω_H and an alternative hypothesis region Ω_k, we have

$$\mathbb{E}_{\theta_0}\phi(X) \leq \alpha \qquad \forall \theta_0 \in \Theta_H, \qquad \mathbb{E}_{\theta_a}\phi(X) \geq \alpha \qquad \forall \theta_a \in \Theta_K.$$

If the power function $\beta_\phi(\theta) := \mathbb{E}_\theta \phi(X)$ is a continuous function of θ, unbiasedness implies

$$\beta_\phi(\theta) = \alpha \; \forall \theta \in \Theta_B, \tag{1.43}$$

where Θ_B is the common boundary of Θ_H and Θ_K; that is, the set of points θ that are points or limit points of both Ω_H and Ω_K. See Section 4.1 of Lehmann and Romano [2005] and Lemma 1.8.5.2.

Definition 1.8.10 (**"Similar on the boundary" tests**)**.** Tests satisfying (1.43) are said to be **similar on the boundary**.

We can prove this in a way similar to Theorem 3.6.1 in Lehmann and Romano [2005], although the proof is a little different.

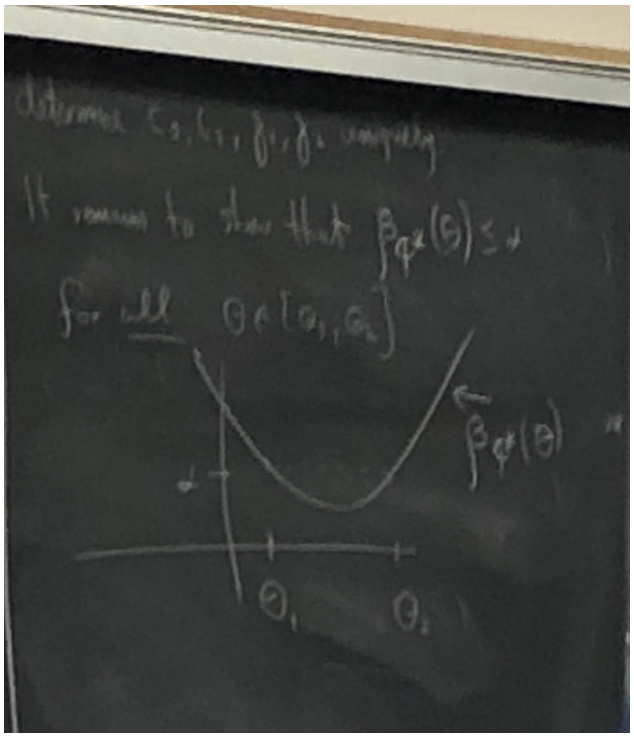

Figure 1.4: Third figure for Example 1.8.4, the desired form of a power function for our selected test.

Theorem 1.8.3.3 (**Theorem 3.7.1 in Lehmann and Romano [2005]**)**.** The most powerful unbiased test for a one-parameter exponential family is

$$\phi^*(x) = \begin{cases} 1, & T(x) < c_1 \text{ or } T(x) > c_2 \\ \gamma_1, & T(x) = c_1 \\ \gamma_2, & T(x) = c_2 \\ 0, & T(x) \in (c_1, c_2) \end{cases}$$

such that

(a) $\mathbb{E}_{\theta_0}(\phi^*(X) = \alpha$.

(b) $\frac{\mathrm{d}}{\mathrm{d}\theta}\beta_{\phi^*}(\theta)\big|_{\theta=\theta_0} = 0$ $(\mathbb{E}_{\theta_0}T(X)\phi^*(X) = \alpha\mathbb{E}_{\theta_0}T(X))$.

Proof of constraint (b). For any unbiased test ϕ of size α,

$$p_\theta(x) = \frac{1}{c(\theta)} e^{\theta T(x)} = \tilde{c}(\theta) e^{\theta T(x)}$$

(letting $\tilde{c}(\theta) = \frac{1}{c(\theta)}$). Then

$$\beta_\phi(\theta) = \int \phi(x) p_\theta(x) d\mu = \int \phi(x) \tilde{c}(\theta) e^{\theta T(x)} \, d\mu$$

so the derivative is

$$\begin{aligned}
\beta'_\phi(\theta) &= \int \phi(x)[\tilde{c}'(\theta) + T(x)\tilde{c}(\theta)]\tilde{c}(\theta) e^{\theta T(x)} \, d\mu \\
&= \frac{\tilde{c}'(\theta)}{\tilde{c}(\theta)} \int \phi(x)\tilde{c}(\theta) e^{\theta T(x)} \, d\mu + \int \phi(x) T(x) \tilde{c}(\theta) e^{\theta T(x)} \, d\mu = \frac{\tilde{c}'(\theta)}{\tilde{c}(\theta)} \mathbb{E}_\theta \phi(X) + \mathbb{E}_\theta \phi(X) T(X) \\
&= \frac{\tilde{c}'(\theta)}{\tilde{c}(\theta)} \alpha + \alpha \mathbb{E}_{\theta_0} T(x) = 0 \implies \frac{\tilde{c}'(\theta)}{\tilde{c}(\theta)} = -E_{\theta_0} T(X)
\end{aligned}$$

Hence,

$$\frac{\tilde{c}'(\theta)}{\tilde{c}(\theta)} \underbrace{\mathbb{E}_{\theta_0} \phi(X)}_{\alpha} + \mathbb{E}_{\theta_0} \phi(X) T(X) = 0$$

implies that $-\alpha \mathbb{E}_{\theta_0} T(X) + \mathbb{E}_{\theta_0} \phi(X) T(X) = 0$ so it follows that $\mathbb{E}_{\theta_0} T(X)\phi(X) = \alpha \mathbb{E}_{\theta_0} T(X)$.

$$\vdots$$

For any unbiased test ϕ, $\beta'_\phi(\theta) = 0$. (We know the test is unbiased: $\phi_\alpha = \alpha$.) Plug $\phi(x) = \alpha$ in to this equation to get

$$\frac{\tilde{c}'(\theta)}{\tilde{c}(\theta)} = -\mathbb{E}_{\theta_0} T(X).$$

□

1.8.4 Locally Most Powerful Tests

Definition 1.8.11 (Locally most powerful test). Assume that $\{\mathbb{P}_\theta, \theta \in \mathbb{R}\}$ is a statistical model, where p_θ is the probability density function corresponding to $\mathbb{P}_\theta$. Moreover, we will assume the model is "smooth;" that is, for any test Φ, the power function $\beta_\Phi(\theta)$ is differentiable and

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \beta_\Phi(\theta) = \int \frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(x) \cdot \Phi(x) \, dx. \tag{1.44}$$

Suppose we want to test $H_0 : \theta \leq \theta_0$ against $H_a : \theta > \theta_0$. A **locally most powerful** test of size α is defined as a solution to the optimization problem

$$\begin{aligned} &\underset{\Phi:\mathbb{R}\to[0,1]}{\arg\max} && \frac{\mathrm{d}}{\mathrm{d}\theta}\beta_\Phi(\theta_0) \\ &\text{subject to} && \beta_\Phi(\theta_0) = \alpha. \end{aligned} \tag{1.45}$$

The intuition of the locally most powerful test is that it maximizes the power in a small neighborhood of θ_0 by maximizing the slope of the power function at θ_0.

Proposition 1.8.4.1 (**Math 541B Homework 2 Problem 5**)**.** The solution to (1.45) always exists and is of the form

$$\Phi^*(x) = \begin{cases} 1, & \frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(x)\big|_{\theta=\theta_0} > kp_{\theta_0}(x), \\ \gamma, & \frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(x)\big|_{\theta=\theta_0} = kp_{\theta_0}(x), \\ 0, & \frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(x)\big|_{\theta=\theta_0} < kp_{\theta_0}(x), \end{cases}$$

where $\gamma \in [0,1]$ and k are determined by the size α of the test.

Proof. First we will show that the suggested test always exists and is a size α test under appropriate specifications for k and γ. Then we will show that under the Generalized Neyman-Pearson Lemma (Theorem 1.8.3.2), this implies that the suggested test is the solution to (1.45). Consider the test

$$\phi^*(x) = \begin{cases} 1, & \frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(x)\big|_{\theta=\theta_0} > kp_{\theta_0}(x) \\ \gamma, & \frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(x)\big|_{\theta=\theta_0} = kp_{\theta_0}(x) \\ 0, & \frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(x)\big|_{\theta=\theta_0} < kp_{\theta_0}(x) \end{cases} \tag{1.46}$$

with $\gamma \in [0,1]$ and $k \in \mathbb{R}$. Observe that

$$\begin{aligned} \mathbb{E}_{\theta_0}\phi^*(X) &= \mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0} > kp_{\theta_0}(X)\right) + \gamma\mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0} = kp_{\theta_0}(X)\right) \\ &= \mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0} \Big/ \, p_{\theta_0}(X) > k\right) + \gamma\mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0} \Big/ \, p_{\theta_0}(X) = k\right), \end{aligned} \tag{1.47}$$

where division by $p_{\theta_0}(X)$ is permissible because we are evaluating this expectation over $\mathbb{P}_{\theta_0}$ so we only need to include x such that $p_{\theta_0}(x) > 0$[1]. Define

$$k := \inf_{c\in\mathbb{R}}\left\{c : \mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0} > cp_{\theta_0}(X)\right) \leq \alpha\right\} = \inf_{c\in\mathbb{R}}\left\{c : \mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0} \Big/ \, p_{\theta_0}(X) > c\right) \leq \alpha\right\}. \tag{1.48}$$

[1] Another way of thinking about this is that the test will not reject for any x such that $p_{\theta_0}(x) = 0$ since with probability 1 these values of x will not be evaluated by the test in the first place.

Suppose that

$$\mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0}\Big/\, p_{\theta_0}(X)=k\right)=0. \tag{1.49}$$

Then we do not need to find γ because ϕ^* is already fully defined by this $c(\alpha)$: that is, it holds that the minimum of the set in (1.48) exists and is the infimum, so

$$\mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0}\Big/\, p_{\theta_0}(X)>k\right)=\alpha$$

and we are done with an arbitrary γ, say $\gamma=0$. The case where (1.49) does not hold is more complicated. Let

$$a(m):=\mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0}>mp_{\theta_0}(X)\right)=\mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0}\Big/\, p_{\theta_0}(X)>m\right).$$

Define the random variable

$$Y:=\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0}\Big/\, p_{\theta_0}(X)$$

Then since $a(m)$ is the probability that Y exceeds m under $\mathbb{P}_{\theta_0}$, $1-a(m)$ can be considered as the cdf of Y as a function of m, with $\lim_{m\to-\infty}(1-a(m))=0$ and $\lim_{m\to\infty}(1-a(m))=1$. This means that $1-a(m)$ is nondecreasing right-continuous, so $a(m)$ is nonincreasing and right-continuous. As a consequence,

$$\mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0}\Big/\, p_{\theta_0}(X)=k\right)=[1-a(k)]-\lim_{y\to k^-}[1-a(y)]=\lim_{y\to k^-}a(y)-a(k)=:a(k^-)-a(k),$$

(where the last step simply introduces a simpler notation) and we have

$$\mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0}\Big/\, p_{\theta_0}(X)=k\right)\neq 0 \iff a(k^-)-a(k)\neq 0. \tag{1.50}$$

If (1.50) holds,

$$\mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X)\Big|_{\theta=\theta_0}\Big/\, p_{\theta_0}(X)>k\right)=a(k)<\alpha.$$

Therefore in general (without any assumptions about whether $\mathbb{P}_{\theta_0}(Y=k)=0$) the size of ϕ as defined by setting (1.47) equal to α is

$$\mathbb{E}_{\theta_0}(\phi^*(X)) = \mathbb{P}_{\theta_0}(Y > k) + \gamma \mathbb{P}_{\theta_0}(Y = k) = a(k) + \gamma \left[a(k^-) - a(k)\right].$$

Because of this, it is clear that defining

$$\begin{aligned}\gamma := \begin{cases} \frac{\alpha - a(k)}{a(k^-) - a(k)}, & \mathbb{P}_{\theta_0}(Y = k) \neq 0 \\ 0, & \mathbb{P}_{\theta_0}(Y = k) = 0 \end{cases} &= \begin{cases} \frac{\alpha - a(k)}{a(k^-) - a(k)}, & a(k^-) - a(k) \neq 0 \\ 0, & a(k^-) - a(k) = 0 \end{cases} \\ &= \frac{\alpha - \mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(X)\big|_{\theta=\theta_0} \Big/ \, p_{\theta_0}(X) > k \right)}{\mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(X)\big|_{\theta=\theta_0} \Big/ \, p_{\theta_0}(X) > k^- \right) - \mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(X)\big|_{\theta=\theta_0} \Big/ \, p_{\theta_0}(X) > k \right)} \end{aligned} \tag{1.51}$$

when $\mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(X)\big|_{\theta=\theta_0} \Big/ \, p_{\theta_0}(X) > k^- \right) - \mathbb{P}_{\theta_0}\left(\frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(X)\big|_{\theta=\theta_0} \Big/ \, p_{\theta_0}(X) > k \right) \neq 0$ and 0 otherwise leads to (1.47) equaling α.

So we have that the test defined in (1.46) with k as defined in (1.48) and γ as defined in (1.51) always exists and is a size α test. Finally, note that

$$\phi^*(x) p_{\theta_0}(x) = \begin{cases} p_{\theta_0}(x), & \frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(x)\big|_{\theta=\theta_0} > k p_{\theta_0}(x) \\ \gamma p_{\theta_0}(x), & \frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(x)\big|_{\theta=\theta_0} = k p_{\theta_0}(x) \\ 0, & \frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(x)\big|_{\theta=\theta_0} < k p_{\theta_0}(x) \end{cases}$$

Therefore under this test, the problem we are hoping to solve (1.45) can be expressed as

$$\begin{aligned} \underset{\phi:\mathbb{R}\to[0,1]}{\text{maximize}} \quad & \frac{\mathrm{d}}{\mathrm{d}\theta}\beta_\phi(\theta) \\ \text{subject to} \quad & \mathbb{E}_{\theta_0}\phi(X) = \alpha \end{aligned} = \begin{aligned} \underset{\phi:\mathbb{R}\to[0,1]}{\text{maximize}} \quad & \int \frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(x) \cdot \phi(x) \, dx \\ \text{subject to} \quad & \int_{-\infty}^{\infty} \phi(x) p_{\theta_0}(x) \, d\mu = \alpha, \end{aligned} \tag{1.52}$$

where we used (1.44). Therefore, by the generalized Neyman-Pearson Lemma (Theorem 1.8.3.2), (1.46) is the solution to (1.52) (and, equivalently, (1.45)).

□

1.8.5 Similarity and Completeness (Section 4.3 of Lehmann and Romano [2005])

Example 1.8.5. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$. Find the UMP unbiased test for $H_0 : \sigma^2 = \sigma_0^2$, $H_a : \sigma^2 \neq \sigma_0^2$. Assuming that n is large, find the approximate values $c_1(\alpha)$, $c_2(\alpha)$.

Before solving this we will prove a lemma that we have been using implicitly.

Lemma 1.8.5.1. Suppose that $T(X)$ is sufficient for $\{p_\theta, \theta \in \Theta\}$, and let ϕ be a test. Denote

$$\psi(X) := \mathbb{E}_\theta[\phi(X) \mid T(X)].$$

Then

(a) $\psi(X)$ is a test.

(b) $\mathbb{E}_\theta \psi(X) = \mathbb{E}_\theta \phi(X)$.

Proof. (a) Since $T(X)$ is sufficient, the expectation over θ does not matter because we are conditioning on $T(X)$ anyway. That is, the distribution of $\phi(X)$ conditional on $T(X)$ does not depend on θ. Therefore the expectation over θ does not either. (A test takes values between 0 and 1 and doesn't depend on θ, only depends on the data.)

(b)

$$\mathbb{E}_\theta \psi(X) = \mathbb{E}_\theta \left(\mathbb{E}_\theta [\phi(X) \mid T(X)] \right) = \mathbb{E}_\theta \phi(X).$$

□

Lemma 1.8.5.2 (**Lemma 4.1.1 in Lehmann and Romano [2005]**)**.** If the distributions P_θ are such that the power function of every test is continuous, and if ϕ_0 is UMP among all tests satisfying (1.43) (restated here:)

$$\beta_\phi(\theta) = \alpha \qquad \forall \theta \in \omega \tag{1.53}$$

(where ω is the common boundary of the rejection and acceptance regions; that is, the set of points θ that are points or limit points of both regions; often called the boundary family) and is a level α test of H, then ϕ_0 is UMP unbiased. **(short version: if the power function of every test is continuous, then the UMP test among the class of similar on the boundary (SOB) tests of size α is UMP unbiased of size α.)**

Proof. For distributions such that the power function of every test is continuous, the class of SOB tests of size α contains the class of unbiased tests (since being SOB of size α is necessary to be unbiased if the power function of every test is continuous). Therefore the UMP test among the class of SOB tests ϕ^* is at least as powerful as any unbiased test. Further, ϕ^* is unbiased because it is uniformly at least as powerful as $\phi(x) := \alpha$ (since this is an SOB test).

□

Remark 40. By Theorem 2.7.1 in Lehmann and Romano [2005], the power function for a (one-parameter) exponential family is continuous and differentiable.

Definition 1.8.12 (**boundedly complete**)**.** $T(X)$ is **boundedly complete** if and only if for any function G such that $|G| \leq 1$ (bounded),

$$\{\mathbb{E}_\theta G(T) = 0 \qquad \forall \theta \in \Theta\} \implies G(T) = 0 \text{ a.e.}$$

So if a statistic is complete, it is boundedly complete (if it is complete, the equation holds for all functions G, not just bounded ones). Example:

$$X \in \{-1, 0, 1, 2, \ldots\}, \qquad \mathbb{P}_\theta(X = k) = \begin{cases} \theta, & k = -1 \\ (1-\theta)^2\theta^k, & k \in \{0, 1, \ldots\} \end{cases}$$

Then $T(X) = X$ is boundedly complete but not complete.

We want to be able to work with distributions that are more than one dimensional. We can do this by reducing the family to a one-parameter family. We do this in the following way, with tests with Neyman structure. Assume we have $X \sim P_\theta, \theta \in \Theta$. The problem is $H_0 : \theta \in \Theta_0, H_a : \theta \in \Theta_a$, where $\Theta_B := \overline{\Theta}_0 \cap \overline{\Theta}_a$ (the intersection of the closures of these sets).

Example: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, $H_0 : \mu \leq 0, H_a : \mu > 0$. Then if we plot $\mathbb{R}^2$ with μ as the horizontal axis and σ^2 as the vertical axis, the acceptance region is Quadrant II (including the top part of the vertical axis) and the rejection region is Quadrant I (not including the top part of the vertical axis). The boundary is the vertical axis ($\Theta_B = \{\mathcal{N}(0, \sigma^2), \sigma^2 > 0\}$).

Approach: take a complete sufficient statistic for the boundary family. We will condition on it and then find tests of size α.

