Math Review Notes—Mathematical Statistics

Gregory Faletto

Contents

Last updated July 3, 2019

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], 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 (**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 $X_{(1)} := \min_{1 \leq i \leq n} X_i$, and for any $2 \leq i \leq n$, inductively define

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

so that

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

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

Definition 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 (**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 $i \geq 1$, define $p_i := \Pr(X \leq x_i)$. Then for any $1 \leq i, j \leq n$,

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

Proof. Note that $\{X_{(j)} \leq x_i\}$ 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. 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 for the cumulative distribution function.

□

Proposition 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 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 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 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 3 since the mean of such a beta distribution is $\frac{j}{n+1}$.

□

Proposition 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 1.3 (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 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. 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 1.4 (**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 1.5 (**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 1.6 (**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 6. 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 7 (**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.

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

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

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

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

b. Since X and Y are independent,

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

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

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

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

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

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

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

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

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

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

c.

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

Since individual observations from X and Y are independent,

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

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 1.7 (**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 8 (**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

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

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

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

Using the results from Proposition 6, we have

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

□

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.

□

Lemma 9. 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, $\langle v_i, v_j\rangle = 0 \;\forall\; 1 \le i < j \le m$.

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

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

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

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

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

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

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

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

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

□

Proposition 10 (**Proposition 4.7 in 541A notes**). Let $n \geq 2$ be an integer. Let $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ with $X_1 - \mu, \dots, X_n - \mu$ so that $\mu = 0$. Also divide by σ so that $\sigma = 1$. Note that $\overline{X}$ is independent of all random variables $X_2 - \overline{X}, \dots, X_n - \overline{X}$ by Lemma 9 because for example

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

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?)

(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 $S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$. 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 11 (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 12 (**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,

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

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

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 (Remark 4.10 in 541A notes)**.** If $X_1, \dots, X_n$ 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 13 (**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 14 (**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

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

Since the populations are normal, we know

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

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

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

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

□

Proposition 15 (**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.

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

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

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

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

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

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

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

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

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

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).

□

1.2.1 The Delta Method

Theorem 16 (Delta Method, Theorem 4.14 in 541A notes, 5.5.24 in Casella and Berger). 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,

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

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 (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 $t > 0$,

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

$$\iff \Pr(\sqrt{n}|h(Y_n - \theta)| > t) \leq \Pr(|Y_n - \theta| > t/\sqrt{n}) + \Pr(\sqrt{n}|h(Y_n - \theta)| > t \cap |Y_n - \theta| \leq t/\sqrt{n}). \tag{3}$$

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 (2).

□

Lemma 17. Under the same assumptions and notation as in Theorem 16,

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

Proof. We will examine the behavior of the right side of (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 18 (**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 19 (**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

$$f(Y_n) = f(\theta) + f'(\theta)(Y_n - \theta) + \frac{1}{2}f''(Z_n)(Y_n - \theta)^2 = f(\theta) + \frac{1}{2}f''(Z_n)(Y_n - \theta)^2 \tag{4}$$

where the second equality follows because $f'(\theta) = 0$. As in the proof of Theorem 16, $Z_n \xrightarrow{p} \theta$. 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 (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 20. 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. If F^{-1} does not exist, it can still work if you construct a generalized inverse of F as follows:

Proposition 21 (**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 1.1 (**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. 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 $U_1, \ldots, U_n$ with i.i.d. $X_1, \ldots, X_n$ 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 22 (**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 1.8 (**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 $Y := g(X_1, \dots, X_n)$ 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 $(X_1, \dots, X_n)$ 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 $X_1, \dots, X_n$.

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

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 $X_1, X_2, \dots, X_n \sim$ 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 1.2 (**Example 5.2 in 541A notes**). Let $X_1, \dots, X_n$ be a random sample of size n from a Bernoulli distribution with parameter $0 < \theta < 1$. Then $Y := X_1 = \dots + X_n$ is sufficient for θ.

