Math Review Notes—Mathematical Statistics

Gregory Faletto

Contents

Last updated February 21, 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, Statistics 100B at UCLA taught by Nicolas Christou, ISE 620 at USC taught by Sheldon Ross, and a few other sources I cite within the text.

1.1 Order Statistics (ISE 620)

Definition 1.1 (Order statistics (from Math 541A, more precise)). Let $X : \Omega \to \mathbb{R}$ be a random variable. Let $X_1, \ldots, X_n$ 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

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

so that

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

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

Definition 1.2 (Order statistics (from ISE 620, more informal)). Let $X_1, \ldots X_n \sim iid \; F$ with $F' = f$. Define $X_{(1)}$ as the smallest among $X_1, \ldots, X_n$, $X_{(2)}$ as the 2nd smallest, and so on, up to $X_{(n)}$, the largest of the group. We call $X_{(1)}, \ldots, X_{(n)}$ the **order statistics** of $X_1, \ldots, X_n$.

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

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

$\square$

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 $n - j$.

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

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

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

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

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.

□

1.2 Random Samples

Definition 1.3 (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 1.4 (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 1.5 (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 5. Suppose we have a random sample of size n from an i.i.d. distribution $X_1, X_2, \ldots, X_n$ with $\mathbb{E}(X_1) = \mu \ in \mathbb{R}$, $\operatorname{Var}(X_1) = \sigma^2 < \infty$. Then

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

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

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,

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

□

Proposition 6 (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. $\operatorname{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$.

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,

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

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

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

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

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

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

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

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

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

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

c.

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

Since individual observations from X and Y are independent,

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

and

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

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

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

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

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

□

Definition 1.6 (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 7 (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 $\mathbb{E}(X_1) = \mu\ in \mathbb{R}$, $\mathrm{Var}(X_1) = \sigma^2 < \infty$. Then $\mathbb{E}(S^2) = \sigma^2$. Further, S^2 is a consistent estimator of σ^2.

Proof. We have

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

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

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

Using the results from Proposition 5, 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.

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

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 8. 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

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

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

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

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

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

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

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

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

□

Proposition 9 (Proposition 4.7 in 541A notes). Let $n \geq 2$ be an integer. Let $X_1, \ldots, X_n$ be a random sample from the Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Let $\overline{X}$ be the sample mean and let S be the sample standard deviation. Then

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

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

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

Proof. (i) Replace $X_1, \ldots, X_n$ with $X_1 - \mu, \ldots, X_n - \mu$ so that $\mu = 0$. Also divide by σ so that $\sigma = 1$. Note that $\overline{X}$ is independent of all random variables $X_2 - \overline{X}, \ldots, X_n - \overline{X}$ by Lemma 8 because for example

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

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

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

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

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

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

(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 10 (**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 11 (**Proposition 4.9 in 541A notes**)**.** Let X be a standard Gaussian random variable. Let Y be a χ_p^2 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

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

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

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

$\vdots$ skipped this stuff in class proof

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

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

We use the change of variables formula:

$$\int\int_{\phi(U)} f(x,y) dx dy = \int\int_U f(\phi(a,b)) |\operatorname{Jac}\phi(a,b)| da db$$

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

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

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

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,

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

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 $X_1, \ldots, X_n$ is a random sample from a Gaussian distribution with mean $\mu \in \mathbb{R}$, standard deviation $\sigma < 0$, then

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

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

Proposition 12 (**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 13 (**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 14 (**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. $\mathrm{Var}(\bar{X} - \bar{Y})$

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

Proof. a.

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

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

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

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

b. Since X and Y are independent,

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

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

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

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

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

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

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

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

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

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

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 15 (**Delta Method, Theorem 4.14 in 541A notes, 5.5.24 in Casella and Berger**)**.** Let $\theta\in\mathbb{R}$. Let $Y_1,Y_2,\dots$ 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 $f'(\theta)\neq 0$ (note: this assumption seems to not be required in new proof?). 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 16. Under the same assumptions and notation as in Theorem 15,

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

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 17 (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 $\mathbb{E}(|X_n|^{1+\epsilon}) \leq c, \quad \forall\, n \geq 1$. 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 18 (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 15, $Z_n \xrightarrow{p} \theta$. 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 19. 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. See Exercise 4.20 in the Math 541A notes.

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

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

$\hat{}$this 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 1.7 (Sufficient Statistic; definition 5.1 in Math 541A notes). Suppose $X_1, \ldots, X_n$ is a sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of distributions (such as an exponential family). Let $g : \mathbb{R}^n \to \mathbb{R}^k$ so that $Y := g(X_1, \ldots, X_n)$ is a statistic. We say that Y is a **sufficient statistic** for θ if for every $y \in \mathbb{R}^k$ and for every $\theta \in \Theta$, the conditional distribution of $(X_1, \ldots, X_n)$ given $Y = y$ (with respect to probabilities given by f_θ) does not depend on θ. That is, Y provides sufficient information to determine θ from $X_1, \ldots, X_n$.