Definition 1.8.13 (**Neyman Stucture**)**.** Let $T(X)$ be sufficient for the boundary family $\{p_\theta, \theta \in \Theta_B\}$. Test ϕ has **Neyman structure with respect to** T if and only if for some $\alpha \in [0, 1]$

$$\mathbb{E}_\theta[\phi(X) \mid T(X)] = \alpha, \qquad \forall \theta \in \Theta_B.$$

Note that ϕ has Neyman structure implies it is SOB of size α (SOB means "similar on the boundary:" the power function is essentially constant on the boundary. See Definition 1.8.10 and section 4.1 of Lehmann and Romano [2005].).

Theorem 1.8.5.3. Suppose T is a sufficient statistic for the boundary family of a SOB test ϕ of size α. Then ϕ has Neyman structure if and only if T is boundedly complete.

Proof. First we will prove this direction: suppose that T is sufficient and boundedly complete. Because ϕ is similar on the boundary (SOB), we have

$$\mathbb{E}_\theta \phi(X) = \alpha \qquad \forall \theta \in \Theta_B.$$

But then (letting $G(T) := \mathbb{E}_\theta[\phi(X) \mid T]$ and using what we know about T)

$$\begin{aligned} \alpha = \mathbb{E}_\theta \phi(X) = \mathbb{E}_\theta \mathbb{E}_\theta[\phi(X) \mid T(X)] = \mathbb{E}_\theta G(T) &\implies \mathbb{E}_\theta(G(T) - \alpha) = 0, \qquad \forall \theta \in \Theta_B \\ &\implies G(T) - \alpha = 0 \text{ a.s.} \qquad \implies G(T) = \alpha \text{ a.s.} \qquad \forall \theta \in \Theta_B \end{aligned}$$

where the last line followed from the bounded completeness of G. (Recall from Definition 1.8.13 that ϕ has Neyman structure precisely if $\mathbb{E}_\theta[\phi(X) \mid T] = \alpha$ for all $\theta \in \Theta_B$. Note that we only need bounded completeness since we must have $0 \leq |G| \leq 1 \forall T$.)

Next, assume an SOB test ϕ of size α has Neyman structure; that is, for a sufficient statistic T for the boundary family,

$$\mathbb{E}_\theta[\phi(X) \mid T] = \alpha \qquad \forall \theta \in \Theta_B.$$

We will prove by contradiction that then T must be boundedly complete. Suppose T is not boundedly complete. Then there must exist some function ψ such that $|\psi| \leq 1$ and $\mathbb{E}_\theta \psi(T) = 0 \; \forall \theta \in \Theta_B$ but $\psi(T)$ is not identically 0. Define $\phi := \alpha + c\psi(T)$, where we choose $c \neq 0$ such that $0 \leq |\phi| \leq 1$ so that ϕ is a test. Note that $\mathbb{E}_\theta \phi = \alpha$ for all $\theta \in \Theta_B$, so ϕ is SOB of size α. But

$$\mathbb{E}(\phi(x) \mid T(x)) = \alpha + c\psi(T(x)) \neq \alpha.$$

Therefore by contradiction, we have that T is boundedly complete.

□

Remark 41. We already know from Lemma 1.8.5.2 that if the power function is continuous for all tests (e.g. one parameter exponential family), then the UMP SOB test of size α ϕ^* is the UMP unbiased test of size α. That is, the UMP unbiased test of size α is the solution to the equation

$$\begin{aligned} &\underset{\phi:\mathbb{R}\to[0,1]}{\arg\max} && \mathbb{E}_\theta(\phi \mid T) && \forall \theta \in \Theta_A \\ &\text{subject to} && \beta_\phi(\theta) = \alpha && \forall \theta \in \Theta_B \end{aligned} \tag{1.54}$$

where T is a sufficient statistic for the boundary family, Θ_A is the rejection region, and Θ_B is the boundary family.

Under Theorem 1.8.5.3, if T is boundedly complete, then any test $\tilde{\phi}$ is an SOB test of size α if and only if it has Neyman structure. So if the power function is continuous for all tests (e.g. one parameter exponential family) and T is a boundedly complete sufficient statistic for the boundary family, then the UMP test in the class of tests with Neyman structure is the UMP unbiased test. That is, the UMP unbiased test is the solution to (1.54).

Theorem 1.8.5.4. Assume that $X \sim P_\theta$ where

$$p_\theta(x) = \frac{1}{c(\theta)} h(x) \cdot \exp\left\{ \sum_{j=1}^{k} \theta_j T_j(x) \right\}$$

(canonical exponential family). Then $T(X) = (T_1(X), \ldots, T_k(X))$ is a sufficient statistic for the family $\{p_\theta, \theta \in \Theta\}$ provided that Θ has non-empty interior. (See also Proposition 1.3.2.4.)

See also Theorem 1.3.4.3 for complete statistics in the exponential family.

Conditioning method: assume that $\beta_\phi(\theta)$ is continuous for any test ϕ.

(1) The UMP unbiased test ϕ^* must be SOB (because as discussed in the proof of Lemma 1.8.5.2, SOB is necessary for unbiasedness if $\beta_\phi(\theta)$ is continuous for all tests).

(2) UMP SOB test of size α is UMP unbiased test of size α (by Lemma 1.8.5.2).

(3) Neyman structure implies that it suffices to find the test that maximizes the conditional power $\mathbb{E}_\theta(\phi \mid T)$ for all $\theta \in \Theta_a$ (see Remark 41).

Example 1.8.6. $X \sim \text{Poisson}(\mu)$, $Y \sim \text{Poisson}(\nu)$. $H_0 : \mu \leq \nu$, $H_a : \nu > \nu$. Find the UMP unbiased test.

Solution

$$p_{\mu,\nu}(x,y) = \frac{e^{-(\mu+\nu)}\mu^x\nu^y}{x!y!} = e^{-(\mu+\nu)}e^{x\log\mu + y\log\nu} \cdot \frac{1}{x!y!} = \frac{1}{x!y!}e^{-(\mu+\nu)}e^{x\log(\mu/\nu)+(x+y)\log\nu}$$

Let $\xi := \log(\mu/\nu), \eta = \log\nu$. Then $H_0 : \xi \leq 0$, $H_a : \xi > 0$. We can use an SOB thing where $\Theta_B = \{(\xi,\eta) : \eta = 0\}$ is the boundary family. Then $T(X,Y) = X + Y$ is complete sufficient for the boundary family $\{p_{\xi,\eta} : \xi = 0\}$ by Theorem 1.8.5.4 (Proposition 1.3.2.4). Let's condition everything on $T = X+Y$. We need to find $\mathbb{P}(X = x \mid T = t)$. Note that

$$\mathbb{P}_{\mu,\nu}(X = x \mid T = t) = \frac{\mathbb{P}_{\mu,\nu}(X = x, Y = t - x)}{\mathbb{P}_{\mu,\nu}(T = t)} = \frac{\mu^x}{x!}e^{-\mu} \cdot \frac{\nu^{t-x}}{(t-x)!}e^{-\nu} \cdot \frac{t!e^{\mu+\nu}}{(\mu+\nu)^t}$$

where we used $T \sim \text{Poisson}(\mu + \nu)$ (exercise; use characteristic functions)

$$= \binom{t}{x}\left(\frac{\mu}{\mu+\nu}\right)^x\left(\frac{\nu}{\mu+\nu}\right)^{t-x}$$

so $\{X \mid T = t\} \sim \text{Bin}(t, \mu/(\mu+\nu))$.

Note that $\mu \leq \nu \iff p \leq 1/2$, so for our new problem, we have $X \sim \text{Binom}(t,p)$, and we will test $H_0' : p \leq 1/2, H_a' : p > 1/2$. For this problem we can find the UMP test (not just the UMP unbiased test) because the binomial family has a monotone likelihood ratio with respect to $T(X) = X$ when $p \in (0,1)$ and t fixed (very straightforward to check).

Then

$$\phi(x) = \begin{cases} 1, & x > c(t) \\ \gamma, & x = c(t) \\ 0, & x < c(t) \end{cases}$$

is UMP. To find $c(t,\alpha)$, use

$$\mathbb{E}_{p=1/2}\phi(X) = \alpha \iff \sum_{x>c(t)}\binom{t}{x}(1/2)^t + \gamma\binom{t}{c(t)}(1/2)^{c(t)} = \alpha. \tag{1.55}$$

(hard to compute in practice; need a computer). So the test for the original problem is

$$\phi(X,Y) = \begin{cases} 1, & X > c(t) \\ \gamma, & x = c(t) \\ 0, x < c(t) \end{cases}$$

where $t = X + Y$ and $c(t)$ solves (1.55).

Example 1.8.7. Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, i.i.d. We want to test $H_0 : \sigma^2 \geq \sigma_0^2, H_a : \sigma^2 < \sigma_0^2$. Find the UMP unbiased test.

Solution

We have

$$p_{\mu,\sigma^2}(x_1, \dots, x_n) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(\frac{\mu}{\sigma^2} \sum_{j=1}^n x_j - \frac{1}{2\sigma^2} \sum_{j=1}^n x_j^2 \right) \cdot e^{-n\mu/(2\sigma^2)}.$$

Therefore $T = (\sum_{j=1}^n X_j, \sum_{j=1}^n X_2^2)$ is a complete sufficient statistic for (μ, σ^2). Let

$$\overline{X}_n := n^{-1} \sum_{j=1}^n X_j, \qquad S^2 := \frac{1}{n-1} \sum_{j=1}^n (X_j - \overline{X}_n)^2.$$

Then $(\overline{X}, S^2)$ is in one to one correspondence with T. (Why? Use

$$\frac{1}{n} \sum_{j=1}^n (y_j - \overline{y}_j)^2 = \frac{1}{n} \sum_{j=1}^n y_j^2 - (\overline{y}_n)^2$$

for any numbers $y_1, \dots, y_n$.) So $(\overline{X}_n, S^2)$ is also complete sufficient for (μ, σ^2), so we only need to consider tests that are functions of this complete sufficient statistic. The boundary family in this case Θ_B consists of all normal distributions with variance equal to σ_0^2; that is, all normal distributions $\mathcal{N}(\mu, \sigma_0^2)$, $\mu \in \mathbb{R}$. For this family, $\overline{X}$ is a complete sufficient statistic, so we can condition everything on $\overline{X}$. Hence, we need to find the conditional distribution of S^2 given $\overline{X}$. Since $\overline{X}$ and S^2 are independent by Basu's Theorem (Theorem 1.3.4.2; see Proposition 1.2.0.6 and example 1.3.8), the problem becomes

$$H_0 : \sigma^2 \geq \sigma_0^2; \qquad H_a, \sigma^2 < \sigma_0^2$$

based on our observation S^2. We know that

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

by Proposition 1.2.0.6. Let $\tilde{\sigma}^2 := \sigma^2/(n-1)$. Then the pdf of S^2 is

$$p_{\tilde{\sigma}^2}(x) = \left(\frac{1}{\tilde{\sigma}^2} \right)^{(n-1)/2} x^{(n-1)/2-1} e^{-x/(2\tilde{\sigma}^2)}$$

This means that

$$\frac{p_{\tilde{\sigma}_1^2}(x)}{p_{\tilde{\sigma}_2^2}(x)} = \psi_{\tilde{\sigma}_1^2,\tilde{\sigma}_2^2}(x)$$

is monotone, so $\tilde{\sigma}^2$ has a monotone likelihood ratio. Then by Karlin-Rubin (Theorem 1.8.3.1), the UMP unbiased test is

$$\phi^* = \begin{cases} 1, & S^2 \leq c \\ 0, & S^2 > c. \end{cases}$$

To find the test of size α, take $c := \frac{\sigma_0^2}{n-1}(w_\alpha^{(n-1)}$ where $w_\alpha^{(n-1)}$ is the $(1-\alpha)$ quantile of χ^2_{n-1}.

Remark 42. This may be UMP in general, although we can't prove it with current tools (??). So it is at least UMP unbiased.

Example 1.8.8. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, i.i.d. We want to test $H_0 : \mu \geq \mu_0, H_a : \mu < \mu_0$ or $H_0 : \mu = \mu_0, H_a : \mu \neq \mu_0$. Find the UMP unbiased tests for each of these cases.

Solution

Consider tests ϕ that depend on $T = (\overline{X}, \sum_{j=1}^n X_j^2)$. The boundary family is $\Theta_B = \{\mathcal{N}(\mu_0, \sigma^2), \sigma^2 > 0\}$. A complete sufficient statistic for Θ_B is then

$$S_{\mu_0}^2 := \sum_{j=1}^n (X_j - \mu_0)^2.$$

If we consider

$$T' = \left(\frac{\overline{X} - \mu_0}{\sqrt{S_{\mu_0}^2}}, S_{\mu_0}^2 \right),$$

note that it is one-to-one correspondence with T, so T' is complete sufficient as well. Observe that

(1)
$$\frac{\overline{X} - \mu_0}{\sqrt{S_{\mu_0}^2}} \perp\!\!\!\perp S_{\mu_0}^2$$

(Why? See proof of Proposition **??**.)

(2) For

$$W = \frac{\overline{X} - \mu_0}{\sqrt{S_{\mu_0}^2}}, \qquad T'' = \frac{(\overline{X} - \mu_0)}{\sqrt{S^2}}\sqrt{n},$$

we have

$$T'' = \frac{\sqrt{(n-1)n}}{\sqrt{1-nW^2}}W$$

(exercise).

Then

$$g(w) = \left| \frac{\sqrt{(n-1)n}}{\sqrt{1-nw^2}}w \right|$$

is a monotone function of w. Moreover, by the characterization of a Student's t distribution

$$T = \frac{Z}{\sqrt{V/v}} \implies T \sim t_v$$

where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2_v$, and $Z \perp\!\!\!\perp V$, T'' has, under the null hypothesis that $\mu = \mu_0$, a Student's t distribution with $n-1$ degrees of freedom. Therefore we can conclude that the UMP unbiased test has the form

$$\phi^* = \begin{cases} 1, & |T''| > c \\ 0, & |T''| < c \end{cases}$$

where we choose c to make the test have size α.

1.8.6 Permutation Tests (section 5.8 in Lehmann and Romano [2005], p. 188 of pdf)

Suppose we have $X_1, \dots, X_1 \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. and $Y_1, \dots, Y_m \sim \mathcal{N}(\eta, \sigma^2)$ i.i.d. We would like to test $H_0 : \mu = \eta$ against $H_a : \eta - \mu = \Delta > 0$.

Alternatively: $H_0 : X_1, \dots, X_n, Y_1, \dots, Y_m$ i.i.d. with density f, H_a : the joint density of $X_1, \dots, X_n, Y_1, \dots, Y_m$ has density $f(x_1) \cdots f(x_n) f(y_1 - \Delta) \cdots f(y_m - \Delta), \Delta > 0$.

We are looking for a SOB test of size α. Hence, a test ϕ must satisfy $\mathbb{E}_f \phi = \alpha$ for any f. Define $\mathbb{Z}_1 = X_1, \dots, \mathbb{Z}_n = X_n, \mathbb{Z}_{n+1} = Y_1, \dots, \mathbb{Z}_{n+m} = Y_m$.

Theorem 1.8.6.1. Consider the following statistical model: $\{p_f : f \text{ is density on } \mathbb{R} \text{ with respect to Lesbegue measure.}\}$ Suppose $X_1, \dots, X_n \sim p_f$ i.i.d. Then $(X_{(1)}, \dots, X_{(n)})$ is a complete sufficient statistic. (That is, for the problem we are considering, the order statistics $(\mathbb{Z}_{(1)}, \dots, \mathbb{Z}_{(n+m)})$ are complete sufficient under H_0.)

Proof. Sufficiency:

$$\mathbb{P}_f\left((X_1, \dots, X_n) \in A \mid X_{(1)}, \dots, X_{(n)}\right) = \frac{1}{n!} \sum_{\sigma \in S_n} I\left\{X_{\sigma(1)}, \dots, X_{\sigma(n)} \in A\right\}$$

which does not depend on any parameters in f, so this proves sufficiency. Completeness: consider the following parametric family

$$\mathbb{P}_{\Theta_1,\ldots,\Theta_n}(x_1,\ldots,x_n) = c(\Theta_1,\ldots,\Theta_n)\exp\left\{\theta_1\sum_{i=1}^n x_j + \theta_2\sum_{i=1}^n x_j^2 + \ldots + \theta_n\sum_{i=1}^n x_j^n - \sum_{i=1}^n x_i^{2n}\right\}.$$

From Theorem 1.3.4.3, we know that

$$Y = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2, \ldots, \sum_{i=1}^n x_i^n\right)$$

is complete. We will establish this intermediate fact: let

$$W := \left(\underbrace{\sum_{j=1}^n x_j}_{w_1}, \underbrace{\sum_{i<j} x_i x_j}_{w_2}, \sum_{i<j<k} x_i x_j x_k, \ldots, \underbrace{x_1 x_2 \cdots x_n}_{w_n}\right).$$

This are symmetric polynomials. We will show that the order statistics T are in one-to-one correspondence with W. Consider the polynomial $p(x) = \prod_{j=1}^n (x - x_j)$. (Of course this polynomial is invariant to permutation of the x_j.) Then the coefficients of the polynomial are given by

$$p(x) = \prod_{j=1}^n (x - x_j) = x^n - w_1 x^{n-1} + w_2 x^{n-2} - \ldots + (-1)^n w_n,$$

so $(x_{(1)}, \ldots x_{(n)})$ is in one-to-one correspondence with W. Finally, the fact that W is in one-to-one correspondence with V follows from Newton's Identities: you can establish by induction that

$$v_k - w_1 v_k + w_2 v_{k-2} - \ldots + (-1)^k w_{k-1} v_k + (-1)^k w_k = 0$$

for $k = 1, 2, \ldots, n$. For example, consider $k = 2$:

$$W = (x_1 + x_2, x_1 x_2), \qquad , V = (x_1 + x_2, x_1^2 + x_2^2)$$

can always write $x_1 x_2 = \left[(x_1 + x_2)^2 - (x_1^2 + x_2^2)\right]/2$, so knowing W is equivalent to knowing V. Newton's Identities induct on this and show that this is true for any power k.

$$\vdots$$

(See also Example 4.3.4 in Lehmann and Romano [2005].)

□

Now we will apply the conditioning method by conditioning on $T = (\mathbb{Z}_{(1)}, \dots, \mathbb{Z}_{(n+m)})$. The new problem is then

$$H_0' : (\mathbb{Z}_{(1)}, \dots, \mathbb{Z}_{(n+m)}) \sim f^{(0)}_{\mathbb{Z}|T} \text{ are uniform on order statistics}), \qquad H_a' : (\mathbb{Z}_{(1)}, \dots, \mathbb{Z}_{(n+m)}) \sim f^{(a)}_{\mathbb{Z}|T}$$

Under H_0, $\mathbb{P}_{H_0}\left((\mathbb{Z}_{(1)}, \dots, \mathbb{Z}_{(n+m)}) = (z_1, \dots, z_{n+m})\right) = 1/(n+m)!$ for any specific permutation $(z_1, \dots, z_{n+m})$. Let's fix a specific alternative and find the most powerful test against this alternative. Any alternative is of the form $(\mathbb{Z}_{(1)}, \dots, \mathbb{Z}_{(n+m)})$ has joint density h, where h is fixed.

so test is

$$H_0 : X_1, \dots, X_n, Y_1, \dots, Y_m \text{ are i.i.d.;} \qquad H_a : X_1, \dots, X_n, Y_1, \dots, Y_m \text{ have joint density } h$$

It turns out that conditional on the order statistics (so that the only randomness left is the way in which they are permuted), the density is given by (1.58).