Proposition 23 (**Example 5.2 in 541A notes**). Let $X_1, \dots, X_n$ be a random sample of size n from a Bernoulli distribution with parameter $0 < \theta < 1$. Let $Y := X_1 = \dots + X_n$. 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. 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 $x_1, \dots, x_n \in [0, 1]$. Let $0 \leq y \leq n$ be an integer. Then Y is binomial with parameters n and θ. We may assume $y = x_1 + \dots + x_n$, 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 (**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

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

we have

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

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

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

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

□

Theorem 24 (**Theorem 6.2.2 in Casella and Berger; 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 25 (**Factorization Theorem, Theorem 5.4 in 541A notes**). 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 $t : \mathbb{R}^n \to \mathbb{R}^k$, so $Y := t(X_1, \ldots, X_n)$ 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

$$f_\theta(x) = g_\theta(t(x))h(x), \quad \forall\, x \in \mathbb{R}^n, \quad \forall \theta \in \Theta. \tag{5}$$

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 (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

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

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

To use the Factorization Theorem (Theorem 5) 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 5) this $(\overline{x})$ is a sufficient statistic.

Remark. 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. 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 23, $(X_1 + \ldots + X_n)^2$ 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 26. 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 5).

□

Definition 1.9 (**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. 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 27 (**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 28 (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 5). So using the Factorization Theorem, Y is sufficient.

⋮

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 $Z = u(X_1, \dots, X_n)$ be any other sufficient statistic. We need to eventually show that Y is a function of Z. By the Factorization Theorem (Theorem 5), 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 29 (**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 $w(\theta); = (w_1(\theta), \dots, w_k(\theta))$. Assume that the following subset of $\mathbb{R}^k$ is k-dimensional:

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

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

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

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

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

Proof. First note that $t(X)$ is sufficient by the Factorization Theorem (Theorem 5) 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 28). 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 28.

□

Remark. 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 30 (**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

$$\sup_{A\subseteq\mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \sup_{A\subseteq\mathbb{R}} \{\mathbb{P}_1(A) - \mathbb{P}_2(A)\}. \tag{6}$$

(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 (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

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

$$\iff \int_{\mathbb{R}} |f_1(x) - f_2(x)|dx = \int_A (f_1(x) - f_2(x))dx + \int_{\mathbb{R}\setminus A} (f_2(x) - f_1(x))dx. \tag{7}$$

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 (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 (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

$$\sup_{A\subseteq\mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \sup_{A\subseteq\mathbb{Z}} \{\mathbb{P}_1(A) - \mathbb{P}_2(A)\}. \tag{8}$$

(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 (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

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

$$\iff \sum_{z\in\mathbb{Z}} |\mathbb{P}_1(z) - \mathbb{P}_2(z)| = \sum_{z\in A} \mathbb{P}_1(z) - \mathbb{P}_2(z) + \sum_{z\in\{\mathbb{Z}\setminus A\}} (\mathbb{P}_2(z) - \mathbb{P}_1(z)). \tag{9}$$

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 (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 (9), we have

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

□

1.3.3 Ancillary Statistics

Definition 1.10 (**Ancillary Statistic**). 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 $Y = t(X_1, \ldots, X_n), t : \mathbb{R}^n \to \mathbb{R}^m$ is **ancillary** for θ if the distribution of Y does not depend on θ.

Example 1.4 (**Example 5.15 from 541A notes**). Let $X_1, \ldots, X_n$ 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 5, Theorem 5.4 in Math 541A notes) (**usually the easiest way to prove sufficiency**) since if $t(x) := (x_{(1)}, \dots, x_{(n)})$, 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 28 (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.

$$\frac{f_\theta(x)}{f_\theta(y)} = \prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1 + (x_i - \theta)^2} \Bigg/ \prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1 + (y_i - \theta)^2} = \prod_{i=1}^{n} [1 + (y_i - \theta)^2] \Bigg/ \prod_{i=1}^{n} [1 + (x_i - \theta)^2] \tag{10}$$