Remark. 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

$X_1, X_2, \ldots, X_n \sim$ i.i.d. f_θ where $\theta \in (\mu, \sigma^2), \mu \in \mathbb{R}$, where $f_\theta \sim \mathcal{N}(\mu, \sigma^2)$, then

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

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

Proposition 21 (**Example 5.2 in 541A notes**)**.** Let $X_1, \ldots, X_n$ be a random sample of size n from a Bernoulli distribution with parameter $0 < \theta < 1$. Let $Y := X_1 = \ldots + X_n$. Then

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

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

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

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

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

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

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

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

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

□

Example 1.3 (**Example 5.3 in 541A notes**)**.** Let $X_1, \ldots, X_n$ be a sample of size n from a Gaussian distribution with known variance $\sigma^2 > 0$ and unknown mean $\mu \in \mathbb{R}$. Then $Y := (X_1, \ldots, X_n)/n$ is a sufficient statistic for μ.

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

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

Since

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

we have

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

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

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

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

$\square$

Theorem 22 (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

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

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 have found the equation in the theorem 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 $r_\theta(z) := \mathbb{P}_\theta(t(X) = z) \;\; \forall\, z \in \mathbb{R}^k$. Also define $t^{-1}t(x) := \{y \in \mathbb{R}6n; t(y) = t(x)\} \;\; \forall x \in \mathbb{R}^n$. 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

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

Since this expression does not contain θ, Y is sufficient.

$\square$

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

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 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 21, $(X_1 + \ldots + X_n)^2$ 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 23. 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

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

So

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

□

Definition 1.8 (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.

Theorem 24 (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. So using the Factorization Theorem, Y is sufficient.

$$\vdots$$

Part we did in class on Friday 02/15: evidently (according to Goldstein) this shows that the statistic is minimal but not necessarily sufficient. Let $Z = u(X_1, \ldots, X_n)$ be any other sufficient statistic. We need to eventually show that Y is a function of Z. By the Factorization Theorem, 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 25 (**Proposition Larry Goldstein gave in class**)**.** If $\{f_\theta : \theta \in \Theta\}$, $\Theta \in \mathbb{R}^n$, f_θ all densities or all mass functions. Then a minimal sufficient statistic exists.

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.

□

1.4 Point Estimation

1.4.1 Method of Moments

1.4.2 Maximum likelihood estimator

Proposition 26 (**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 $\text{Var}(\hat{p} = p(1-p)/n$.

Proof. a. Bernoulli random variable:

$$P(X_i = x) = p^x (1-p)^{1-x}$$

Assuming independent samples,

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

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

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

$$\frac{1}{\hat{p}}\sum_{i=1}^n X_i = \frac{1}{1-\hat{p}}\sum_{i=1}^n (1-X_i)$$

$$(1-\hat{p})\sum_{i=1}^n X_i = \hat{p}\sum_{i=1}^n (1-X_i)$$

$$\sum_{i=1}^n X_i = \hat{p}\sum_{i=1}^n (X_i + 1 - X_i)$$

$$\sum_{i=1}^n X_i = n\hat{p}$$

$$\boxed{\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i}$$

b.

$$\text{Var}(\hat{p}) = \frac{1}{n^2}\text{Var}\bigg(\sum_{i=1}^{n} X_i\bigg)$$

Since X_i are independent, we can write this as

$$\frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i)$$

And since X_i is Bernoulli, $\text{Var}(X_i) = p(1-p)$.

$$= \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{1}{n^2}np(1-p) = \boxed{\frac{p(1-p)}{n}}$$

Cramer-Rao lower bound:

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

$$\frac{\partial}{\partial p}\log(p^x(1-p)^{1-x}) = \frac{\partial}{\partial p}(x\log(p) + (1-x)\log(1-p)) = \frac{x}{p} - \frac{1-x}{1-p}$$

$$\frac{\partial^2 \log(f(X;\theta))}{\partial\theta^2} = \frac{\partial}{\partial p}(\frac{x}{p} - \frac{1-x}{1-p}) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

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

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

$$\implies \text{Var}(\hat{p}) \geq 1/\bigg(-n(\frac{-1}{p(1-p)})\bigg) = \frac{p(1-p)}{n} = \text{Var}(\hat{p})$$

c. (1) **Unbiased:**

$$E\bigg(\frac{X}{n}\bigg) = \frac{np}{p} = p$$

(2) $\text{Var}(\hat{\theta}) \to 0$ **as** $n \to \infty$ **:**

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

Therefore $\frac{X}{n}$ is a consistent estimator of p.

d.