Theorem 1.8.6.2 (**Theorem 5.8.1 in Lehmann and Romano [2005]**).

Lemma 1.8.6.3 (**See section 5.9 of Lehmann and Romano [2005], p. 189 of pdf**). Let $(\mathbb{Z}_1, \dots, \mathbb{Z}_{n+m})$ have joint density h. Then for any f,

$$\mathbb{E}\left[f(\mathbb{Z}_1, \dots, \mathbb{Z}_{n+m}) \mid T\right] = \frac{\sum_{\sigma \in S_{n+m}} f(\mathbb{Z}_{\sigma(1)}, \dots, \mathbb{Z}_{\sigma(m_n)}) h(\mathbb{Z}_{\sigma(1)}, \dots, \mathbb{Z}_{\sigma(m_n)})}{\sum_{\sigma \in S_{n+m}} h(\mathbb{Z}_{\sigma(1)}, \dots, \mathbb{Z})} \tag{1.56}$$

where $T = (\mathbb{Z}_{(1)}, \dots, \mathbb{Z}_{(n+m)})$ and S_{n+m} is a collection of all permutations of $\{1, \dots, n_m\}$.

Proof. Want to show that for any symmetric set $\mathcal{A}_0$ (meaning that $(z_1, \dots, z_{n_m}) \in \mathcal{A}_0 \iff (z_{\sigma(1)}, \dots, z_{\sigma(n_m)}) \in \mathcal{A}_0$),

$$\int_{\mathcal{A}_0} f(z_1, \dots, z_{n_m}) h(z_1, \dots, z_{n+m}) \, dz = \mathbb{E} f(\mathbb{Z}_1, \dots, \mathbb{Z}_{n+m}) I\left\{(\mathbb{Z}_1, \dots, \mathbb{Z}_{n+m}) \in \mathcal{A}_0\right\} \tag{1.57}$$

Trying to prove:

$$g(T) = \mathbb{E}[f(X) \mid T] \iff \forall h, \mathbb{E}[f(X)h(T)] = \mathbb{E}[g(T)h(T)].$$

it is always sufficient to prove this for indicator functions because you can approximate any function h as a sum of indicator functions (fact from measure theory). Any set that is measurable with respect to the permutation statistics must be measure theoretic; any set that belongs to the sigma algebra generated by the order statistics must have the property that any permutation is in $\mathcal{A}_0$.

We want to show that (1.57) equals

$$\mathbb{E} f_0(\mathbb{Z}_{(1)}, \dots, \mathbb{Z}_{(n+m)}) I\left\{(\mathbb{Z}_{(1)}, \dots, \mathbb{Z}_{(n+m)}) \in \mathcal{A}_0\right\}$$

where f_0 is given by the right side of (1.56). By permutation invariance, we have

$$\int f_0(z_1, \dots, z_{m+n} h(z_{\sigma(1)}, \dots z_{\sigma(m+n)}) dz = \text{ constant}$$

So we have $(n+m)!$ values that are all equal. So we can write

$$\begin{aligned}
&\int_{\mathcal{A}_0} f_0(z_1, \dots, z_{n+m}) h(z_1, \dots, z_{n+m}) \, dz \\
&\qquad = \int_{\mathcal{A}_0} f_0(z_1, \dots, z_{n+m}) \frac{1}{(n+m)!} \sum_{\sigma \in S_{n+m}} h(z_{\sigma(1)}, \dots, z_{\sigma(n+m)}) \, dz \\
&\qquad = \frac{1}{(n+m)!} \int_{\mathcal{A}_0} \sum_{\sigma \in S_{n+m}} f_0(z_{\sigma(1)}, \dots, z_{\sigma(n+m)}) h(z_{\sigma(1)}, \dots, z_{\sigma(n+m)}) \, dz \\
&\qquad\qquad = \int_{\mathcal{A}_0} f_0(z_{\sigma(1)}, \dots, z_{\sigma(n+m)}) h(z_{\sigma(1)}, \dots, z_{\sigma(n+m)}) \, dz
\end{aligned}$$

since $\mathcal{A}_0$ is (by assumption) permutation symmetric. So this holds for any permutation. In particular, it holds for the identity permutation, so we can write

$$= \int_{\mathcal{A}_0} f_0(z_1, \dots, z_{n+m}) h(z_{\sigma(1)}, \dots, z_{\sigma(n+m)}) \, dz$$

We are interested in $f^{(a)}_{\mathbb{Z}|T}$. Take $f = I\{\mathbb{Z}_1 = z_1, \dots, \mathbb{Z}_{n+m} = z_{n+m}\}$ for a specific permutation $(z_1, \dots, z_{n+m})$ of $(\mathbb{Z}_{(1)}, \dots, \mathbb{Z}_{(n+m)})$. Then

$$f^{(a)}_{\mathbb{Z}|T} = \frac{h(z_1, \dots, z_{n+m})}{\sum_{\sigma \in S_{n+m}} h(z_{\sigma(1)}, \dots, z_{\sigma(n+m)})} \tag{1.58}$$

□

Remark 43. Note that in case where alternative hypothesis is one particular ordering, (1.56) reduces to $1/(n+m)!$ since f is an indicator variable for the particular permutation and the denominator is a sum of the number of all of the permutations.

(next time: apply Neyman-Pearson lemma to (1.58) to show how we finally get the UMP test.) The Neyman-Pearson test:

$$\phi = \begin{cases} 1, & h^{(a)}_{\mathbb{Z}|T}(x_1, \dots, x_n, y_1, \dots, y_m) \Big/ h^{(0)}_{\mathbb{Z}|T}(x_1, \dots, x_n, y_1, \dots, y_m) > c(T) \\ \gamma, & h^{(a)}_{\mathbb{Z}|T}(x_1, \dots, x_n, y_1, \dots, y_m) \Big/ h^{(0)}_{\mathbb{Z}|T}(x_1, \dots, x_n, y_1, \dots, y_m) = c(T) \\ 0, & h^{(a)}_{\mathbb{Z}|T}(x_1, \dots, x_n, y_1, \dots, y_m) \Big/ h^{(0)}_{\mathbb{Z}|T}(x_1, \dots, x_n, y_1, \dots, y_m) < c(T) \end{cases} \tag{1.59}$$

(using data in the specific order, seeing if the likelihood of the data in this specific order is large). Note that the denominator of (1.58) is a function of T (the order statistics) so it can be absorbed into $c(T)$. Therefore this test (1.59) is equivalent to the simpler test

$$\phi = \begin{cases} 1, & h^{(a)}_{\mathbb{Z}|T}(z_1, \dots, z_{n+m}) > c'(T) \\ \gamma, & h^{(a)}_{\mathbb{Z}|T}(z_1, \dots, z_{n+m}) = c'(T) \\ 0, & h^{(a)}_{\mathbb{Z}|T}(z_1, \dots, z_{n+m}) < c'(T) \end{cases} \tag{1.60}$$

We can control the Type I error rate by setting

$$\mathbb{E}_{H_0}(\phi) = \alpha \iff \alpha = \frac{1}{(n+m)!} \sum_{\sigma \in S_{n+m}} I\left\{h(\mathbb{Z}_{\sigma(1)}, \dots, \mathbb{Z}_{\sigma(n_m)}) > C\right\} + \frac{1}{(n+m)!}\gamma \sum_{\sigma \in S_{n+m}} I\left\{h(\mathbb{Z}_{\sigma(1)}, \dots, \mathbb{Z}_{\sigma(n_m)}) = C\right\} \tag{1.61}$$

Example 1.8.9. Testing hypothesis that treatment is different; assume treatment is random so groups are homogeneous. $H_0 : X_1, \dots, X_n, Y_1, \dots Y_m$ are i.i.d.; $H_a : X_1, \dots, X_n \sim \mathcal{N}(\mu_1, \sigma^2)$ i.i.d., $Y_1, \dots, Y_m \sim \mathcal{N}(\mu_2, \sigma^2)$ i.i.d. Find the permutation test for the problem.

Solution

test rejects when this quantity is large

$$\begin{aligned} f(x_1, \dots, x_n, y_1, \dots, y_m) &= \frac{1}{(2\pi\sigma^2)^{(n+m)/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu_1)^2\right\} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \mu_2)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{(n+m)/2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i - \mu_1)^2 + \sum_{i=1}^{m}(y_i - \mu_2)^2\right]\right\} \end{aligned}$$

note that this is monotone. So test rejects when the sum of squares

$$\begin{aligned} \sum_{i=1}^{n}(x_i - \mu_1)^2 + \sum_{i=1}^{m}(y_i - \mu_2)^2 &= \sum_{i=1}^{n} x_i^2 - 2\mu_1 \sum_{i=1}^{n} x_i + n\mu_1^2 + \sum_{j=1}^{m} y_j^2 - 2\mu_2 \sum_{j=1}^{m} y_j + n\mu_2 \\ &= \sum_{i=1}^{n} x_i^2 + \sum_{j=1}^{m} y_j^2 + n(\mu_1^2 + \mu_2^2) - \underbrace{2\left[\mu_1 \sum_{i=1}^{n} x_i + \mu_2 \sum_{j=1}^{m} y_j\right]}_{\text{only part that depends on specific ordering}} \end{aligned} \tag{1.62}$$

is small. We can take everything that depends on the sufficient statistic (the order statistics) and shift it to the constant in the hypothesis test. (for example, the sum of squares of all of the numbers depends on the order statistic only, and so does the sum of the constant terms. so everything that does not depend on the specific order of the data can be subsumed into the constant term. So we only need the cross terms in the binomial expansion, and then we determine c' by the solution to (1.61).) Note that the sum of squares is small if and only if

$$\mu_1 \sum_{i=1}^{n} x_i + \mu_2 \sum_{j=1}^{m} y_j$$

is large, if and only if

$$\underbrace{(\mu_1 - \mu_2)}_{>0} \sum_{i=1}^{n} x_i + \underbrace{\mu_2 \left(\sum_{i=1}^{n} x_i + \sum_{j=1}^{m} y_j \right)}_{\text{subsumed into } c \text{ since depends only on order statistics}}$$

is large, if and only if $\sum_{i=1}^{n} x_i$ is large. So we can write the test (1.60) as simply

$$\phi = \begin{cases} 1, & \sum_{i=1}^{n} x_i > c(T) \\ \gamma, & \sum_{i=1}^{n} x_i = c(t) \\ 0, & \sum_{i=1}^{n} x_i < c(T) \end{cases}$$

For example, assume $n = m = 2$. Suppose $x_1 = 4, x_2 = 1, y_1 = 7, y_2 = 3$. Consider all possible permutations, but we don't care about switching xs or y (test statistic is invariant to permutations of xs only), so that's $\binom{4}{2} = 6$ possible orderings. The orderings are given in Table 1.1. Then the test will reject if $X_1 + X_2 \geq 11$. Corresponds to situation when large values in begninning are unusuallly high; if so, unlikely data are i.i.d. because if that were true then the large and small values would be distributed more uniformly. Then the size of this test would be $\alpha = 1/6$. (If you want smaller α then you need to use a randomized test. But obviously in real life that makes no sense.)

Table 1.1: Table for Example 1.8.9 in case of $n = m = 2$.

1	3	4	7	$\sum_{i=1}^{2} X_i$
X	X	Y	Y	4
X	Y	X	Y	5
X	Y	Y	X	8
Y	X	X	Y	7
Y	X	Y	X	10
Y	Y	X	X	11

1.8.7 Invariance in Testing (Chapter 6 of Lehmann and Romano [2005])

Example 1.8.10. $X_1, \ldots, X_n \sim \mathcal{N}(\theta_j, 1), j \in [n]$, independent. Consider testing $H_0 : \theta_1 = \theta_2 = \ldots = \theta_n = 0$ against $H_1 : \exists j : \theta_j \neq 0$. That is,

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}, I_n \right).$$

Let $\mathbb{O} \in \mathbb{R}^{n \times n}$ be an orthogonal matrix; that is, $\mathbb{O}^T = \mathbb{O}^{-1}$. Then

$$\mathbb{O}\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \sim \mathcal{N}\left(\mathbb{O}\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}, I_n\right).$$

Then the problem becomes $H_0 : \theta_1' = \theta_2' = \ldots = \theta_n' = 0$ against $H_1 : \exists j : \theta_j' \neq 0$ where

$$\theta_j' = \left\langle \mathbb{O}\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}, e_j \right\rangle.$$

Then a "reasonable" test must satisfy $\phi(X) = \phi(\mathbb{O}X)$ for any orthogonal $\mathbb{O}$. For this to be true, we must have $\phi(X) = \phi(\|X\|_2^2)$ (because we basically need spherical invariance). For such tests, the problem becomes

$$H_0 : \lambda := \sum_{i=1}^n \theta_i^2 = 0, \qquad H_a : \lambda > 0.$$

Here $\|X\|_2^2 = \sum_{i=1}^n X_i^2$ has a non-central χ^2 distribution with non-centrality parameter λ. It turns out that this famly as a monotone likelihood ratio, so the solution is easy to obtain. The UMP is

$$\phi(x) = \begin{cases} 1, & \sum_{i=1}^n x_i^2 \geq c \\ 0, & \sum_{i=1}^n x_i^2 < c. \end{cases}$$

We will use groups for this (see also Section **??**).

Definition 1.8.14. $(G, \cdot)$ is a **group** if

1. $a, b \in G \implies a \cdot b \in G$.
2. There exists $e \in G$ such that $a \cdot e = e \cdot a = a$.
3. For all a, there exists $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e$.
4. For all $a, b, c \in G$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.

Theorem 1.8.7.1. Assume that $\{\mathbb{P}_\theta \theta \in \Theta\}$ is an identifiable statistical model ($\theta_1 \neq \theta_2 \implies p_{\theta_1} \neq \mathbb{P}_{\theta_2}$). Suppose

$$X \sim \mathbb{P}_\theta, \qquad X \in S(S = \mathbb{R}^d, S = \mathbb{Z}^d).$$

Let G be a group of bijections of S (where the group operation is the composition). Then

$$\mathbb{P}_\theta(gX \in A) = \mathbb{P}_\theta(X \in g^{-1}A) = \mathbb{P}_\theta \cdot g^{-1}(A).$$

Notice that in general gX doesn't have to be in A for a general g. So we require the following assumption.

Assumption: for any $g \in G$ there exists $\overline{g} : \Theta \to \Theta$ such that $\mathbb{P}_\theta \cdot g^{-1} = \mathbb{P}_{\overline{g}(\theta)}$.

Setup for today: $\{p_\theta, \theta \in \Theta\}$. $X \sim p_\theta, X \in S$. $H_0 : \theta \in \Theta_0, H_a : \theta \in \Theta_a$. Let $(G, \cdot)$ be a group of bijections on $S = \mathbb{R}^n, \mathbb{Z}^n$. Note that $\mathbb{P}_\theta(gX \in A) = \mathbb{P}_\theta(X \in g^{-1}A)$.

Assumption: The distribution $\mathbb{P}_\theta \cdot g^{-1}$ coincides with $\mathbb{P}_{\overline{g}(\theta)}$. True if and only if $\mathbb{P}_\theta(gX \in A) = \mathbb{P}_{\overline{g}(\theta)}(X \in A)$.

Example 1.8.11. f: fixed pdf. $\{f(\cdot - \theta), \theta \in \mathbb{R}\}$: location family. $g \in G \iff \exists \tau \in \mathbb{R} : g_\tau(x) = x + \tau$. Then $\overline{g}_\tau(\theta). = \theta + \tau$. Indeed,

$$\mathbb{P}_\theta(X + \tau < t) = \mathbb{P}_\theta(X < t - \tau) = \int_{-\infty}^{t-\tau} f(x - \theta)\, dx$$

Let $y = x + \tau$, then write this as

$$= \int_{-\infty}^{t} f(y - (\theta + \tau))\, dy = \mathbb{P}_{\theta+\tau}(X < t).$$

Theorem 1.8.7.2. 1. For all $g \in G$, let $\overline{g}$ be a bijection on Θ.

2. $(\overline{G}, \circ)$ is a group (with respect to composition). Here, $\overline{G} = \{\overline{g}, g \in G\}$.

3. $g \in G \implies \overline{g} \in \overline{G}$ is a group homomorphism.

Proof. We will show some initial results. (See also Lemma 6.1.1 in Lehmann and Romano [2005].)

(a) Injectively, if $\overline{g}(\theta_1) = \overline{g}(\theta_2) \implies \theta_1 = \theta_2$. Assume that $\overline{g}(\theta_1) = \overline{g}(\theta_2)$. Then

$$\mathbb{P}_{\overline{g}(\theta_1)}(X \in A) = \mathbb{P}_{\theta_1}(g(X) \in A) = \mathbb{P}_{\theta_1}(X \in g^{-1}(A))$$

$$\mathbb{P}_{\overline{g}(\theta_2)}(X \in A) = \mathbb{P}_{\theta_2}(g(X) \in A) = \mathbb{P}_{\theta_2}(X \in g^{-1}(A)) = \mathbb{P}_{\theta_1}(X \in g^{-1}(A))$$

Since g is a bijection, this means that $\mathbb{P}_{\theta_1}(X \in B) = \mathbb{P}_{\theta_2}(X \in B)$ for all B. Therefore $\theta_1 = \theta_2$ by identifiability of the model.

(b)
$$\mathbb{P}_\theta(X \in A) = \mathbb{P}_\theta(g(g^{-1}(X)) \in A) = \mathbb{P}_{\overline{g}(\theta)}(g^{-1}(X) \in A) = \mathbb{P}_{\overline{g^{-1}}(\overline{g}(\theta))}(X \in A)$$

Therefore $\overline{g^{-1}}(\overline{g}(\theta)) = \theta \iff (\overline{g})^{-1} = \overline{g^{-1}}$ by identifiability.

(c) **Surjectivity.** Need to show: for all θ, there exists θ' such that $\overline{g}(\theta') = \theta$.
By part (b), $\overline{g}(\overline{g^{-1}}(\theta)) = \theta$ for all θ. So take $\theta' = \overline{g^{-1}}(\theta)$, then we have the result.

Now we can use these to prove what we want. (Should follow easily? especially from part (b).)

1.

2.

3. **look if scribe filled in details?** $g_1 g_2 \to \overline{g}_1 \circ \overline{g}_2$.

□

Definition 1.8.15 (**invariance of tests with respect to groups**)**.** The testing problem $H_0 : \theta \in \Theta_0$ against $H_a : \theta \in \Theta_a$ is invariant with respect to $\overline{G}$ if and only if for all $\overline{g} \in \overline{G}$,

$$\overline{g}(\Theta_0) = \Theta_0, \qquad \overline{g}(\Theta_a) = \Theta_a.$$

benefit of this approach: can get tests that are invariant across e.g. rotation (if G is a group of rotations), reduce problem significantly (only need to worry about e.g. length, or other property that is invariant with respect to group).