Keep x, y fixed, $\theta \in \mathbb{R}$ variable. Then the likelihood ratio (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: $\theta = y_i \pm \sqrt{-i}, 1 \le i \le n$. Roots on bottom: $\theta = x_i \pm \sqrt{-i}, 1 \le i \le n$. So we can see this is true if and only if the vector $(x_1, \dots, x_n)$ is a permutation of $(y_1, \dots, y_n)$, 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 $Z_1, \dots, Z_n$ 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 $X_i = Z_i + \theta,\ \forall 1 \le i \le n, \forall \theta \in \mathbb{R}$. 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 $f : \mathbb{R}^n \to \mathbb{R}$, $f(x_1, \dots, x_n) = x_n - x_1 - c\ \forall\ (x_1, \dots, x_n) \in \mathbb{R}^n$. Then $\mathbb{E}_\theta f(Y) = 0,\ \forall \theta \in \Theta, Y = (X_{(1)}, \dots, X_{(n)})$. Note that $f \neq 0$ (in fact $f(Y) \neq 0$ with probability 1).

□

1.3.4 Complete Statistics

Definition 1.11 (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."

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

Example 1.5 (**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 5). 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 ($x = (x_1,\dots,x_n)$) is simply

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

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

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

Choose

$$h(x) = (1/3)^n, \qquad g_\theta(t(x)) = \begin{cases} 1 & t(x) \in \{\theta, \theta+1, \theta+2\} \times \{\theta, \theta+1, \theta+2\} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Then we have $f_\theta(x) = g_\theta(t(x))h(x)$, as desired. Now we will use Theorem 28 (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 $\theta \in \mathbb{Z}$, $X_i \in \mathbb{Z}\ \forall i \in \{1, \dots, n\}$, $\forall \theta \in \Theta$.) Using the expressions in (11), we can write the equation in (12) as

$$f_\theta(x) = c(x,y)f_\theta(z) \iff g_\theta(t(x))\left(\frac{1}{3}\right)^n = c(x,z)g_\theta(t(z))\left(\frac{1}{3}\right)^n \iff g_\theta(t(x)) = c(x,z)g_\theta(t(z)) \tag{12}$$

I will argue that the equality in (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 (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 1.12 (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:
>
> $$\text{For any } f : \mathbb{R}^k \to \mathbb{R} \text{ such that } \mathbb{E}_\theta f(Y) = 0\ \forall\ \theta \in \Theta, \text{ it holds that } f(Y) = 0.$$

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 1.6 (**Discrete RV example, Example 5.21 in Math 541A notes; return to Example 23**)**.** 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 1.7 (**Continuous RV example; return to Example 1.3**)**.** 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

$$0 = \int_{-\infty}^{\infty} f(y)e^{-y^2/2}e^{y\mu}dy, \qquad \forall \mu \in \mathbb{R} \tag{13}$$

If $f(y) \geq 0$, we are done since (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 (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. 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 31 (**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. 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 27 (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 32 (**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

$$\mathbb{E}_\theta(g(Y)) = \mathbb{E}_\theta(\mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}}) = \mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}}) = \mathbb{P}_\theta(Z \in B). \tag{14}$$

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 (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.

□

1.4 Point Estimation

Definition 1.13 (**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 1.14 (**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 1.15 (**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 1.16 (**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 $\mathrm{Var}_\theta(Y) \leq \mathrm{Var}_\theta(Z)$ for all $\theta \in \Theta$.

Remark. 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).

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 33 (**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. 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 1.17 (**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. 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 34 (**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 (**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. 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 1.8. 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 35 (**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 (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 35, 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 35, $h(Z)$ 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 35: 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

$$\mathbb{E}_\theta h(Z) = g(\theta) \tag{15}$$

by the above remark, (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 (15). For example if Z is binomial the left side of (15) will be the sum of a bunch of numbers. Find the h values that satisfy (15), if possible.

(3) **(Luck method):** Somehow guess the h such that (15) is satisfied.