$$\mathbb{E}(\hat{\sigma^2}) = \mathbb{E}\left[\frac{1}{n}\left(\frac{X}{n}\big(1-\frac{X}{n}\big)\right)\right] = \mathbb{E}\left[\frac{X(n-X)}{n^3}\ \right] = \frac{1}{n^3}\mathbb{E}[nX-(X)^2] = \frac{1}{n^2}\mathbb{E}(X) - \frac{1}{n^3}\mathbb{E}(X^2)$$

$$= \frac{1}{n^2}\cdot np - \frac{1}{n^3}\big(\mathrm{Var}(X) + (E(X))^2\big) = \frac{p}{n} - \frac{np(1-p)}{n^3} - \frac{p^2n^2}{n^3} = \frac{pn - p + p^2 - p^2 n}{n^2}$$

$$= \frac{p(n-1+p-pn)}{n^2} = \frac{p(n-1)(1-p)}{n^2}$$

This is a biased estimator since $\mathrm{Var}(X) = \frac{p(1-p)}{n}$ (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 27 (**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 28 (**Stats 100B homework problem**). Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Poisson(λ) distribution. Then

(a) The maximum likelihood estimator of λ is

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

(b) The variance of the maximum likelihood estimator is

$$\text{Var}(\hat{\lambda}) = \frac{\lambda}{n}$$

(c) The maximum likelihood estimator is a minimum variance unbiased estimator.

(d) The maximum likelihood estimator is consistent.

Proof. (a)

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

Assuming the samples are independent,

$$L = \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} = \left(e^{-n\lambda} \lambda^{\sum_{i=1}^{n} X_i} \right) / \prod_{i=1}^{n} X_i!$$

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

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

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

(b)

$$\text{Var}(\hat{\lambda}) = \text{Var}(\frac{1}{n} \sum_{i=1}^{n} X_i) = \frac{1}{n^2} \text{Var}(\sum_{i=1}^{n} X_i)$$

Since X_i are i.i.d. this can be written as

$$= \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) = \frac{1}{n^2} \sum_{i=1}^{n} \lambda = \frac{\lambda}{n}$$

(c) Cramer-Rao lower bound:

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

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

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

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

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

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

Since $\text{Var}(\hat{\lambda})$ equals the Cramer-Rao lower bound, $\hat{\lambda}$ is a MVUE.

(d) We already know the MLE is unbiased. To show consistency, we show $\text{Var}(\hat{\theta}) \to 0$ as $n \to \infty$.

$$\lim_{n\to\infty} \text{Var}(\hat{\lambda}) = \lim_{n\to\infty} \frac{\lambda}{n} = \boxed{0}$$

Therefore $\hat{\lambda}$ is a consistent estimator of λ.

□

Proposition 29 (Stats 100B homework problem). 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 30 (Stats 100B homework problem). 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 unbiased.

(c) The maximum likelihood estimator of σ^2 has variance

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

and is consistent.

(d) The variance of the maximum likelihood estimator of σ^2 reaches the Cramer-Rao lower bound.

Proof. a. Since sample is i.i.d. $\mathcal{N}(0, \sigma^2)$:

$$L = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{X_i}{\sigma}\right]^2\right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n X_i^2\right)$$

$$\log(L) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2 = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - (\sigma^2)^{-1}\frac{1}{2}\sum_{i=1}^n X_i^2$$

$$\frac{\partial \log(L)}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + (\sigma^2)^{-2}\frac{1}{2}\sum_{i=1}^n X_i^2 = 0$$

$$\frac{\sum_{i=1}^n X_i^2}{(\hat{\sigma^2})^2} = \frac{n}{\hat{\sigma^2}}$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n X_i^2$$

b.

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

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

$$\mathbb{E}\bigg(\frac{1}{n}\sum_{i=1}^n X_i^2\bigg) = \frac{1}{n}\sum_{i=1}^n \ \sigma^2 = \frac{1}{n}n\sigma^2 = \boxed{\sigma^2}$$

c.

$$\mathrm{Var}\bigg(\frac{1}{n}\sum_{i=1}^n X_i^2\bigg) = \frac{1}{n^2}\mathrm{Var}\bigg(\sum_{i=1}^n X_i^2\bigg)$$

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

$$\mathrm{Var}\bigg(\frac{X_i^2}{\sigma^2}\bigg) = 2$$

$$\frac{1}{\sigma^4}\mathrm{Var}(X_i^2) = 2$$

$$\mathrm{Var}(X_i^2) = 2\sigma^4$$

Therefore

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

Test for consistency (already known that estimate is unbiased):

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

So this is a consistent estimator of σ^2.

d. Cramer-Rao lower bound:

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

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

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

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

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

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

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

□

1.4.3 Bayes estimator

1.4.4 EM Alglorithm

1.4.5 Comparison of estimators

1.5 Hypothesis Testing

Proposition 31 (**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:

$$\frac{p_0(1-p_a)}{p_a(1-p_0)} = \frac{p_0 - p_0p_a}{p_a - p_0p_a} < 1$$

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.

□