Definition 1.8.16 (**Orbits**)**.** We say that x_1 is equivalent to x_2 ($x_1 \sim x_2$) if and only if there exists $g \in G$ such that $g(x_1) = x_2$. The equivalence classes are called **orbits** (that is, an orbit contains all elements that are equivalent to one another; the orbit of $x \in S$ is $\{g(x) : g \in G\}$.

Example 1.8.12. If G is a set of rotations, then the orbit is the sphere with radius equal to $\|x\|$.).

1.8.8 Maximal Invariants (Section 6.2 of Lehmann and Romano [2005])

Definition 1.8.17 (**Invariants**)**.** A map $T : S \to \mathbb{R}^d$ is **invariant with respect to** G if and only if

$$x_1 \sim x_2 \implies T(x_1) = T(x_2).$$

T is called a **maximal invariant** if and only if T is invariant and

$$T(x_1) = T(x_2) \implies x_1 \sim x_2;$$

so

$$x_1 \sim x_2 \iff T(x_1) = T(x_2).$$

That is, it is constant on the orbits and for each orbit takes on a different value. All maximal invariants are equivalent in the sense that their sets of constancy coincide.

Definition 1.8.18. Let ϕ be a test. ϕ is **invariant** with respect to G if and only if $\phi(x) = \phi(g(x))$ for all $g \in G, x \in S$.

Definition 1.8.19 (**Invariance of problems**)**.** We say that the problem of testing $H_0 : \theta \in \Theta_0$ against $H_a : \theta \in \Theta_a$ remains invariant under a transformation g if $\overline{g}$ preserves both Θ_0 and Θ_a, so that both

$$\overline{g}\Theta_0 = \Theta_0$$

and

$$\overline{g}\Theta = \Theta$$

hold.

Theorem 1.8.8.1 (**Theorem 6.2.1 in Lehmann and Romano [2005]**). Let T be a maximal invariant with respect to G. Then test ϕ is G-invariant if and only if $\phi(x) = h(T(x))$ for all $x \in S$ and some function h; that is, ϕ must depend on x only through $T(x)$.

Proof. Assume that $\phi(x) = h(T(x))$. Take $g \in G$. Then

$$\phi(g(x)) = h(T(g(x)) = h(T(x)) = \phi(x)$$

(where the third step used the maximal invariance of T with respect to G). Conversely, assume that ϕ is G-invariant. Suppose that x_1, x_2 are such that $T(x_1) = T(x_2)$. Then there exists $g \in G$ such that $x_2 = g(x_1)$. Then $\phi(x_2) = \phi(g(x_1)) = \phi(x_1)$, so ϕ is constant on each orbit, so $\phi = h(T)$ for some function h.

□

Example 1.8.13 (**Location family (like Example 6.2.1(i) in Lehmann and Romano [2005])**). $X_1, \dots, X_n \sim f(\cdot - \theta)$ i.i.d., $\theta \in \mathbb{R}$. Then the invariance transformation is $g(x_1, \dots, x_n) = (x_1 + c, \dots, x_n + c)$ for some $c \in \mathbb{R}$.

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \sim \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} \iff \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} c \\ \vdots \\ c \end{bmatrix}.$$

One maximal invariant is $T = (x_1 - x_2, x_2 - x_3, \dots, x_{n-1} - x_n)$. Another is $T' = (x_1 - x_n, x_2 - x_n, \dots, x_n - x_n)$.

Example 1.8.14 (**Scale family (like Example 6.2.1(ii) in Lehmann and Romano [2005])**). $X_1, \dots, X_n \sim \frac{1}{\theta} f(\frac{\cdot}{\theta})$ i.i.d., $\theta \in \mathbb{R}$. Then an invariance transformation is $g \in G \iff \exists c > 0$ such that $g(x_1, \dots, x_n) = (cx_1, \dots, cx_n)$. (Note that $\overline{g}(\theta) = c\theta$.) Then a maximal invariant is $T = (x_1/x_n, \dots, x_n/x_n)$ (if $x_n \neq 0$).

Example 1.8.15 (**Orthogonal transformation (like Example 6.2.1(iii) in Lehmann and Romano [2005])**). $X \sim \mathcal{N}(0, \Sigma)$, $X \in \mathbb{R}^d$. G is a group of orthogonal transformations: $\langle gu, gv \rangle = \langle u, v \rangle$ for all $g \in G$, $u, v \in \mathbb{R}^d$. G corresponds to the group of orthogonal matrices Observe that $\overline{g}(\Sigma) = g\Sigma g^{-1} = g\Sigma g^T$. Indeed, the covariance matrix of gX is

$$\mathbb{E}(gX)(gX)^T = \mathbb{E}gXX^T g^T = g\mathbb{E}(XX^T)g^T = g\Sigma g^T.$$

A maximal invariant $T(x)$ in this case is $\|X\|_2$.

Example 1.8.16 (**Rank tests (like Example 6.2.2(ii) in Lehmann and Romano [2005])**). Suppose $X_1, \dots, X_n$ are i.i.d. with cdf F that is strictly increasing. Let $g \in G$ if and only if there exists f continuous, strictly monotone such that $g(x_1, \dots, x_n) = (f(x_1), \dots, f(x_n))$. (Then the order of the values is always preserved after applying any transformation.) The maximal invariants are then the ranks defined by $R_j(x_1, \dots, x_n) = |\{k : x_k \leq x_j\}|$. In particular, $x_j = x_{R_j}$.

Example 1.8.17. $X \sim \mathbb{P}_\theta$, $\theta \in \Theta$. $H_0 : \theta \in \Theta_0$, $H_a : \theta \in \Theta_a$. G: a group of bijections. $\overline{G}$: a corresponding group of bijections on Θ. The problem is G-invariant if and only if for all $g \in G$, $g(\Theta_0) \subseteq \Theta_0$, $g(\Theta_a) \subseteq \Theta_a$. Find the UMP G-invariant test.

Solution

Know: if $T(x)$ is the maximal invariant, then any G-invariant test depends on T only. Moreover, the power function of a G-invariant test is $\overline{G}$-invariant: indeed, for all $g \in G$,

$$\beta_\phi(\theta) = \mathbb{E}_\theta \phi(X) = \mathbb{E}_\theta \phi(g(X)) = \mathbb{E}_{\overline{g}(\theta)} \phi(X) = \beta_\phi(\overline{g}(\theta)).$$

Let $\tau(\theta)$ be the maximal invariant with respect to $\overline{G}$, then the distribution of any G-invariant test depends only on $\tau(\theta)$. In many cases, the problem reduces to testing $H_0' : \tau(\theta) \leq \tau_0$ against $H_a' : \tau(\theta) > \tau_0$.

Example 1.8.18. $X_1, \dots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$. $H_0 : \sigma^2 \leq \sigma_0^2$, $H_a : \sigma^2 > \sigma_0^2$. Take G to be a group of parallel shifts $(x_1, \dots, x_n) \sim (x_1 + c, \dots, x_n + c)$. (Then the problem is G-invariant, since the variance isn't affected by a shift in the values.) The sufficient statistic is $(\overline{X}, S^2)$, and a maximal invariant is S^2. Recall that

$$S^2 = \frac{\sigma^2}{n-1} Z,$$

where $Z \sim \chi^2_{n-1}$. We have

$$\overline{G} : (\mu, \sigma^2) \xrightarrow{\overline{g}} (\mu + c, \sigma^2).$$

Note that

$$\left(\frac{1}{n} \sum_{i=1}^n x_i, \frac{1}{n-1} \sum_{i=1}^n \left[X_j - \frac{1}{n} \sum_{i=1}^n X_j \right]^2 \right) \xrightarrow{g} \left(\frac{1}{n} \sum_{i=1}^n x_i + c, \frac{1}{n-1} \sum_{i=1}^n \left[X_j - \frac{1}{n} \sum_{i=1}^n X_j \right]^2 \right)$$

The maximal invariant is S^2:

$$S^2 =^d \frac{\sigma^2}{n-1} W,$$

where $W \sim \chi^2_{n-1}$. The pdf of S^2 is

$$f_{S^2}(x) = \frac{1}{2^{(n-1)/2} \Gamma((n-1)/2)} \exp\left\{ -\frac{x}{2\sigma^2} + \left(\frac{n-3}{2} \right) (\log x - \log \sigma^2) \right\}$$

Note that this has an MLR with respect to $T(X) = X$. Therefore by Karlin-Rubin, the UMP invariant test is

$$\phi^* = \begin{cases} 1, & S^2 \geq c \\ 0, & \text{otherwise.} \end{cases}$$

Example 1.8.19 (Two-sample t - test). $X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma^2)$ i.i.d., $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma^2)$ i.i.d. $H_0 : \mu_1 \geq \mu_2, H_a : \mu_1 < \mu_2$. Note that the ordering of the data is invariant to multiplication by a non-negative number or adding a constant. Therefore a grouping is

$$G : (x_1, \ldots, x_n) \xrightarrow{g} (ax_1 + c, \ldots, ax_n + c), \qquad a > 0, c \in \mathbb{R}.$$

Then the group $\overline{G}$ that acts on the parameters is

$$\overline{G}(\mu_1, \mu_2, \sigma^2) \xrightarrow{\overline{g}} (a\mu_1 + c, a\mu_2 + c, a^2\sigma^2).$$

Sufficient statistic: $(\overline{X}, \overline{Y}, S^2)$, where S^2 is the pooled variance:

$$S^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^{n}[X_i - \overline{X}_n]^2 + \sum_{i=1}^{m}[Y_i - \overline{Y}_m]^2\right)$$

Note that

$$(\overline{X}, \overline{Y}, S^2) \xrightarrow{g} (a\overline{X} + c, a\overline{Y} + c, a^2S^2)$$

and a maximal invariant with respect to G

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{S^2}}.$$

and for the parameters (maximal invariant with respect to $\overline{G}$)

$$\delta = \frac{\tilde{\mu}_1 - \tilde{\mu}_2}{\tilde{\sigma}}$$

So we know the distribution of T can only depend on δ. Note that

$$\text{Var}(\overline{X} - \overline{Y}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2\frac{n+m}{nm}$$

and we can consider

$$T' = \frac{\overline{X} - \overline{Y}}{\sqrt{S^2}}\sqrt{\frac{nm}{n+m}}$$

which has a t distribution. We can also now consider the different test $H_0' : \delta \geq 0$, $H_a : \delta < 0$.

Note: if $\delta \neq 0$, then T' has a non-central Student's t-distribution with non-centrality parameter δ. The pdf is (with $\nu = n + m - 2$)

$$f(x) = \nu^{\nu/2} \frac{\exp\left\{-\frac{\nu\delta^2}{2(x^2+\nu)}\right\}}{\sqrt{\pi}\Gamma(\nu/2)2^{(\nu-1)/2}(x^2+\nu)^{(\nu+1)/2}(x^2+\nu)^{(\nu+1)/2}} \int_0^\infty y^\nu \exp\left\{-\frac{1}{2}\left[6 - \frac{\delta x}{\sqrt{x^2+\nu}}\right]\right\} \, dy$$

But it turns out (not trivial to prove) that T' has a monotone likelihood ratio in δ. The UMP invariant test is therefore

$$\phi^* = \begin{cases} 1, & T' \leq c \\ 0, T' > c \end{cases}$$

where $c = t_{\alpha,n+m-2} < 0$ (standard Student's t, because due to Karlin-Rubin we can just go with the distribution when $\delta = 0$. Only need to look at pdf to establish that it has an MLR.).

1.8.9 Rank Tests (Section 6.8 and 6.9 of Lehmann and Romano [2005])

Example 1.8.20. $X_1, \ldots, X_n$ i.i.d. with cdf F, pdf f. $Y_1, \ldots, Y_m$ i.i.d. with cdf G, pdf g. $H_0 : X_1, \ldots, X_n, Y_1, \ldots, Y_m$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, $H_1 : X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma^2)$ i.i.d., $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma^2)$ i.i.d., $\mu_2 > \mu_1$. A natural "generalization" is the following: if we think the data come from a roughly normal distribution but not exactly, we might want to use a permutation test instead. That is,

$$H_0 : F = G, \qquad H_a : F(z) \leq G(z) \qquad \text{for all } z, \qquad \text{and } G \neq F$$

that is, the Y's are stochastically larger than the X's. (The last condition implies that there exists an x for which the inequality is strict: $F(x) < G(x)$.) The observations remain invariant under monotonic transformations:

$$G : (x_1, \ldots, x_n, y_1, \ldots, y_m) \xrightarrow{g} (h(x_1), \ldots, h(x_n), h(y_1) \ldots, h(y_m))$$

where h is continuous and strictly increasing. Then

$$\overline{G} : (F, G) \xrightarrow{\overline{g}} (F \circ h^{-1}, G \circ h^{-1})$$

where $(F \circ h^{-1})(x) = F(h^{-1}(x)) = \mathbb{P}(h^{-1}(X) \leq x)$, or $h(x)$ has cdf $F \circ h^{-1} = \mathbb{P}(h(x) \leq t)] = \mathbb{P}(X \leq h^{-1}(t)) = F(h^{-1}(t))$. Define

$$Z_1 = X_1, \ldots, Z_n = X_n, Z_{n+1} = Y_1, \ldots, Z_{n+m} = Y_m$$

and let $n + m = N$. Let $Z_{(1)}, \ldots, Z_{(n+m)}$ be the order statistics, and define $R_j = \text{rank}(Z_j) = |\{k : Z_k \leq Z_j\}|$. (Note the relationship $Z_{(R_j)} = Z_j$.) Finally, also note that there are $N!$ orderings possible. Ranks are the maximal invariants. Goal: find the distribution of the ranks, namely find $\mathbb{P}(R = r)$. We will use the following result:

Theorem 1.8.9.1 (**Hoeffding's Formula**). Assume that $\xi_1, \ldots, \xi_N$ are independent random variables and ξ_j has pdf f_j (that is, they might have different distributions). Choose a pdf f_0 such that $f_0 = 0 \implies f_1 = \ldots = f_n = 0$. Then

$$\mathbb{P}(R = r) = \frac{1}{n!}\mathbb{E}_{f_0}\left[\frac{\prod_{i=1}^n f_i(V_{(r_i)})}{\prod_{i=1}^n f_0(V_{(r_i)})}\right]$$

where $V_1, \ldots, V_n$ are i.i.d. with distribution f_0.

In our case, take $f_0 := f$, then

$$\mathbb{P}(R = r) = \frac{1}{n!}\mathbb{E}_{f_0}\left[\frac{\prod_{i=1}^m g_i(V_{(q_i)})}{\prod_{i=1}^m f(V_{(q_i)})}\right]$$

where $g_1, \ldots, g_m$ are the ranks of $Y_1, \ldots Y_m$ among $Z_1, \ldots, Z_N$ (the terms corresponding to x will cancel out). **Claim:** Let $Q = (Q_1 < Q_1 < \ldots < Q_m)$ be the ordered ranks of $Y_1, \ldots, Y_m$ among $Z_1, \ldots, Z_n$. Then Q is sufficient for the distribution of R.

Proof.

$$\mathbb{P}(R = r \mid Q = q) = \frac{\mathbb{P}(R = r, Q = q)}{\mathbb{P}(Q = q)} = \frac{\mathbb{P}(R = r)}{\mathbb{P}(Q = q)}$$

and

$$\begin{aligned}\mathbb{P}(Q = q) = \sum_{r:Q(r)=q} \mathbb{P}(R = r) &= \frac{1}{N!}\sum_{r:Q(r)=q} \mathbb{E}_f\left[\frac{\prod_{i=1}^m g_i(V_{(q_i)})}{\prod_{i=1}^m f(V_{(q_i)})}\right] \\ &= \frac{1}{N!}|\{r : Q(r) = q\}|\mathbb{E}_f\left[\frac{\prod_{i=1}^m g_i(V_{(q_i)})}{\prod_{i=1}^m f(V_{(q_i)})}\right] = \frac{1}{N!}n!m!\mathbb{E}_f\left[\frac{\prod_{i=1}^m g_i(V_{(q_i)})}{\prod_{i=1}^m f(V_{(q_i)})}\right]\end{aligned}$$

Therefore we have

$$\mathbb{P}(R = r \mid Q = q) = \frac{\mathbb{P}(R = r, Q = q)}{\mathbb{P}(Q = q)} = \frac{\mathbb{P}(R = r)}{\mathbb{P}(Q = q)} = \frac{1}{n!m!}$$

which doesn't depend on any parameters, so Q is sufficient for R.

□

Therefore we can consider tests that depend on Q only. Let $U \sim \text{Unif}[0, 1]$. Then $F^{-1}(U)$ has distribution F, since

$$\mathbb{P}(F^{-1}(U) \leq t) = \mathbb{P}(U \leq F(t)) = F(t).$$

Hence

$$\mathbb{P}(R=r)=\frac{1}{n!}\mathbb{E}_f\left[\frac{\prod_{i=1}^m g_i(V_{(q_i)})}{\prod_{i=1}^m f(V_{(q_i)})}\right]=\frac{1}{n!}\mathbb{E}_f\left[\frac{\prod_{i=1}^m g(F^{-1}(U_{(q_i)}))}{\prod_{i=1}^m f(F^{-1}(U_{(q_i)}))}\right]$$

But this is the derivative of a certain expression. Let $\tau(x)=(G\circ F^{-1})(x)$. Then

$$\tau'(x)=G'(F^{-1}(x))(F^{-1})'(x)=\frac{g(F^{-1}(x))}{f(F^{-1}(x))}$$

so

$$\mathbb{P}(R=r)=\frac{1}{n!}\mathbb{E}_f\prod_{i=1}^m\tau'(U_{(q_i)})$$

so τ is the maximal invariant of $\overline{G}$ because the distribution of Q depends only on τ, so it must be the maximal invariant.

Theorem 1.8.9.2 (**Theorem 6.3.2 in Lehmann and Romano [2005]**)**.** If $T(x)$ is invariant under G and if $v(\theta)$ is a maximal invariant under the induced group $\overline{G}$, then the distribution of $T(X)$ depends only on $v(\theta)$.