Example 1.9. 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 5), 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 (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 (15) is satisfied).

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

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

So,

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

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

Example 1.10. 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 (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

$$(1-\theta)^{-n} = \sum_{j=0}^n h(j)\binom{n}{j}a^j = \theta(1-\theta)^{1-n} = \frac{a}{1+a}\left(\frac{1}{1+a}\right)^{1-n} = a(1+a)^{n-2} \tag{16}$$

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

$$\sum_{j=0}^n h(j)\binom{n}{j}a^j = a(1+a)^{n-2} = \text{ (RHS of (16), by binomial theorem) } a\sum_{j=0}^{n-2}\binom{n-2}{j}a^j = \sum_{j=1}^{n-1}\binom{n-2}{j} - 1a^j$$

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,

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

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

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

Example 1.11. Try Method (1). Suppose we have n independent samples $X_1, \ldots, X_n$ 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 35, $W := \mathbb{E}_\theta(Y \mid Z)$ is UMVU for σ^2. Note, $Y = 1$ when $X_1 = X_2 = 1$ and 0 otherwise. So

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

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

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

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

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

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

So, by Theorem 35,

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

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 35), $W_1 = \mathbb{E}_\theta(Y_1 \mid Z)$, $W_2 = \mathbb{E}(Y_2 \mid Z)$ 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 35) 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 36 (**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. 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 1.18 (**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.

Remark. 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 1.12 (**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 37 (**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,

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

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. 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 38 (**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 39 (**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 $\Theta \subseteq \mathbb{R}$. Let $t : \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X)$ be a statistic. For any $\theta \in \Theta$ let $g(\theta) := \mathbb{E}_\theta Y$. 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. For a one-parameter family of distributions, the equality case of Theorem 39 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 1.13 (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 39 (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 1.19 (**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

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

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

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

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. $t(X)$ can depend on Ψ.

Proposition 40 (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 1.14 (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 35, 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 5).

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

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

$$= \frac{1}{\sqrt{(2\pi)^n}} \left[(k+1)^n \right]^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T \left((k+1)\boldsymbol{I}_n \right)^{-1} \boldsymbol{x} \right) \tag{17}$$

which is the density of an n-dimensional multivariate Gaussian random variable with mean $\mathbf{0}$ 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

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

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

$$\implies \mathbb{E}\left(\left[\frac{k+1}{k}\mu_i - X_i\right]^2 \mid \boldsymbol{X}\right) = 1 \iff \mathbb{E}\left(\left[\frac{k+1}{k}\right]^2 \mu_i^2 + X_i^2 - 2\frac{k+1}{k}\mu_i X_i \mid \boldsymbol{X}\right) = 1 \tag{18}$$

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 $\boldsymbol{X}$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$, where

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

$$\iff \mu_i - X_i \mid \boldsymbol{X} \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \implies \mathbb{E}\left([\mu_i - X_i]^2 \mid \boldsymbol{X}\right) = 1 \iff \mathbb{E}\left(\mu_i^2 + X_i^2 - 2\mu_i X_i \mid \boldsymbol{X}\right) = 1 \quad (19)$$

So,

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

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

So,

$$\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}} = 1 + \boldsymbol{X}^T\boldsymbol{X} - \left(\boldsymbol{X}^T\boldsymbol{X} - n\right) = 1 + n. \quad (20)$$

(d) If the true distribution of $\boldsymbol{\mu}$ is the prior from part (b), then the marginal distribution of $\boldsymbol{X}$ is (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 $i \in \{1, \dots, n\}$.) Per the suggestion, let

$$\hat{D} = \hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}} = \left[\frac{k}{k+1}\right]^2 \left(1 + \boldsymbol{X}^T\boldsymbol{X}\right) - \left(\boldsymbol{X}^T\boldsymbol{X} - n\right) = \frac{k^2 - (k^2 + 2k + 1)}{(k+1)^2}\boldsymbol{X}^T\boldsymbol{X} + \left[\frac{k}{k+1}\right]^2 + n$$

$$= -\frac{2k+1}{(k+1)^2}\boldsymbol{X}^T\boldsymbol{X} + \left[\frac{k}{k+1}\right]^2 + n \quad (21)$$

Since from (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 1.21 (**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 1.15 (**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 1.22 (**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 1.16 (**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. The method of moments estimator is not necessarily unbiased.

1.4.6 Maximum likelihood estimator

Remark. Under some reasonable assumptions, the maximum likelihood estimator is consistent. See Keener for details.

Proposition 41 (**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. One example of a function whose log likelihood has no maximum value is $\exp(-e^{-\theta})$ (as in an extreme value distribution).

Remark. 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 42 (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** $n \to \infty$ **:**

$$\lim_{n\to\infty} \text{Var}\left(\frac{X}{n}\right) = \lim_{n\to\infty} \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) = \lim_{n\to\infty} \frac{1}{n^2} \cdot np(1-p) = \lim_{n\to\infty} \frac{p(1-p)}{n} = \boxed{0}$$

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 43 (**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 44 (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:

$$\text{Var}(\hat{\lambda}) \geq 1/\bigg(-n\mathbb{E}\bigg[\frac{\partial^2 \log(f(X;\lambda))}{\partial \lambda^2}\bigg]\bigg)$$

$$\log(f(X;\lambda)) = \log\left(\frac{\lambda^{X_i}e^{-\lambda}}{X_i!}\right) = X_i\log(\lambda) - \lambda - \log(X_i!)$$

$$\frac{\partial}{\partial \lambda}\log(f(X;\lambda)) = \frac{1}{\lambda}X_i - 1$$

$$\frac{\partial^2 \log(f(X;\lambda))}{\partial \lambda^2} = -\frac{1}{\lambda^2}X_i$$

$$\mathbb{E}\bigg[\frac{\partial^2 \log(f(X;\lambda^2))}{\partial \lambda}\bigg] = -\frac{1}{\lambda^2}\mathbb{E}(X_i) = -\frac{1}{\lambda^2}\lambda = -\frac{1}{\lambda}$$

$$\implies \text{Var}(\hat{\lambda}) \geq 1/\bigg(-n\mathbb{E}\bigg[\frac{\partial^2 \log(f(X;\lambda))}{\partial \lambda^2}\bigg]\bigg) = \frac{1}{n/\lambda} = \boxed{\frac{\lambda}{n} = \text{Var}(\hat{\lambda})}$$

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 45 (**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,

$$L = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n \exp(-\lambda \sum_{i=1}^{n} X_i)$$

$$\log(L) = n\log(\lambda) + -\lambda \sum_{i=1}^{n} X_i$$

$$\frac{\mathrm{d}\log(L)}{\mathrm{d}\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0$$

$$\implies \hat{\lambda} = n/\sum_{i=1}^{n} X_i = \frac{1}{\bar{X}}$$

□

Proposition 46 (**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

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) = \frac{2\sigma^4}{n}$$

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):

$$\lim_{n\to\infty} \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) = \lim_{n\to\infty} \frac{2\sigma^4}{n} = \boxed{0}$$

So this is a consistent estimator of σ^2.

d. Cramer-Rao lower bound:

$$\text{Var}(\hat{\theta}) \geq 1/\left(-n\mathbb{E}\left[\frac{\partial^2 \log(f(X;\theta))}{\partial\theta^2}\right]\right)$$

$$\log(f(X;\theta)) = \log\left[\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left[\frac{X_i}{\sigma}\right]^2\right)\right] = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}X_i^2(\sigma^2)^{-1}$$

$$\frac{\partial}{\partial\sigma^2}\left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}X_i^2(\sigma^2)^{-1}\right) = -\frac{1}{2\sigma^2} + \frac{1}{2}(X_i)^2(\sigma^2)^{-2}$$