1.8.10 Likelihood Ratio Tests (Section 12.4.4 of Lehmann and Romano [2005])

Last time:

$$\phi^*(x)=\begin{cases}1, & \left(\frac{\partial}{\partial\theta}p_\theta(x)\big|_{\theta=\theta_0}\right)\Big/p_{\theta_0}(x)>c\\ \gamma, & \left(\frac{\partial}{\partial\theta}p_\theta(x)\big|_{\theta=\theta_0}\right)\Big/p_{\theta_0}(x)=c\\ 0, & \left(\frac{\partial}{\partial\theta}p_\theta(x)\big|_{\theta=\theta_0}\right)\Big/p_{\theta_0}(x)<c\end{cases}$$

We got

$$\frac{\partial}{\partial\theta}\mathbb{P}(Q=q)=\frac{1}{\binom{N}{m}}\sum_{j=1}^m\alpha(q_j)$$

where

$$\alpha(q_j)=\mathbb{E}_\theta\left[\frac{f'(V_{(q_j)})}{f(V_{(q_j)})}\right]$$

Suppose $X_1,\dots,X_n$ i.i.d. distributed according to $P_\theta\in\{P_\theta,\theta\in\Theta\}$. Suppose P_θ has either density function or mass function p_θ. We will test $H_0:\theta\in\Theta_0$ against $H_a:\theta\in\Theta_a$. The likelihood ratio is

$$L_n(\theta,x_1,\dots,x_n)=\prod_{j=1}^n p_\theta(x_j)$$

and we will use the log likelihood

$$\ell_n(\theta) = \log L_n(\theta) = \sum_{j=1}^{n} \log p_\theta(x_j).$$

Then the maximum likelihood estimator (if it is unique) is

$$\hat{\theta}_n := \arg\max_{\theta \in \Theta} l_n(\theta).$$

Definition 1.8.20. Assumption: $\Theta \subseteq \mathbb{R}^d$. Define

$$T_n := \frac{\sup_{\theta \in \Theta_a} L_n(\theta)}{\sup_{\theta \in \Theta_0} L_n(\theta)}$$

Note that this quantity is well-defined even if the MLE is not unique. The likelihood ratio test (LRT) is

$$\phi = \begin{cases} 1, & T_n > c \\ \gamma, & T_n = c \\ 0, & T_n < c \end{cases}$$

for some $c, \gamma > 0$. For simplicity, assume that both

$$\hat{\theta}_n^0 := \arg\max_{\theta \in \Theta_0} L_n(\theta), \qquad \hat{\theta}_n^a := \arg\max_{\theta \in \Theta_a} L_n(\theta)$$

exist. Then

$$T_n := \frac{L_n(\hat{\theta}_n^a)}{L_n(\hat{\theta}_n^0)}.$$

Remark 44. In this simplest case,

$$\Theta_0 = \{\theta_0\}, \qquad \Theta_a = \{\theta_a\}.$$

Then

$$T_n = \frac{\prod_{j=1}^n p_{\theta_a}(x_j)}{\prod_{j=1}^n p_{\theta_0}(x_j)},$$

and the likelihood ratio test coincides with the Neyman-Pearson test.

Define

$$\tilde{T}_n := \frac{\sup_{\theta\in\Theta} L_n(\theta)}{\sup_{\theta\in\Theta_0} L_n(\theta)} = \frac{L_n(\hat{\theta}_n)}{L_n(\theta_0)} = \max\{T_n, 1\}$$

and define

$$\Lambda_n := \log \tilde{T}_n \geq 0.$$

Then the LRT can be equivalently stated as

$$\phi = \begin{cases} 1, & \Lambda_n > c \\ \gamma, & \Lambda_n = c \\ 0, & \Lambda_n < c \end{cases}$$

Example 1.8.21 (**Example 12.4.6 in Lehmann and Romano [2005], p. 524 of pdf**). Let $X \sim$ Multinomial$(n, p_1, \ldots, p_k)$. Assume that $\Theta_0 = \{p_0\}$, where $p_0 = (p_0^{(1)}, \ldots, p_0^{(k)})$ and $\Theta_a \in \Delta_k = \{p \neq p_0\}$. (For example, the experiment could be throwing a dice with k faces n times.) Then we have

$$p_\theta(x_1, \ldots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}.$$

Then

$$\Lambda_n = \log\left(\frac{\sup_{\theta\in\Theta} L_n(\theta)}{\sup_{\theta\in\Theta_0} L_n(\theta)}\right) = \log\left(\frac{L_n(\hat{p}_n)}{L_n(p_0)}\right).$$

The MLE is (exercise)

$$\hat{p}_n = \left(\frac{X_1}{n}, \ldots, \frac{X_k}{n}\right) = n\sum_{i=1}^k \frac{X_i}{n} \log\left(\frac{\hat{p}_i}{p_{0,i}}\right) = n\sum_{i=1}^k \hat{p}_i \log\left(\frac{\hat{p}_i}{p_{0,i}}\right)$$

Therefore

$$\Lambda_n = \log \prod_{i=1}^k \left(\frac{X_i}{np_{0,i}}\right)^{x_i}$$

(which is exactly n times the KL divergence between the maximum likelihood estimator and the null hypothesis value.) Now note that

$$\hat{p}_n = n\sum_{i=1}^k (\hat{p}_i - p_{0,i} + p_{0,i}) \log\left(1 + \frac{\hat{p}_i - p_{0,i}}{p_{0,i}}\right)$$

Under H_0, $\hat{p}_i$ is consistent, so $\hat{p}_i - p_{0,i} = o_{\mathbb{P}}(1)$, meaning that for every $\epsilon > 0$ $\mathbb{P}(|\hat{p}_i - p_{0,i}| > \epsilon) \to 0$ as $n \to \infty$. Since $\log(1+x) = x - x^2/2 + o(x^2)$, we have

$$\Lambda_n = n\sum_{i=1}^{k}(\hat{p}_i \pm p_{0,i})\left[\frac{\hat{p}_i - p_{0,i}}{p_{0,i}} - \frac{(\hat{p}_i - p_{0,i})^2}{2p_{0,i}^2} + o_{\mathbb{P}}((\hat{p}_i - p_{0,i})^2\right]$$

$$= n\sum_{i=1}^{k}\left[\frac{(\hat{p}_i - p_{0,i})^2}{2p_{0,i}} + (\hat{p}_i - p_{0,i}) - \underbrace{\frac{1}{2}\frac{(\hat{p}_i - p_{0,i})^3}{2p_{0,i}^2}}_{o_{\mathbb{P}}((\hat{p}_i - p_{0,i})^2)} - \frac{1}{2}\frac{(\hat{p}_i - p_{0,i})^2}{2p_{0,i}} + o_{\mathbb{P}}((\hat{p}_i - p_{0,i})^2\right]$$

$$= \frac{n}{2}\sum_{i=1}^{k}\left[\frac{(\hat{p}_i - p_{0,i})^2}{p_{0,i}} + \underbrace{(\hat{p}_i - p_{0,i})}_{\mathcal{O}_{\mathbb{P}}(1/\sqrt{n})} + \underbrace{o_{\mathbb{P}}((\hat{p}_i - p_{0,i})^2}_{o_{\mathbb{P}}(1/n)}\right] = \frac{n}{2}\sum_{i=1}^{k}\frac{(\hat{p}_i - p_{0,i})^2}{p_{0,i}} + o_{\mathbb{P}}(1)$$

Hence

$$2\Lambda_n = n\sum_{i=1}^{k}\underbrace{\frac{(\hat{p}_i - p_{0,i})^2}{p_{0,i}}}_{\text{Pearson's } \chi^2 \text{ statistic}} + o_{\mathbb{P}}(1)$$

Under H_0, $2\Lambda_n$ is asymptotically distributed as χ^2_{k-1}.

(Work related to derivation of asymptotics:) Further, as $n \to \infty$,

$$\sqrt{n}\left(\begin{bmatrix} X_1/n \\ \vdots \\ X_k/n \end{bmatrix} - \begin{bmatrix} p_1 \\ \vdots \\ p_k \end{bmatrix}\right) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

where $\Sigma_{ii} = p_i(1-p_i)$, $\Sigma_{ij} = -p_i p_j$, $i \neq j$ (exercise). Σ can be written as

$$\Sigma = \Gamma^{1/2}(I_k - \sqrt{p}\sqrt{p}^T)\Gamma^{1/2}$$

where

$$\sqrt{p} = \begin{bmatrix} \sqrt{p_1} \\ \vdots \\ \sqrt{p_k} \end{bmatrix}, \qquad \Gamma = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_k \end{bmatrix}$$

(Notice that $\|\sqrt{p}\| = 1$. The representation using $I_k - \sqrt{p}\sqrt{p}^T$ is very helpful when figuring out the asymptotics.)

Example 1.8.22 (Related to Example 12.4.5 in Lehmann and Romano [2005], p. 523 of pdf). Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\theta, I_d), \theta \in \mathbb{R}^d$. $H_0 : \theta \in L$, where L is a linear subspace of dimension k, and $H_a : \theta \notin L$. We have

$$L_n(\theta) = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{\sum_{j=1}^n \|x_j - \theta\|_2^2}{2}\right\}$$

so

$$\Lambda_n = \log\left(\frac{\sup_{\theta\in\mathbb{R}^d} L(\theta)}{\sup_{\theta\in L} L_n(\theta)}\right) = \log\left(\frac{L_n(\overline{X}_n)}{\sup_{\theta\in L} L_n(\theta)}\right) = \frac{1}{2}\inf_{\theta\in L}\left[\sum_{j=1}^n \|x_j - \theta\|_2^2 - \sum_{j=1}^n \|x_j - \overline{X}_n\|_2^2\right]$$

Claim: $2\Lambda_n$ has χ^2_{d-k} distribution. Indeed, let Π_L be the orthogonal projection onto L. For all $z \in \mathbb{R}^d$,

$$\|z\|_2^2 = l\|\Pi_l z\|_2^2 = \|\Pi_{L^\perp} z\|_2^2$$

where $L^\perp$ is the orthogonal complement of L. Then

$$2\Lambda_n = \inf_{\theta\in L}\left[\sum_{j=1}^n \|\Pi_L(x_j - \theta)\|_2^2 + \sum_{j=1}^n \|\Pi_{L^\perp}(x_j - \theta)\|_2^2 - \sum_{j=1}^n \|\Pi_L(x_j - \overline{X}_n)\|_2^2 - \sum_{j=1}^n \|\Pi_{L^\perp}(x_j - \overline{X}_n)\|_2^2\right]$$

Note that $\Pi_{L^\perp}\theta = 0$ for any $\theta \in L$. Then

$$\sum_{j=1}^n \|\Pi_L(x_j - \theta)\|_2^2$$

is minimized by $n^{-1}\sum_{i=1}^n \Pi_L x_j = \Pi_l \overline{X}_n$. Hence

$$\begin{aligned} 2\Lambda_n &= \underbrace{\sum_{j=1}^n \|\Pi_L(x_j - \theta)\|_2^2 - \sum_{j=1}^n \|\Pi_L(x_j - \overline{X}_n)\|_2^2}_{0} + \sum_{j=1}^n \|\Pi_{L^\perp} x_j\|_2^2 - \sum_{j=1}^n \|\Pi_{L^\perp}(x_j - \overline{X}_n)\|_2^2 \\ &= 2n\sum_{j=1}^n \angle\Pi_{L^\perp} x_j, \Pi_{L^\perp}\overline{X}_n\rangle - n\|\Pi_{L^\perp}\overline{X}_n\|_2^2 = 2n\|\Pi_{L^\perp}\overline{X}_n\|_2^2 - n\|\Pi_{L^\perp}\overline{X}_n\|_2^2 = n\|\Pi_{L^\perp}\overline{X}_n\|_2^2. \end{aligned}$$

Finally, observe that for $Z = \sqrt{n}(\overline{X}_n - \theta) \sim \mathcal{N}(0, I_d)$ under H_0,

$$n\|\Pi_{L^\perp}\overline{X}_n\|_2^2 = \|\Pi_{L^\perp} Z\|_2^2$$

(since $\theta \in L$). $\Pi_{L^\perp} Z$ is normal, and its covariance has eigenvalues all equal to 1 (since this is a projection matrix). Therefore $\|\Pi_{L^\perp} Z\|_2^2 \sim \chi^2_{d-k}$ as claimed.

Heuristic derivation of the asymptotics of the likelihood ratio test: Suppose $X_1, \ldots, X_n$ are i.i.d. from $\mathbb{P}_\theta$, $\theta \in \Theta \subseteq \mathbb{R}^d$. $\mathbb{P}_\theta$ has density/pmf p_θ. The family $\{p_\theta, \theta \in \Theta\}$ is *regular* (informally: support

of p_θ does not depend on θ and $\log p_\theta(x)$ is smooth with respect to θ; can do Taylor expansions, etc. The weakest form of regularity conditions is known as QMD (quadratic mean differentiability)).

Theorem 1.8.10.1 (Wilk's Theorem (see Section 12.4.2 of Lehmann and Romano [2005], p. 518 of pdf)). Assume that we are interested in testing $H_0 : \theta \in \Theta_0 = L$, $H_a : \theta \in \Theta \setminus \Theta_0$, where $\Theta_0 = L$ is an affine subspace of dimension k. **note: in 11/11 notes we defined $L := \Theta_0 - \theta_0$ so that Θ_0 is an affine subspace and L is a linear subspace (θ_0 is the offset). Not sure if this is different from notes on 11/08 or if I misunderstood on that lecture.** Then

$$2\Lambda_n = 2\log\left(\frac{\sup_\Theta L_n(\theta)}{\sup_{\theta\in\Theta_0} L_n(\theta)}\right) \xrightarrow{d} \chi^2_{d-k}$$

under H_0.

Proof. Define

$$L_n(\theta) = \prod_{j=1}^n p_\theta(x_j), \qquad \hat{\theta}_n = \arg\max_{\theta\in\Theta} L_n(\theta), \qquad \hat{\theta}_L := \arg\max_{\theta\in\Theta_0} L_n(\theta), \qquad f_\theta(x) = \log p_\theta(x),$$
$$f'_\theta(x) = \frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(x) = \nabla \log p_\theta(x), \qquad \hat{u}_n = \sqrt{n}(\hat{\theta}_n - \theta_0), \qquad \hat{u}_{n,L} = \sqrt{n}(\hat{\theta}_L - \theta_0).$$

Let

$$Z_n(u) := \sum_{j=1}^n \left[f_{\theta_0+u/\sqrt{n}}(x_j) - f_{\theta_0}(x_j)\right].$$

Note that

$$\Lambda_n = \log\frac{\prod_{j=1}^n p_{\hat{\theta}_n}(x_j)}{\prod_{j=1}^n p_{\hat{\theta}_L(x_j)}} = \sum_{j=1}^n \left[f_{\hat{\theta}_n}(x_j) - f_{\hat{\theta}_L}(x_j)\right] = \sum_{j=1}^n \left[f_{\hat{\theta}_n}(x_j) - f_{\theta_0}(x_j)\right] - \sum_{j=1}^n \left[f_{\hat{\theta}_L}(x_j) - f_{\theta_0}(x_j)\right]$$
$$= Z_n(\hat{u}_n) - Z_n(\hat{u}_{n,L}) \quad (1.63)$$

(since $\hat{\theta}_n = \theta_0 + \hat{u}_n/\sqrt{n}$).

Then we can write the Taylor expansion as

$$f_{\theta_0+u/\sqrt{n}}(x) = f_{\theta_0}(x) + \left\langle f'_{\theta_0}(x), \frac{u}{\sqrt{n}}\right\rangle + \frac{1}{2}\left\langle \underbrace{f''_{\theta_0}(x)}_{\text{Hessian}}\frac{u}{\sqrt{n}}, \frac{u}{\sqrt{n}}\right\rangle + \underbrace{R_{\theta_0}\left(\frac{u}{\sqrt{n}}, x\right)\frac{\|u\|_2^2}{n}}_{\to 0 \text{ as } u/\sqrt{n}\to 0 \text{ (or } n\to\infty)}$$

$$Z + n(u) = \left\langle \frac{1}{\sqrt{n}}\sum_{j=1}^n f'_{\theta_0}(x_j), u\right\rangle + \frac{1}{2}\left\langle \frac{1}{n}\sum_{j=1}^n f''_{\theta_0}(x_j)u, u\right\rangle + \underbrace{\left[\sum_{j=1}^n R_{\theta_0}\left(\frac{u}{\sqrt{n}}, X_j\right)\right]\frac{\|u\|_2^2}{n}}_{o_{\mathbb{P}}(1)}$$

(randomness comes from fact that x_j are random variables with distribution p_{θ_0}). Note that

$$\frac{1}{n}\sum_{j=1}^{n} f''_\theta(X_j) \to \mathbb{E} f''_\theta(X)$$

with probability 1 by the Strong Law of Large Numbers (applied element-wise in the matrix). Moreover,

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n} f'_{\theta_0}(X_j) \xrightarrow{d} \mathcal{N}(0, I(\Theta_0)) \tag{1.64}$$

where $I(\theta_0)$ is the Fisher information matrix. Indeed,

$$\mathbb{E} f'_\theta(X_j) = 0$$

$$f'_{\theta_0}(X_j) = \nabla \log p_\theta(x_j) = \begin{pmatrix} \frac{\frac{\partial}{\partial \theta_1} p_\theta(X_j)}{p_\theta(X_j)} \\ \vdots \\ \frac{\frac{\partial}{\partial \theta_d} p_\theta(X_j)}{p_\theta(X_j)} \end{pmatrix}$$

Then (using regularity assumptions)

$$\mathbb{E}\left[\frac{\frac{\partial}{\partial \theta_1} p_\theta(x_1)}{p_\theta(x_1)}\right] = \int_{\mathbb{R}^d} \frac{\frac{\partial}{\partial \theta_1} p_\theta(y)}{p_\theta(y)} p_\theta(y)\ dy = \frac{\partial}{\partial \theta_1}\int_{\mathbb{R}^d} p_\theta(y)\ dy = \frac{\partial}{\partial \theta_1} 1 = 0.$$

By the Central Limit Theorem,

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n} f'_\theta(X_j)$$

is approximately normal with covariance equal to the Fisher information

$$\mathbb{E}_{\theta_0} f'_\theta(X)\left(f'_\theta(X)\right)^T = \mathbb{E}_{\theta_0}(\nabla \log p_{\theta_0}(x))(\nabla \log p_{\theta_0}(x))^T.$$

Hence

$$Z_n(u) = \langle A_n, u\rangle + \frac{1}{2}\langle B_n u, u\rangle + o_{\mathbb{P}}(1)$$

where

$$A_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} f'_{\theta_0}(X_j), \qquad B_n = \frac{1}{n}\sum_{j=1}^{n} f''_{\theta_0}(X_j).$$

Also,

$$Z_n(u) = \langle A_n, u\rangle - \frac{1}{2}\langle I(\theta_0)u, u\rangle + o_{\mathbb{P}}(1)$$

above from 11/11 lecture; below are my notes from 11/08 lecture which may be wrong?

$$Z_n(u) = \langle A_n, u\rangle + \frac{1}{2}\langle I(\theta_0)u, u\rangle + o_{\mathbb{P}}(1)$$

where $I(\theta_0) := \mathbb{E}f'_{\theta_0}(x)(f'_{\theta_0}(x))^T$ (the Fisher information matrix) since $\mathbb{E}f''_\theta(X_1) = -I(\theta_0)$ (exercise; can see by integration by parts. Do in one-dimensional case and then can see how it works in general.).