$$\frac{\partial^2 \log(f(X;\theta))}{\partial(\sigma^2)^2} = \frac{1}{2}(\sigma^2)^{-2} - X_i^2(\sigma^2)^{-3}$$

$$\mathbb{E}\left[\frac{\partial^2 \log(f(X;\theta^2))}{\partial\theta}\right] = \mathbb{E}\left[\frac{1}{2}(\sigma^2)^{-2} - X_i^2(\sigma^2)^{-3}\right] = \frac{1}{2\theta^4} - \frac{1}{\theta^6}\mathbb{E}(X_i^2) = \frac{1}{2\theta^4} - \frac{\theta^2}{\theta^6} = -\frac{1}{2\theta^4}$$

$$\implies \text{Var}(\hat{\sigma^2}) \geq 1/\left(-n\mathbb{E}\left[\frac{\partial^2 \log(f(X;\theta))}{\partial\theta^2}\right]\right) = \frac{1}{n/(2\theta^4)} = \boxed{\frac{2\theta^4}{n}}$$

Therefore the variance of this estimator is equal to the Cramer-Rao lower bound.

Alternative solution (Math 541A):

$$I_X(\sigma) = I_{(X_1,\ldots,X_n)}(\sigma) = \text{(by Proposition 37 (Proposition 6.21 in Math 541A notes))}\ nI_{X_1}(\sigma)$$

$$= n\text{Var}_\sigma\left(\frac{\mathrm{d}}{\mathrm{d}\sigma}\log\left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(X_1-\mu)^2}{2\sigma^2}\right)\right.\right. \text{(by definition of Fisher information)}$$

$$= n\text{Var}_\sigma\left[-\frac{1}{\sigma} - \frac{\mathrm{d}}{\mathrm{d}\sigma}\left(\frac{(-X_1-\mu)^2}{2\sigma^2}\right)\right] = n\sigma^{-6}\text{Var}_\sigma\big((X_1-\mu)^2\big)$$

$$= n\sigma^{-6}\big(\mathbb{E}(X_1-\mu)^4 - \big[\mathbb{E}(X_1-\mu)^2\big]^2\big) = n\sigma^{-6}\sigma^4(3-1) = \frac{2n}{\sigma^2}.$$

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,

$$g'(\sigma) = \frac{2\sigma(n-1)}{n} \quad |g'(\sigma)|^2 = \frac{4\sigma^2(n-1)^2}{n^2}$$

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 1.17 (**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 47 (**Functional Equivariance of the MLE; Proposition 6.49 from Lecture Notes, Theorem 7.2.10 in Casella and Berger**). 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 (p.320).)

□

Proposition 48 (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.

$$f(x_i; \theta) = \frac{1}{2}e^{-|x_i - \theta|}$$

$$\implies L = \prod_{i=1}^{n} \frac{1}{2}e^{-|x_i-\theta|} = 2^{-n} \exp\left(-\sum_{i=1}^{n} |x_i - \theta| \right)$$

$$\implies \log(L) = -n\log(2) - \sum_{i=1}^{n} |x_i - \theta| \implies \frac{\mathrm{d}\log(L)}{\mathrm{d}\theta} = -\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta}|x_i - \theta| = -\sum_{i=1}^{n} \operatorname{sgn}(x_i - \theta)$$

since $\frac{\mathrm{d}|x|}{\mathrm{d}x} = \operatorname{sgn}(x)$. Next set this equal to 0 and solve:

$$-\sum_{i=1}^{n} \operatorname{sgn}(x_i - \hat{\theta}_{MLE}) = 0$$

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 (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 (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

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 1.23 (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 49 (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 50 (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 Hypothesis Testing

Proposition 51 (**Stats 100B Homework problem**)**.** Let $Y_1, Y_2, \ldots, Y_n$ be the outcomes of n independent Bernoulli trials. Then by the Neyman-Pearson lemma, 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.

□

References

G. Casella and R. Berger. *Statistical Inference*. Duxbury Resource Center, June 2001. ISBN 0534243126.