One-dimensional case:

$$\int \left(\frac{\partial}{\partial\theta}\log p_\theta(x)\right)^2 p_\theta(x)\ dx = \int \frac{[p'_\theta(x)]^2}{p_\theta(x)}\ dx$$

We want to show that this equals

$$-\int \frac{\partial^2}{\partial\theta^2}\log[p_\theta(x)]p_\theta(x)\ dx = -\int \frac{\partial}{\partial\theta}\left[\frac{p'_\theta(x)}{p_\theta(x)}\right]p_\theta(x)\ dx = -\int \frac{p''_\theta(x)p_\theta(x) - (p'_\theta(x))^2}{p_\theta(x)}\ dx$$
$$\iff \int \frac{p''_\theta(x)p_\theta(x)}{p_\theta(x)}\ dx = 0$$

By regularity,

$$\frac{\partial}{\partial\theta}\underbrace{\int p_\theta(x)\ dx}_{=1} = \int \frac{\partial}{\partial\theta}p_\theta(x)\ dx = 0.$$

Idea: since $\hat{u}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$ maximizes $Z_n(u)$, it must be close to $\tilde{u}_n$ that maximizes $\langle A_n, u\rangle - (1/2)\langle I(\theta_0)u, u\rangle$ (since they only differ by $o_{\mathbb{P}}(1)$). (to show formally: need to show that $o_{\mathbb{P}}(1)$ term is in fact $o_{\mathbb{P}}(1)$ over all u. takes a lot of work.) Note that $\tilde{u}_n$ must solve $A_n = I(\theta_0)\tilde{u}_n$ because $\langle A_n, u\rangle - (1/2)\langle I(\theta_0)u, u\rangle$ is concave so it is maximized where its gradient equals 0 (exercise). Since $I(\theta_0)$ is invertible, this is given by $\tilde{u}_n = I^{-1}(\theta_0)A_n$. By (1.64), $A_n \xrightarrow{d} \mathcal{N}(0, I(\theta_0))$. Let $Y_n := I^{-1/2}(\theta_0)A_n$; then $Y_n \xrightarrow{d} \mathcal{N}(0, I_d)$. Similarly, $\hat{u}_L$ must be close to $\tilde{u}_L$ that maximizes

$$\langle A_n, u\rangle - \frac{1}{2}\langle I(\theta_0)u, u\rangle$$

over all $u \in L$. Note that

$$\langle A_n, u\rangle - \frac{1}{2}\langle I(\theta_0)u, u\rangle = \langle I^{1/2}(\theta_0)Y_n, u\rangle - \frac{1}{2}\langle I^{1/2}(\theta_0)u, I^{1/2}(\theta_0 u\rangle = \langle I^{1/2}(\theta_0)Y_n, u\rangle - \frac{1}{2}\|I^{1/2}(\theta_0)u\|_2^2$$
$$= \langle I^{1/2}(\theta_0)u, Y_n\rangle - \frac{1}{2}\|I^{1/2}(\theta_0)u\|_2^2 - \frac{1}{2}\|Y_n\|_2^2 + \frac{1}{2}\|Y_n\|_2^2 = -\frac{1}{2}\|I^{1/2}(\theta_0)u - Y_n\|_2^2 + \frac{1}{2}\|Y_n\|_2^2.$$

Hence $\tilde{u}_n$ minimizes $\|I^{1/2}(\theta_0)u - Y_n\|_2^2$ over $u \in L$. Let $L(\theta_0)$ be the image of L under $I^{1/2}(\theta_0)$. $L(\theta_0)$ is a linear subspace of dimension k. The minimizer of $\|v - Y_n\|_2^2$ over $v \in L(\theta_0)$ is

$$\tilde{v} := \text{proj}_{L(\theta_0)} Y_n = \Pi_{L(\theta_0)} Y_n.$$

Hence $\tilde{u}_{n,L} = I^{-1/2}(\theta_0)\Pi_{L(\theta_0)}Y_n$ and $\tilde{u}_n = I^{-1}(\theta_0)A_n = I^{-1/2}(\theta_0)Y_n$. Then using (1.63),

$$\begin{aligned}\Lambda_n = Z_n(\hat{u}_n) - Z_n(\hat{u}_{n,L}) &= Z_n(\tilde{u}_n) - Z_n(\tilde{u}_{n,L}) + o_{\mathbb{P}}(1) \\ &= \langle A_n, \tilde{u}_n\rangle - \frac{1}{2}\langle I(\theta_0)\tilde{u}_n, \tilde{u}_n\rangle - \langle A_n, \tilde{u}_L\rangle + \frac{1}{2}\langle I(\theta_0)\tilde{u}_L, \tilde{u}_L\rangle + o_{\mathbb{P}}(1)\end{aligned}$$

And

$$\begin{aligned}\langle A_n, \tilde{u}_n\rangle = \langle I^{1/2}(\theta_0)Y_n, I^{-1/2}(\theta_0)Y_n\rangle &= \|Y_n\|_2^2, \\ \frac{1}{2}\langle I(\theta_0)\tilde{u}_l, \tilde{u}_L\rangle = \frac{1}{2}\langle I(\theta_0)I^{-1/2}(\theta_0)Y_n, I^{-1/2}(\theta_0)Y_n\rangle &= \frac{1}{2}\|Y_n\|_2^2, \\ \langle A_n, \tilde{u}_L\rangle = \langle I^{1/2}(\theta_0)Y_n, I^{-1/2}(\theta_0)\Pi_{L(\theta_0)}Y_n\rangle = \langle Y_n, \Pi_{L(\theta_0)}Y_n\rangle &= \|\Pi_{L(\theta_0)}Y_n\|_2^2, \\ \frac{1}{2}\langle I(\theta_0)\hat{u}_L, \hat{u}_L\rangle &= \frac{1}{2}\|\Pi_{L(\theta_0)}Y_n\|_2^2.\end{aligned}$$

So

$$\Lambda_n = \frac{1}{2}\|Y_n\|_2^2 - \frac{1}{2}\|\Pi_{L(\theta_0)}Y_n\|_2^2$$

and by (1.64) we then have

$$2\Lambda_n = \|\Pi_{L(\theta_0)^\perp}Y_n\|_2^2 \xrightarrow{d} \chi^2_{d-\dim(L(\theta_0))} = \chi^2_{d-k}$$

where we used the general fact that

$$\langle x, \Pi_L x\rangle = \langle \Pi_L x + \Pi_{L^\perp}x, \Pi_L x\rangle = \|\Pi_L x\|_2^2.$$

□

Remark 45. Theorem 1.8.10.1 is often applied in significance tests for the coefficients in logistic regression.

Remark 46. If Θ_0 is a halfspace and θ_0 is on the boundary, then the χ^2 asymptotics don't hold anymore (because no matter how far we "zoom in", we can't get it to locally "look like" an affine subspace). That is, we can get this to work for $H_0 : \theta = \theta_0$ but not for $H_0 : \theta \leq \theta_0$.

1.8.11 Bahadur's Relative Efficiency (Section 10.4 of Serfling [1980])

Setup: $X_1, \dots, X_n$ i.i.d. from $\mathbb{P}_\theta$, $\theta \in \Theta$. $\{\phi_n\}_{n \geq 1}$ is a sequence of tests for testing $H_0 : \theta \in \Theta_0, H_a : \theta \in \Theta_a$. Let $\theta' \in \Theta_a$. Want to define

$$n(\underbrace{\alpha}_{\text{size}}, \underbrace{\gamma}_{\text{power}}, \theta') := \min \left\{ n \geq 1 : \sup_{\theta \in \Theta_0} \beta_\theta(\phi_n) \leq \alpha, \beta_{\theta'}(\phi_n) \geq \gamma \right\},$$

the smallest sample size required to achieve size α and power γ in the case of alternative θ'. If $\{\phi_n^{(1)}\}_{n \geq 1}$, $\{\phi_n^{(2)}\}_{n \geq 1}$ are two sequences of tests, consider the ratio

$$\lim_{\alpha \to 0} \frac{n^{(2)}(\alpha, \gamma, \theta')}{n^{(1)}(\alpha, \gamma, \theta')},$$

(if the limit exists), which is called **Bahadur's relative efficiency**. (Clearly we prefer if n is smaller, so we will choose sequence (1) if this ratio is greater than 1 and (2) otherwise). We will be interested in tests of the form

$$\phi = \begin{cases} 1, & T_n \geq C_n \\ 0, & T_n < C_n, \end{cases}$$

where $T_n = T_n(X_1, \dots, X_n)$ is a test statistic. (Note that likelihood ratio tests are of this form.) We will make the following assumptions:

1. Under $\mathbb{P}_\theta$, $T_n \xrightarrow{p} \mu(\theta)$.
2. $-\frac{2}{n} \log(\mathbb{P}_\theta(T_n \geq t)) \xrightarrow{n \to \infty} \nu(t, \theta)$ and $\mu(\theta)$, $\nu(t, \theta)$ are continuous functions.

Second assumption tells us about the asymptotic behavior of the power function for any θ. Strict assumption; doesn't hold in many cases, does hold in many problems involving normal distribution.

Example 1.8.23. Recall Example 1.8.21 with $X \sim \text{Multinomial}(n, \theta_1, \dots, \theta_k)$. We got the likelihood ratio statistic **(this expression below has errors, I couldn't read it)**

$$2\Lambda_n = n \sum_{j=1}^k \hat{\theta}_j \log(\hat{\theta}_j / \theta), \qquad \hat{\theta}_j = x_j / n$$

$$\iff \frac{2\Lambda_n}{n} = \sum_{j=1}^k \hat{\theta}_j \log(\hat{\theta}_j / \theta_{0,j}.$$

Under $\mathbb{P}_\theta$, $\hat{\theta}_j \xrightarrow{p} \theta_j$ by the Law of Large Numbers.

$$\frac{2\Lambda_n}{n} \xrightarrow{p} \sum_{j=1}^n \theta_j \log(\theta_j / \theta_{n,j}) = KL(\mathbb{P}_\theta || \mathbb{P}_{\theta_0}) = \mu(\theta)$$

second assumption is harder to check

Example 1.8.24. $X_1, \dots, X_n \sim \mathcal{N}(\theta, I_d)$. Test $H_0 : \theta \in \Theta_0$ where Θ_0 is an affine subspace, $H_a : \theta \notin \Theta_0$. In Theorem 1.8.10.1, we obtained

$$2\Lambda_n = \|\Pi_{L^\perp} \overline{X}_n\|_2^2 \cdot n,$$

where L is the linear subspace corresponding to Θ_0. Then

$$\frac{2\Lambda_n}{n} = \|\Pi_{L^\perp} \overline{X}_n\|_2^2.$$

Under $\mathbb{P}_\theta$, $\overline{X}_n \xrightarrow{p} \theta$, so

$$\|\Pi_{L^\perp} \overline{X}_n\|_2^2 \xrightarrow{p} \|\Pi_{L^\perp} \theta\|_2^2.$$

Example 1.8.25. $X_1, \dots, X_n \sim \mathcal{N}(\theta, \sigma^2 I_n)$, $\sigma^2 > 0$ is known. Want to test $H_0 : \theta = \mu_0, H_a : \theta \neq \mu_0$. Then can see that likelihood ratio test is (for some $c > 0$)

$$\phi(x) = \begin{cases} 1, & \left| \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma} \right| \geq c, \\ 0, & \left| \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma} \right| < c \end{cases}$$

Note that using the test statistic

$$T_n' = \frac{\overline{X}_n - \mu_0}{\sigma}$$

is equivalent to $\frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma}$ (after adjusting c appropriately). Under $\mathcal{N}(\theta, \sigma^2)$, $\overline{X}_n \xrightarrow{p} \theta$, hence $T_n' \xrightarrow{p} (\theta - \mu_0)/\sigma := \mu(\theta)$. Then

$$\mathbb{P}_\theta(T_n' \geq t) = \mathbb{P}_\theta\left(\frac{\overline{X}_n - \mu_0}{\sigma} \geq t\right) = \mathbb{P}_\theta\left(\frac{\overline{X}_n - \theta}{\sigma} \geq t + \frac{\mu_0 - \theta}{\sigma}\right) = \mathbb{P}_\theta\left(\underbrace{\frac{\sqrt{n}(\overline{X}_n - \theta)}{\sigma}}_{\mathcal{N}(0,1)} \geq \sqrt{n}\left[t + \frac{\mu_0 - \theta}{\sigma}\right]\right)$$
$$= \exp\left\{-\frac{1}{2}\left[\sqrt{n}\left(t + \frac{\mu_0 - \theta}{\sigma}\right)\right]^2 + o(1)\right\} \approx \exp\left\{-\frac{1}{2}\left[\sqrt{n}\left(t + \frac{\mu_0 - \theta}{\sigma}\right)\right]^2\right\}$$

where the last step used the result of the following exercise (well-known bound): Let $\phi(t) = \mathbb{P}(\xi \geq t)$ where $\xi \sim \mathcal{N}(0, 1)$. (So $\phi(t) = \int_t^\infty (2\pi)^{-1/2} e^{-x^2/2}\, dx$.) Then

1. $\phi(t) \leq e^{-t^2/2}$ for any t.
2. $\phi(t) \geq \frac{1}{\sqrt{2\pi}} \frac{t}{1+t^2} e^{-t^2/2}$.

In particular, as $t \to \infty$,

$$\frac{t^2}{2} \leq -\log \phi(t) \leq \frac{t^2}{2} - \underbrace{\log\left(\frac{1}{\sqrt{2\pi}} \frac{t}{1+t^2}\right)}_{o(t) \text{ as } t \to \infty}$$

Returning to the problem,

$$-\frac{2}{n} \log\left[\mathbb{P}_\theta(T_n' \geq t)\right] = \left(t + \frac{\mu_0 - \theta}{\sigma}\right)^2 / \nu(\theta, t) + \underbrace{o(1)}_{\to 0 \text{ as } n \to \infty}.$$

(can see how in practice checking second assumption is hard.) Assume that $\Theta_0 = \{\theta_0\}$. Let $\hat{\alpha}$ be the p-value, i.e.

$$\hat{\alpha}_n = \sup_{\theta_0 \in \Theta_0} \left\{ \underbrace{\mathbb{P}_{\theta_0}(T_n \geq t)}_{G(t)} \big|_{t=T_n} \right\} = \sup_{\theta_0 \in \Theta_0} \{G(T_n)\}$$

(probability of observing an outcome more extreme than the current one). Note: When $\Theta_0 = \{\theta_0\}$, $\hat{\alpha}_n = \mathbb{P}_{\theta_0}(T_n \geq t)|_{t=T_n}$. Also note that the test of size α rejects if and only if $\hat{\alpha}_n \leq \alpha$. By our assumption,

$$\hat{\alpha}_n = \exp\left\{-\frac{n}{2}\left[\nu(T_n, \theta_0) + o(1)\right]\right\} = \exp\left\{-\frac{n}{2}\left[\underbrace{\nu(\mu(\theta), \theta_0)}_{\text{"Bahadur's Slope"}} + o_{\mathbb{P}}(1)\right]\right\}$$

because $T_n \xrightarrow{p} \mu(\theta)$.

Claim: for any $\gamma < 1$,

$$\lim_{\alpha \to 0} n(\alpha, \gamma, \theta') = \frac{\log(1/\alpha)}{\nu(\mu(\theta))} + o(1).$$

$$n(\alpha, \gamma, \theta') = \frac{\log(1/\alpha)}{\nu(\mu(\theta'), \theta')} + o(1).$$

Hence Bahadur's relative efficiency is approximately equal to

$$\frac{\nu_1(\mu_1(\theta'), \theta')}{\nu_2(\mu_2(\theta'), \theta')}$$

"the larger the function ν is, the smaller the sample size is, so the tests with higher Bahadur relative efficiencies." Likelihood ratio tests achieve the asymptotic highest possible Bahadur relative efficiency. Related at its core to the fact that methods based on maximum likelihood reduce some notion of KL divergence between distributions. So achieve information-theoretic lower limits on best possible solutions (since

KL-divergence comes from information theory). Note that if BRE is less than 1, $\{\phi_n^{(2)}\}$ is more efficient, which is true if $\nu_2(\mu_2(\theta'),\theta') > \nu_1(\mu_1(\theta'),\theta')$.

Proof of claim. ϕ_n rejects when $\hat{\alpha}_n \leq \alpha$. Hence

$$\beta_{\phi_n}(\theta) = \mathbb{P}_\theta(\hat{\alpha}_n \leq \alpha) = \mathbb{P}_\theta\left(\exp\left\{-\frac{n}{2}(\nu(\mu(\theta_0),\theta_0) + o_{\mathbb{P}}(1)\right\} \leq \alpha\right) = \mathbb{P}_\theta\left(n \geq \frac{2\log(1/\alpha)}{\nu(\mu(\theta_0),\theta_0) + o_{\mathbb{P}}(1)}\right)$$

We want to choose n such that this probability is greater than or equal to γ. Take $\epsilon > 0$ and assume that

$$n = (1+\epsilon)\frac{2\log(1/\alpha)}{\nu(\mu(\theta_0),\theta_0)}$$

Then

$$\mathbb{P}_\theta\left(n \geq \frac{2\log(1/\alpha)}{\nu(\mu(\theta_0),\theta_0) + o_{\mathbb{P}}(1)}\right) = \mathbb{P}_\theta\left(n \geq \frac{2\log(1/\alpha)}{\nu(\mu(\theta_0),\theta_0)[1 + o_{\mathbb{P}}(1)]}\right)$$
$$= \mathbb{P}_\theta\left(1 + \epsilon \geq \frac{1}{1 + o_{\mathbb{P}}(1)}\right) \to 1 \text{ as } \alpha \to 0.$$

So $n(\alpha,\gamma,\theta) < (1+\epsilon)\frac{2\log(1/\alpha)}{\nu(\mu(\theta_0),\theta_0)}$. Formally,

$$\lim_{\alpha\to 0} \frac{n(\alpha,\gamma,\theta)}{\frac{2\log(1/\alpha)}{\nu(\mu(\theta_0),\theta_0)}} \leq 1 + \epsilon \qquad \forall \epsilon > 0.$$

Now take

$$n = (1-\epsilon)\frac{2\log(1/\alpha)}{\nu(\mu(\theta_0),\theta_0)}$$

Then

$$\mathbb{P}_\theta\left(1 - \epsilon \geq \frac{1}{1 + o_{\mathbb{P}}(1)}\right) \to 0 \text{ as } \alpha \to 0,$$

so

$$\lim_{\alpha\to 0} \frac{n(\alpha,\gamma,\theta)}{\frac{2\log(1/\alpha)}{\nu(\mu(\theta_0),\theta_0)}} = 1 \qquad \forall \gamma < 1.$$

□

Exercise 7. Assume that $\Theta_0 = \{\theta_0\}$. Under H_0, $\mathbb{P}_{\theta_0}(\hat{\alpha}_n \leq t) \leq t$, $t \in [0,1]$. When $\mathbb{P}_{\theta_0}$ has density with respect to Lebesgue measure (i.e. $d\mathbb{P}_{\theta_0}(x) = p_{\theta_0}(x)\, dx$), then under $\mathbb{P}_{\theta_0}$ $\hat{\alpha}_n \sim \text{Uniform}[0,1]$. (Hint: if X has density function $F(x)$, then $F'(x) \sim \text{Uniform}[0,1]$ when F is continuous.)

Theorem 1.8.11.1. Let $\{T_n\}_{n\geq 1}$ be any sequence of test statistics, and $\{\phi_n\}_{n\geq 1}$ corresponding tests, $\hat{\alpha}_n$ associated p-values. Then for all $\theta' \in \Theta_a$,

$$\limsup_{n\to\infty}\left[-\frac{2}{n}\log(\hat{\alpha}_n) \leq 2\,\text{KL}(p_{\theta'}||p_{\theta_0})\right]$$

with probability 1, where $\text{KL}(p_{\theta'}||p_{\theta_0})$ is the KL divergence. (Result also holds for composite null and alternative hypotheses.)

Proof. Let

$$\Lambda_n := \sum_{j=1}^{n}\log\left(\frac{p_\theta(x_j)}{p_{\theta_0}(x_i)}\right).$$

Take $B > A > 0$. Let

$$\mathcal{E}_n = \left\{\hat{\alpha}_n \leq e^{-nB}, \frac{\Lambda_n}{n} \leq A\right\}.$$

Then

$$\begin{aligned}
\mathbb{P}_\theta(\mathcal{E}_n) &= \int p_\theta(x_1)\cdots p_\theta(x_n)\mathbb{1}\left\{\hat{\alpha}_n \leq e^{-nB}, \Lambda_n \leq nA\right\}\, d\mu_1(x_1)\ldots d\mu_n(x_n) \\
&= \int \underbrace{\frac{p_\theta(x_1)\cdots p_\theta(x_n)}{p_{\theta_0}(x_1)\cdots p_{\theta_0}(x_n)}}_{e^{\Lambda_n}} \mathbb{1}\left\{\hat{\alpha}_n \leq e^{-nB}, \Lambda_n \leq nA\right\} p_{\theta_0}(x_1)\cdots p_{\theta_0}(x_n)\, d\mu_1(x_1)\ldots d\mu_n(x_n) \\
&\leq e^{nA}\int \mathbb{1}\left\{\hat{\alpha}_n \leq e^{-nB}\right\} p_{\theta_0}(x_1)\cdots p_{\theta_0}(x_n)\, d\mu_1(x_1)\ldots d\mu_n(x_n) = e^{nA}\underbrace{\mathbb{P}_{\theta_0}\left(\hat{\alpha}_n \leq e^{-nB}\right)}_{\leq e^{-nB}\ \text{(exercise)}} \\
&\leq e^{nA}e^{-nB} = e^{-n(\underbrace{B-A}_{>0})}
\end{aligned}$$

But

$$\sum_{n\geq 1}\mathbb{P}(\mathcal{E}_n) \leq \sum_{n\geq 1} e^{-n(B-A)} < \infty$$

By Borel-Cantelli lemma, on event Ω_1 of $\mathbb{P}_\theta$ probability 1 for n large enough, $\overline{\mathcal{E}}_n$ occur, where

$$\overline{\mathcal{E}}_n = \left\{\hat{\alpha}_n > e^{-nB}\right\} \cup \left\{\Lambda_n > nA\right\}.$$

Then by the Strong Law of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p_\theta(x_j)}{p_{\theta_0}(x_j)}\right) \xrightarrow{n\to\infty} \mathbb{E}_\theta \log\left(\frac{p_\theta(x_j)}{p_{\theta_0}(x_j)}\right) = KL(p_\theta||p_{\theta_0})$$

with probability 1. If $A > KL(p_\theta||p_{\theta_0})$, then there exists an event Ω_2 such that $\mathbb{P}_\theta(\Omega_2) = 1$ such that on Ω_2,

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p_\theta(x_j)}{p_{\theta_0}(x_j)}\right) < A$$

for n large enough (if and only if $\Lambda_n < nA$). On $\Omega_1 \cap \Omega_2$, for all n large enough

$$\hat{\alpha}_n > e^{-nB} \iff -\frac{2}{n}\log(\hat{\alpha}_n) < 2B$$

for n large enough, and $\mathbb{P}_\theta(\Omega_1 \cap \Omega_2) = 1$. But B can be taken arbitrarily close to $KL(p_\theta||p_{\theta_0})$. Therefore the only possibility is

$$\limsup_{n\to\infty}\left[-\frac{2}{n}\log(\hat{\alpha}_n)\right] \leq 2KL(p_\theta||p_{\theta_0}).$$

□

This upper bound is sharp, as we will show next. In particular, we will show that the upper bound from the previous theorem is achieved by likelihood ratio tests. Assumption: Θ_a is finite. Recall that

$$\tilde{\Lambda}_n := \log\frac{\max_{\theta\in\Theta}\prod_{j=1}^{n}p_\theta(x_j)}{\prod_{i=1}^{n}p_{\theta_0}(x_j)} = \log\frac{\max_{\theta\in\Theta}L_n(\theta)}{L_n(\theta_0)}.$$

Theorem 1.8.11.2. Let

$$\hat{\alpha}_n := \mathbb{P}_{\theta_0}\left(\underbrace{\frac{\tilde{\Lambda}_n}{n}}_{T_n} \geq t\right)\Big|_{t=\tilde{\Lambda}/n}.$$

Then for all $\theta \in \Theta_a$,

$$-\frac{2}{n}\log(\hat{\alpha}_n) \xrightarrow{a.s.} 2KL(p_\theta||p_{\theta_0}).$$

Remark 47. The likelihood ratio test achieves the information theoretic upper bound from the previous theorem. So likelihood ratio tests are asymptotically optimal. (Not necessarily UMP, UMP unbiased, UMP invariant, etc.; weaker claim.)

Proof.

$$\begin{aligned}
\mathbb{P}_{\theta_0}(\tilde{\Lambda}_n \geq nt) = \mathbb{P}_{\theta_0}\left(\log \frac{\max_{\theta\in\Theta} L_n(\theta)}{L_n(\theta_0)} \geq nt\right) &= \mathbb{P}_{\theta_0}\left(\max_{\theta\in\Theta} \sum_{j=1}^{n} \left\{(\log \frac{p_\theta(x_j)}{p_{\theta_0}(x_j)}\right\} \geq nt\right) \\
&\leq (\text{by union bound and assumption of finite } \Theta_a) \sum_{\theta\in\Theta} \mathbb{P}_{\theta_0}\left(\sum_{j=1}^{n} \log \frac{p_\theta(x_j)}{p_{\theta_0}(x_j)} \geq nt\right) \\
&= \sum_{\theta\in\Theta} \mathbb{P}_{\theta_0}\left(\exp\left\{\sum_{j=1}^{n} \log \frac{p_\theta(x_j)}{p_{\theta_0}(x_j)}\right\} \geq e^{nt}\right) \\
\leq (\text{by Markov's inequality}) \sum_{\theta\in\Theta} \underbrace{\mathbb{E}_{\theta_0}\left(\prod_{j=1}^{n} \frac{p_\theta(x_j)}{p_{\theta_0}(x_j)}\right)}_{=1} e^{-nt} &= \sum_{\theta\in\Theta} \underbrace{\int p_\theta(x_j) p_{\theta_0}(x_j)}_{=1}\, d\mu\; e^{-nt} = |\Theta| e^{-nt}.
\end{aligned}$$

where $|\Theta|$ is the cardinality of Θ. This implies

$$-\frac{2}{n}\log \mathbb{P}_{\theta_0}(\tilde{\Lambda}_n \geq nt) \geq 2t - 2\frac{\log|\Theta|}{n},$$

so by the definition of $\hat{\alpha}_n$ (plugging in $t = \tilde{\Lambda}/n$),

$$-\frac{2}{n}\log \hat{\alpha}_n \geq 2\frac{\tilde{\Lambda}_n}{n} - 2\frac{\log|\Theta|}{n}.$$

Note that

$$\tilde{\Lambda}_n = \log \frac{\max_{\theta\in\Theta} L_n(\theta)}{L_n(\theta_0)} \geq \log \frac{L_n(\theta')}{L_n(\theta_0)} \qquad \forall \theta' \in \Theta.$$

Take θ corresponding to the distribution of $X_1, \ldots, X_n$. Then

$$2\frac{\tilde{\Lambda}_n}{n} \geq \frac{2}{n}\sum_{i=1}^{n} \log \frac{p_\theta(x_j)}{p_{\theta_0}(x_j)} \xrightarrow{n\to\infty} KL(p_\theta || p_{\theta_0})$$

by the Law of Large Numbers. Hence as $n \to \infty$

$$\lim_{n\to\infty} \in \left(-\frac{2}{n}\log(\hat{\alpha}_n \geq 2KL(p_\theta||p_{\theta_0})\right) \implies \lim_{n\to\infty} -\frac{2}{n}\log(\hat{\alpha}_n) = 2KL(p_\theta||p_{\theta_0})$$

for any $\theta \in \Theta_a$.

□

1.9 Confidence Intervals

1.9.1 Connection Between Testing and Confidence Sets

Definition 1.9.1. Assume we have a statistical model $X \sim p_\theta, \theta \in \Theta$. A random set $C(X)$ is a $1-\alpha$ **confidence set** or **confidence region** ($\alpha \in (0,1)$) if and only if $\mathbb{P}_\theta(\theta \in C(X)) \geq 1-\alpha \ \forall \theta \in \Theta$.

Remark 48. An acceptance region of a non-randomized test ϕ is the set of all values of the test statistic for which the null hypothesis is not rejected.

Theorem 1.9.1.1. Let $A(\theta_0)$ be the acceptance region of a non-randomized test for testing $H_0 : \theta = \theta_0$; $H_a : \theta \in K(\theta_0)$ at level α (e.g., $K(\theta_0) = \{\theta \neq \theta_0\}$). Define $C(x) := \{\theta : x \in A(\theta)\}$. Then $C(X)$ is a $1-\alpha$ confidence set.

Proof. By definition, $x \in A(\theta) \iff \theta \in C(x)$. Therefore for all $\theta \in \Theta$ $\mathbb{P}_\theta(X \in A(\theta)) = \mathbb{P}_\theta(\theta \in C(X))$. The result follows since by assumption the test is of size α, so $\mathbb{P}_\theta(X \in A(\theta)) \geq 1-\alpha$.

□

Example 1.9.1. Suppose $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. Consider $H_0 : \mu = \mu_0, H_a : \mu \neq \mu_0$. Then (homework problem) the test

$$\phi = \begin{cases} 1, & \left|\frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sqrt{S^2}}\right| \geq t_{n-1,1-\alpha/2} \\ 0, \left|\frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sqrt{S^2}}\right| < t_{n-1,1-\alpha/2} & \end{cases}$$

is the UMP test.

(a) Find the acceptance region.

Solution

(a)

$$A(\mu_0) = \left\{x : \left|\frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sqrt{S^2}}\right| < t_{n-1,1-\alpha/2}\right\}$$

Note that

$$\left|\frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sqrt{S^2}}\right| < t_{n-1,1-\alpha/2} \iff -t_{n-1,1-\alpha/2} < \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sqrt{S^2}} < t_{n-1,1-\alpha/2}$$
$$\iff \overline{x}_n - \frac{\sqrt{s^2}t_{n-1,1-\alpha/2}}{\sqrt{n}} < \mu_0 < \overline{x}_n + \frac{\sqrt{s^2}t_{n-1,1-\alpha/2}}{\sqrt{n}}$$

So

$$C(X) = \left[\overline{X}_n - \frac{\sqrt{S^2}t_{n-1,1-\alpha/2}}{\sqrt{n}}, \overline{X}_n + \frac{\sqrt{S^2}t_{n-1,1-\alpha/2}}{\sqrt{n}}\right]$$

is a $1-\alpha$ confidence set.

Definition 1.9.2. A $(1-\alpha)$ confidence set $C(X)$ is **unbiased** if and only if $\mathbb{P}_\theta(\theta' \in C(X)) \leq 1-\alpha$ for all $\theta \neq \theta'$.

An unbiased confidence set can be obtained by inverting an unbiased confidence test.

Theorem 1.9.1.2. For each θ_0, let $A^*(\theta_0)$ be the acceptance region of a UMP unbiased test of size α for testing the null hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_a : \theta = \theta_0$. Let $C^*(X)$ be the corresponding confidence set. Then for any other unbiased $(1-\alpha)$ confidence set $C(X)$,

$$\mathbb{P}_\theta(\theta_0 \in C^*(X)) \leq \mathbb{P}_\theta(\theta_0 \in C(X)) \qquad \forall \theta \in \Theta.$$

Proof.

$$\mathbb{P}_\theta(\theta_0 \in C^*(X)) = \mathbb{P}_\theta(X \in A^*(\theta_0))$$

Note that for all $\theta \neq \theta_0$, $\mathbb{P}_\theta(X \in A^*(\theta_0))$ is the probability that the test will accept the null hypothesis given that the alternative hypothesis is true (probability of a Type II error, or 1 - the power). The UMP test minimizes this, so we have for any other rejection region $A(\theta_0) := \{x : \theta_0 \in C(x)\}$,

$$\mathbb{P}_\theta(X \in A^*(\theta_0)) \leq \mathbb{P}_\theta(X \in A(\theta_0)) = \mathbb{P}_\theta(\theta_0 \in C(X))$$

□

Theorem 1.9.1.3 (Pratt). Suppose $X \sim p_\theta, \theta \in \Theta$. Assume that $C(X) = [L(X), U(X)]$ where $L(x), U(x)$ are increasing functions. Then for all θ_0,

$$\mathbb{E}[U(X) - L(X)] = \mathbb{E}|C(X)| = \int_{\theta \neq \theta_0} \underbrace{\mathbb{P}_{\theta_0}(\theta \in C(X))}_{\text{probability of a Type II error}} \, d\theta$$

That is, if we know that $C(X)$ is an interval of this form, then in expectation the minimizer of the integral on the right (the interval corresponding to the UMP test) results in the shortest possible confidence region.

Proof.

$$\begin{aligned}\mathbb{E}_{\theta_0}[U(X) - L(X)] &= \int_{\mathbb{R}} [U(X) - L(X)] p_{\theta_0}(x) \, dx = \int_{\mathbb{R}} \int_{L(x)}^{U(x)} dt \; p_{\theta_0}(x) \, dx = \int_{\mathbb{R}} \int_{U^{-1}(t)}^{L^{-1}(t)} p_{\theta_0}(x) \, dx \, dt \\ &= \int_{\mathbb{R}} \mathbb{P}_{\theta_0}(U^{-1}(t) \leq x \leq L^{-1}(t)) \, dt = \int_{\mathbb{R}} \mathbb{P}_{\theta_0}(t \in C(x)) \, dt = \int_{t \neq \theta_0} \mathbb{P}_{\theta_0}(t \in C(x)) \, dt\end{aligned}$$

See Figure 1.5 for an explanation of the change of integrands in the third step (either way finds the probability-weighted area of the whole region between the two curves in all of $\mathbb{R}^2$).

□

Example 1.9.2 ("The Rule of Threes"). Suppose $X_1, \ldots, X_n \sim \text{Ber}(p)$, i.i.d. Assume that $\sum_{i=1}^n x_i = 0$. Find the 95% upper confidence bound for p. (it is approximately equal to $3/n$).

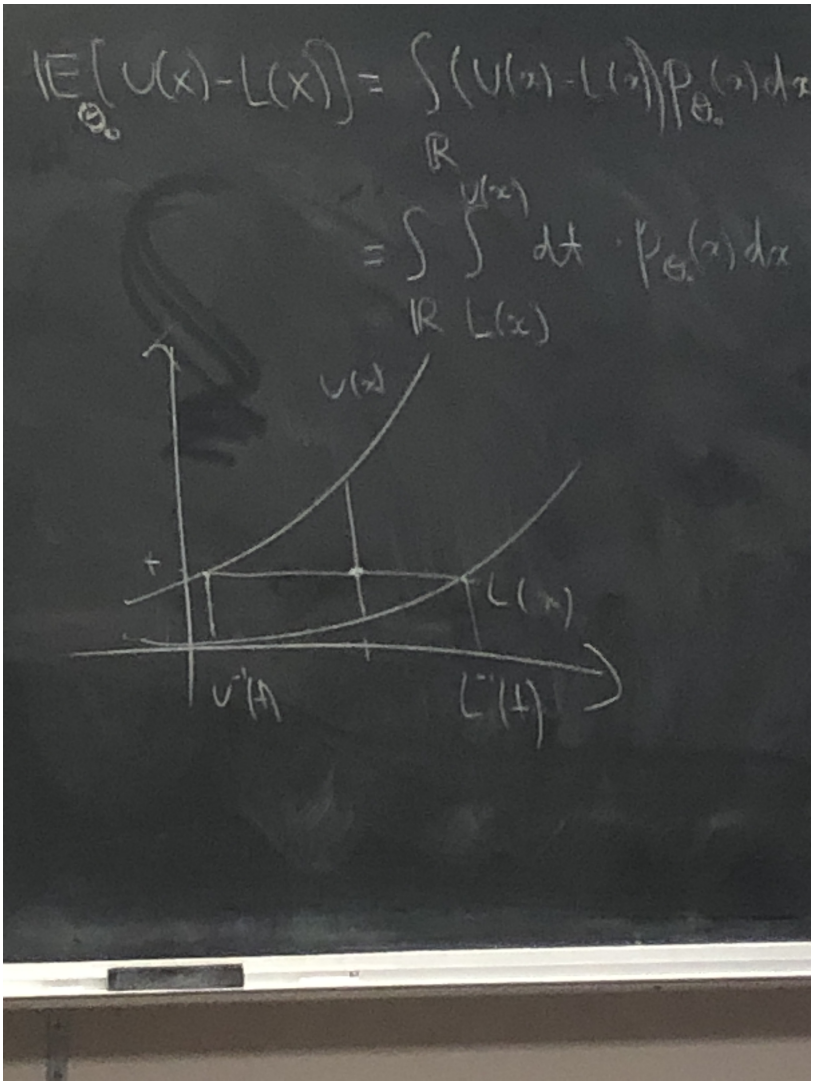

Figure 1.5: Visual depiction of change of integrands in proof of Theorem 1.9.1.3.

Solution

Consider testing $H_0 : p \geq p_0$ against $H_a : p < p_0$. We know that $X = \sum_{i=1}^n X_j$ is a complete sufficient statistic and $X \sim \text{Binomial}(n, p)$, so $\mathbb{P}_p(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$, a monotone likelihood ratio with respect to $T(X) = X$. Therefore the UMP (nonrandomized) test is

$$\phi^* = \begin{cases} 1, & X \leq c \\ 0, & X > c \end{cases}$$

(since this test is nonrandomized, it will only be UMP for specific values of α) where $c = c(p_0, \alpha) \in \mathbb{Z}_+$ solves

$$\alpha = \mathbb{P}(\text{Type I error}) = \mathbb{P}_{p_0}(X \leq c) = \sum_{j=0}^{c(p_0,\alpha)} \binom{n}{j} p_0^j (1-p_0)^{n-j}$$

To find the 95% upper confidence bound for p, we want to know for which values of p_0 is 0 in the acceptance region ($0 \in \text{AR}(p_0)$). Note that for this hypothesis test with null hypothesis $p \geq p_0$, the acceptance regions are nested, meaning that $p_1 < p_2 \implies \{\text{the rejection region for } p_1\} \subseteq \{\text{the rejection region for } p_2\} \iff \text{AR}(p_2) \subseteq \text{AR}(p_1)$. Hence we need to find the largest p such that $0 \in \text{AR}(p)$. This p must satisfy

$$\alpha < \mathbb{P}_p(X = 0) = \sum_{j=0}^{0} \binom{n}{j} p^j (1-p)^{n-j} \iff (1-p)^n > \alpha \iff p < 1 - \alpha^{1/n}.$$

Then α is small implies $\alpha^{1/n} = e^{\log(\alpha)/n} \approx 1 + \log(\alpha)/n$. Using this, we get $p < -\log(\alpha)/n$. When $\alpha = 0.05$, $-\log(\alpha) = \log(1/0.05) \approx 3$.

Exercise 8. What if we observe 1 success in n trials? Find an equation for a $(1-\alpha)$ upper confidence bound for p.

1.10 U-Statistics

1.10.1 Basic Definitions and Properties [Serfling, 1980, Sections 5.1 and 5.2], [DasGupta, 2008, Section 15.1]

While the limiting distribution of linear statistics is known from the canonical Central Limit Theorem (under suitable moment conditions), limiting distributions of nonlinear statistics are more difficult. However, a class of nonlinear statistics called **U-Statistics** have a common CLT.

Definition 1.10.1 (**U-Statistics; definition drawing from Section 5.1.1 of Serfling [1980, p. 190 of pdf, p. 172 of book] and DasGupta [2008, Chapter 15, p. 240 of pdf, p. 225 of book]**). Let $X_1, X_2, \ldots$ be independent observations on a distribution F. (They may be vector-valued, but usually we will consider the real-valued case.) Consider a parametric function $\theta = \theta(F)$ for which there is an unbiased estimator. That is, $\theta(F)$ may be represented as

$$\theta(F) = \mathbb{E}_F\left[h\left(X_1, \ldots, X_m\right)\right] = \int \cdots \int h(x_1, \ldots, x_m) dF(x_1) \ldots dF(x_m)$$

for some function $h = h(x_1, \ldots, x_m)$, called a **kernel**. Without loss of generality, we may assume h is symmetric[2]. Of course, $h(X_1, \ldots, X_m)$ is an unbiased estimator for $\theta(F)$, but a better unbiased estimate should exist if $n > m$ because $h(X_1, \ldots, X_m)$ does not use all of the sample data.

In particular, for any kernel h, the corresponding U**-Statistic of order** m **with kernel** h for estimation of θ on the basis of a sample $X_1, \ldots, X_n$ of size $n \geq m$ is obtained by averaging the kernel h symmetrically over the observations:

$$U_n = U(X_1, \ldots, X_n) := \frac{1}{\binom{n}{m}} \sum_c h(X_{i_1}, \ldots, X_{i_m}),$$

where $\sum_c$ denotes summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \ldots, i_m\}$ from $[n]$. Clearly, U_n is an unbiased estimate of θ.

Example 1.10.1 (**Example 15.2 in DasGupta [2008]**)**.** Let $m = 2$ and $h(x_1, x_2) = (x_1 - x_2)^2/2$. One can show that

$$\frac{1}{\binom{n}{2}} \sum_{i<j} \frac{(X_i - X_j)^2}{2} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

That is, the sample variance is a U-statistic.

A U-statistic may be represented as the result of conditioning the kernel on the order statistic; that is, for $n \geq m$,

$$U_n = \mathbb{E}\left[h(X_1, \ldots, X_m) \mid \boldsymbol{X}_{(n)}\right]$$

where $\boldsymbol{X}_{(n)}$ denotes the order statistic. This implies that any statistic $S = S(X_1, \ldots, X_n)$ for unbiased estimation of $\theta = \theta(F)$ may be improved by the corresponding U-statistic.

Theorem 1.10.1.1 (**Optimality property of** U**-statistics; Serfling [1980, Section 5.14, p. 194 of pdf, p. 176 of book]**)**.** Let $S = S(X_1, \ldots, X_n)$ be an unbiased estimator of $\theta(F)$ based on a sample $X_1, \ldots, X_n$ from the distribution F. Then the corresponding U-statistic is also unbiased and

$$\mathrm{Var}_F(U) \leq \mathrm{Var}_F(S),$$

with equality if and only if $\mathbb{P}_F(U = S) = 1$.

Proof. The kernel associated with S is

[2] To see why this is without loss of generality, suppose h is not symmetric. Then replace h with the symmetric kernel $(1/m!) \sum_p h(x_{i_1}, \ldots x_{i_m})$ where $\sum_p$ denotes summation over the $m!$ permutations $(i_1, \ldots, i_m)$ of $(1, \ldots, m)$.

$$\frac{1}{n!}\sum_{p} S(x_{i_1},\dots,x_{i_n}),$$

which in this case $m = n$ is the U-statistic associated with itself. That is, the U-statistic associated with S may be expressed as $U = \mathbb{E}(S \mid \boldsymbol{X}_{(n)})$. Therefore

$$\mathbb{E}_F(U^2) = \mathbb{E}_F\left(\left[\mathbb{E}_F\left(S \mid \boldsymbol{X}_{(n)}\right)\right]^2\right) \leq \mathbb{E}_F\left(\mathbb{E}_F\left(S^2 \mid \boldsymbol{X}_{(n)}\right)\right) = \mathbb{E}_F(S^2),$$

where the inequality follows from the fact that

$$\mathrm{Var}_F\left(S \mid \boldsymbol{X}_{(n)}\right) = \mathbb{E}_F\left(S^2 \mid \boldsymbol{X}_{(n)}\right) - \left[\mathbb{E}_F\left(S \mid \boldsymbol{X}_{(n)}\right)\right]^2 \geq 0.$$

with equality if and only if $\mathbb{E}_F(S \mid \boldsymbol{X}_{(n)})$ equals S with probability 1. Since $\mathbb{E}_F(U) = \mathbb{E}_F(S)$, the proof is complete.

□

Remark 49. Since $\boldsymbol{X}_{(n)}$ is sufficient for any family of distributions containing F, the U-statistic is the result of conditioning on a sufficient statistic. Therefore this result is simply a special case of the Rao-Blackwell theorem (Theorem 1.4.2.1). This further implies that if $\boldsymbol{X}_{(n)}$ is complete sufficient then the U-statistic is the MVUE of θ by Lehmann-Scheffe (Theorem 1.4.2.3).

Next we will find the variance of a U-statistic. In order to do this, we will establish some notation that will be useful.

Definition 1.10.2 (**Notation from Section 5.1.5 of Serfling [1980]**). For a symmetric kernel $h(x_1,\dots,x_m)$ satisfying $\mathbb{E}_F|h(X_1,\dots,X_m)| < \infty$, we define the associated functions

$$h_c(x_1,\dots,x_c) := \mathbb{E}_F\left[h(x_1,\dots,x_c,X_{c+1},\dots,X_m)\right] = \mathbb{E}_F\left[h(X_1,\dots,X_m) \mid (X_1,\dots,X_c) = (x_1,\dots,x_c)\right]$$

for each $c \in [m-1]$, and define $h_m := h$. Note that for $1 \leq c \leq m-1$,

$$h_c(x_1,\dots,x_c) = \mathbb{E}_F\left[h_{c+1}(x_1,\dots,x_c,X_{c+1})\right].$$

Recall $\theta(F) = \mathbb{E}_F[h(X_1,\dots,X_m)]$. Define

$$\tilde{h} := h - \theta(F)$$

and

$$\tilde{h}_c := h_c - \theta(F), \quad c \in [m].$$

Remark 50. Note that since by Definition 1.10.1

$$U_n = \frac{1}{\binom{n}{m}} \sum_c h(X_{i_1}, \dots, X_{i_m}),$$

it holds that

$$\begin{aligned} U_n - \theta &= \binom{n}{m}^{-1} \sum_c h(X_{i_1}, \dots, X_{i_m}) - \binom{n}{m}^{-1} \binom{n}{m} \theta \\ &= \binom{n}{m}^{-1} \sum_c \left[h(X_{i_1}, \dots, X_{i_m}) - \theta \right] \\ &= \binom{n}{m}^{-1} \sum_c \tilde{h}(X_{i_1}, \dots, X_{i_m}). \end{aligned} \tag{1.65}$$

Definition 1.10.3 (From Section 5.2.1 of Serfling [1980]). Consider a symmetric kernel satisfying $E_F[h^2(X_1, \dots, X_m)] < \infty$. Note that

$$\mathbb{E}_F \tilde{h}_c(X_1, \dots, X_c) = \mathbb{E}_F \left(\mathbb{E}_F \left[h(x_1, \dots, x_c, X_{c+1}, \dots, X_m) \right] - \theta \right) = 0, \qquad \forall c \in [m].$$

Define $\zeta_0 := 0$ and for $c \in [m]$

$$\zeta_c := \mathrm{Var}_F \left[h_c(X_1, \dots, X_c) \right] = \mathbb{E}_F \left[\tilde{h}_c^2(X_1, \dots, X_c) \right].$$

Example 1.10.2 (Example A in Section 5.2.1 of Serfling [1980]; extension of Example 1.10.1). Again, let $m = 2$ and $h(x_1, x_2) = (x_1 - x_2)^2/2$. We have already shown in Example 1.10.1 that this U-statistic corresponds to the sample variance; that is, $\mathbb{E}_F[h(X_1, X_2)] = \sigma^2(F) =: \theta(F)$. We have

$$\begin{aligned} \tilde{h}(x_1, x_2) &= h(x_1, x_2) - \sigma^2, \\ h_1(x) &= \mathbb{E}_F \left[h(X_1, X_2) \mid X_1 = x \right] = \mathbb{E}_F \left[\frac{1}{2}(x - X_2)^2 \right] = \frac{1}{2} \mathbb{E}_F \left[x^2 - 2xX_2 + X_2^2 \right] \\ &= \frac{1}{2} \left(x^2 - 2x\mu + \sigma^2 + \mu^2 \right), \\ \tilde{h}_1(x) &= \frac{1}{2} \left(x^2 - 2x\mu + \sigma^2 + \mu^2 \right) - \sigma^2 = \frac{1}{2} \left(x^2 - 2x\mu - \sigma^2 + \mu^2 \right) = \frac{1}{2} \left([x - \mu]^2 - \sigma^2 \right), \\ \zeta_1 &= \mathbb{E}_F \left[\tilde{h}_1^2(X_1) \right] = \mathbb{E}_F \left[\frac{1}{2} \left([X_1 - \mu]^2 - \sigma^2 \right) \right]^2 = \frac{1}{4} \mathrm{Var}_F \left[(X_1 - \mu)^2 \right] = \frac{1}{4} \left(\mu_4 - \sigma^4 \right), \\ \mathbb{E}(h^2) &= \frac{1}{4} \mathbb{E} \left[(X_1 - X_2)^2 \right]^2 = \frac{1}{4} \mathbb{E} \left[(X_1 - \mu) - (X_2 - \mu) \right]^4 \\ &= \frac{1}{4} \sum_{j=0}^{4} \binom{4}{j} (-1)^{4-j} \mathbb{E} \left[(X_1 - \mu)^j \right] \mathbb{E} \left[(X_2 - \mu)^{4-j} \right] = \frac{1}{4} \left(2\mu_4 + 6\sigma^4 \right), \\ \zeta_2 &= \mathbb{E}_F \left[\tilde{h}_2^2(X_1, X_2) \right] = \mathbb{E}_F \left[\left(h(X_1, X_2) - \sigma^2 \right)^2 \right] = \mathbb{E}_F \left[h^2(X_1, X_2) - 2\sigma^2 h(X_1, X_2) + \sigma^4 \right] \\ &= \mathbb{E}_F \left[h^2(X_1, X_2) \right] - \sigma^4 = \frac{1}{4} \left(2\mu_4 + 6\sigma^4 \right) - \sigma^4 = \frac{1}{2} \left(\mu_4 + \sigma^4 \right). \end{aligned}$$

Proposition 1.10.1.2 (**Problem 5.P.3(i) in Serfling [1980]**).

$$0 = \zeta_0 \leq \zeta_1 \leq \ldots \leq \zeta_m = \text{Var}_F(h) < \infty.$$

Proposition 1.10.1.3 (**Problem 5.P.4 in Serfling [1980]**). Consider two sets $\{a_1, \ldots, a_m\} \subset [n]$ and $\{b_1, \ldots, b_m\} \subset [n]$ of distinct integers. Let $c = |\{a_1, \ldots, a_m\} \cap \{b_1, \ldots, b_m\}|$ be the number of integers common to both. Then by symmetry of $\tilde{h}$ and independence of $\{X_1, \ldots, X_n\}$,

$$\mathbb{E}_F\left[\tilde{h}(X_{a_1}, \ldots, X_{a_m})\tilde{h}(X_{b_1}, \ldots, X_{b_m})\right] = \zeta_c.$$

Remark 51. Note also that the number of distinct choices for two such sets having exactly c elements in common is $\binom{n}{m}\binom{m}{c}\binom{n-m}{m-c}$. Also, since $\mathbb{E}_F\tilde{h} = 0$, we have that

$$\zeta_c = \text{Cov}\left(\tilde{h}(X_{a_1}, \ldots, X_{a_m}), \tilde{h}(X_{b_1}, \ldots, X_{b_m})\right) = \text{Cov}\left(h(X_{a_1}, \ldots, X_{a_m}), h(X_{b_1}, \ldots, X_{b_m})\right)$$

$$h_c(x_1, \ldots, x_c) = \mathbb{E}_F\left[h_{c+1}(x_1, \ldots, x_c, X_{c+1})\right].$$

Now we can find the variance of a U-statistic. We have

$$\begin{aligned}
\text{Var}_F(U_n) &= \mathbb{E}_F\left[(U_n - \theta)^2\right] \\
\text{(using (1.65))} \quad &= \mathbb{E}_F\left[\left(\binom{n}{m}^{-1}\sum_c \tilde{h}(X_{i_1}, \ldots, X_{i_m})\right)^2\right] \\
&= \binom{n}{m}^{-2}\sum_c\sum_c \mathbb{E}_F\left[\tilde{h}(X_{a_1}, \ldots, X_{a_m})\tilde{h}(X_{b_1}, \ldots, X_{b_m})\right] \\
\text{(by Proposition 1.10.1.3)} \quad &= \binom{n}{m}^{-2}\sum_{c=0}^{n}\binom{n}{m}\binom{m}{c}\binom{n-m}{m-c}\zeta_c \\
\text{(using } \zeta_0 = 0\text{)} \quad &= \binom{n}{m}^{-1}\sum_{c=1}^{n}\binom{m}{c}\binom{n-m}{m-c}\zeta_c. \qquad (1.66)
\end{aligned}$$

Lemma 1.10.1.4 (**Lemma A in Section 5.2.1 of Serfling [1980]**). The variance of U_n is given by (1.66) and satisfies

(i)
$$\frac{m^2}{n}\zeta_1 \leq \text{Var}_F(U_n) \leq \frac{m}{n}\zeta_m,$$

(ii)
$$(n+1)\text{Var}_F(U_{n+1}) \leq n\text{Var}_F(U_n), \qquad \text{and}$$

(iii)
$$\text{Var}_F(U_n) = \frac{m^2\zeta_1}{n} + \mathcal{O}(n^{-2}), \qquad n \to \infty.$$

1.10.2 Asymptotics [Serfling, 1980, Section 5.3], [DasGupta, 2008, Section 15.2]

Asymptotic theory for U-statistics is not straightforward since the summands are in general dependent. Hajek had the idea of projecting U onto the class of linear statistics of the form $n^{-1}\sum_{i=1}^n h(X_i)$. It turns out that the projection is the dominant part and determines the limiting distribution of U.

Definition 1.10.4 (**Hajek projection; Definition from Serfling [1980, Section 5.3.1, p. 206 of pdf, p. 188 of book] and Wager and Athey [2018, Section 3.3]**). Assume $\mathbb{E}_F|h| < \infty$. Then the **Hajek projection** of the U-statistic U_n is defined as

$$\hat{U}_n := \theta + \sum_{i=1}^n (\mathbb{E}_F[U_n \mid X_i] - \theta) = \sum_{i=1}^n \mathbb{E}_F[U_n \mid X_i] - (n-1)\theta$$

Note that it is a sum of i.i.d. random variables. In terms of the function $\tilde{h}_1$, we have

$$\hat{U}_n = \theta + \frac{m}{n}\sum_{i=1}^n \tilde{h}_1(X_i).$$

DasGupta [2008] uses a slightly different convention to define U statistics (their definition projects $U - \theta$ rather than U, so their definition equals the above definition minus θ).

Definition 1.10.5 (**Hajek projection; Definition 15.2 in DasGupta [2008]**). The **Hajek projection** of $U - \theta$ onto the class of statistics $\sum_{i=1}^n h(X_i)$ is

$$\hat{U} = \hat{U}_n = \sum_{i=1}^n \mathbb{E}_F(U - \theta \mid X_i).$$

Theorem 1.10.2.1 (**Theorem 15.1 in DasGupta [2008]**). Suppose the kernel h is twice integrable; i.e., $\mathbb{E}h^2 < \infty$. Then

$$\sqrt{n}(U - \hat{U}) \xrightarrow{p} 0$$

and

$$\sqrt{n}(U - \theta) \xrightarrow{D} \mathcal{N}(0, m^2\zeta_1).$$

Definition 1.10.6 (**Definition 6 in Wager and Athey [2018]**). U is $\nu(s)$**-incremental** if

$$\frac{\text{Var}\left[\hat{U}\right]}{\text{Var}\left[U\right]} \gtrsim \nu(s),$$

where

$$f(s) \gtrsim g(s) \qquad \iff \qquad \liminf_{s\to\infty}\left(\frac{f(s)}{g(s)}\right) \geq 1.$$

(That is, U is $\nu(s)$-incremental if

$$\liminf_{s\to\infty}\left(\frac{\text{Var}\left[\hat{U}\right]/\text{Var}\left[U\right]}{\nu(s)}\right) \geq 1.)$$

1.11 Influence Functions (Section 6.6.1 of Serfling [1980])

1.12 M-Estimators (Chapter 7 of Serfling [1980])

Many estimation methods are based on minimization of some function of the observations $\{X_i\}$ and the unknown parameter θ, e.g. the least squares estimator

$$\hat{\theta} = \arg\min_{\theta}\left\{d(\theta; X_1, \ldots, X_n)\right\} = \arg\min_{\theta}\left\{\sum_{i=1}^{n}(X_i-\theta)^2\right\}.$$

Similarly, the least absolute values estimator of θ is given by

$$\hat{\theta} = \arg\min_{\theta}\left\{\sum_{i=1}^{n}|X_i-\theta|\right\}.$$

Maximum likelihood estimation may be regarded as an approach of this type as well. Typically the problem of minimizing a function of data and a parameter reduces to a problem involving solving a system of equations for an estimator $\hat{\theta}$. Statistics given as solutions of equations are called **M-statistics.**

1.13 Distance Correlation

Bibliography

G. Casella and R. Berger. *Statistical Inference*. Duxbury Resource Center, June 2001. ISBN 0534243126.

A. DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer New York, 2008. ISBN 9780387759715. URL https://books.google.com/books?id=sX4_AAAAQBAJ.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005. ISBN 0-387-98864-5.

R. Serfling. *Approximation theorems of mathematical statistics*. Wiley series in probability and mathematical statistics : Probability and mathematical statistics. Wiley, New York, NY [u.a.], [nachdr.] edition, 1980. ISBN 0471024031. URL http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+024353353&sourceid=fbw_bibsonomy.

S. Wager and S. Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242, jul 2018. ISSN 0162-1459. doi: 10.1080/01621459.2017.1319839. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1319839.