

Quiz Session 6: Multinomial Distribution and MLE for Simple Linear Regression

Yikun Zhang

November 16, 2022

Problem 1 (Midterm problem in Autumn 2018, 2019). Suppose that a sample of size n is taken at random with replacement from the population of all UW students. Each student in the sample is recorded as either male (M) or female (F) and as either a Washington resident (R) or non-resident (N). The data are presented in a two-way contingency table as:

$$\begin{array}{cc} & \begin{matrix} R & N \end{matrix} \\ \begin{matrix} M \\ F \end{matrix} & \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} \end{array}.$$

That is, X_{11} is the number of students in the sample who are both M and R , X_{12} is the number of students in the sample who are both M and N , etc. Thus, $X_{11} + X_{12} + X_{21} + X_{22} = n$. Let

$$\begin{array}{cc} & \begin{matrix} R & N \end{matrix} \\ \begin{matrix} M \\ F \end{matrix} & \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \end{array}$$

denote the corresponding population proportions, that is, p_{11} is the proportion of students in the UW population who are both M and R , p_{12} is the proportion of students in the population who are both M and N , etc. Hence, $p_{11} + p_{12} + p_{21} + p_{22} = 1$.

- What is the distribution of $(X_{11}, X_{12}, X_{21}, X_{22})$?
- What is the conditional distribution of $(X_{11}, X_{12})|X_{11} + X_{12}$? What is the conditional correlation $\text{Corr}(X_{11}, X_{12})|X_{11} + X_{12}$?
- Find $\text{Corr}(X_{11}, X_{12})|X_{11} + X_{12} + X_{21}$.
- What is the conditional distribution of $(X_{11}, X_{12})|X_{11} + X_{21}$? What is the conditional correlation $\text{Corr}(X_{11}, X_{12})|X_{11} + X_{21}$?

Solution. (a) Based on the “sampling with replacement” setting, we know that

$$(X_{11}, X_{12}, X_{21}, X_{22}) \sim \text{Multinomial}_4(n; p_{11}, p_{12}, p_{21}, p_{22}).$$

(b) According to the calculations and results in Section 7.2 of Lecture 7 notes (see also Chapter 7 in [Perlman 2020a](#)), we know that the conditional distribution of $(X_{11}, X_{12})|X_{11} + X_{12}$ is

$$(X_{11}, X_{12})|X_{11} + X_{12} \sim \text{Multinomial}_2\left(X_{11} + X_{12}; \frac{p_{11}}{p_{11} + p_{12}}, \frac{p_{12}}{p_{11} + p_{12}}\right).$$

Given $X_{11} + X_{12}$, X_{11} and X_{12} is negatively linear correlated, so $\text{Corr}(X_{11}, X_{12})|X_{11} + X_{12} = -1$.

(c) Recall from Section 7.2 of Lecture 7 notes that for any multinomial random vector $(X_1, \dots, X_k) \sim \text{Multinomial}_k(n; p_1, \dots, p_k)$, the covariance between any two components X_i, X_j with $1 \leq i \neq j \leq k$ can be computed as:

$$\text{Cov}(X_i, X_j) = \frac{1}{2} [\text{Var}(X_i + X_j) - \text{Var}(X_i) - \text{Var}(X_j)]$$

$$\begin{aligned}
&= \frac{1}{2} [n(p_i + p_j)(1 - p_i - p_j) - np_i(1 - p_i) - np_j(1 - p_j)] \\
&= -np_i p_j.
\end{aligned}$$

Hence, given that

$$(X_{11}, X_{12}, X_{21}) | X_{11} + X_{12} + X_{21} \sim \text{Multinomial}_3 \left(X_{11} + X_{12} + X_{21}; \frac{p_{11}}{p_{11} + p_{12} + p_{21}}, \frac{p_{12}}{p_{11} + p_{12} + p_{21}}, \frac{p_{21}}{p_{11} + p_{12} + p_{21}} \right),$$

we can obtain from the above results as:

$$\text{Cov}(X_{11}, X_{12}) | X_{11} + X_{12} + X_{21} = -(X_{11} + X_{12} + X_{21}) \left(\frac{p_{11}}{p_{11} + p_{12} + p_{21}} \right) \left(\frac{p_{12}}{p_{11} + p_{12} + p_{21}} \right).$$

Additionally,

$$\text{Var}(X_{11} | X_{11} + X_{12} + X_{21}) = (X_{11} + X_{12} + X_{21}) \left(\frac{p_{11}}{p_{11} + p_{12} + p_{21}} \right) \left(\frac{p_{12} + p_{21}}{p_{11} + p_{12} + p_{21}} \right)$$

and

$$\text{Var}(X_{12} | X_{11} + X_{12} + X_{21}) = (X_{11} + X_{12} + X_{21}) \left(\frac{p_{12}}{p_{11} + p_{12} + p_{21}} \right) \left(\frac{p_{11} + p_{21}}{p_{11} + p_{12} + p_{21}} \right).$$

Therefore,

$$\begin{aligned}
\text{Corr}(X_{11}, X_{12}) | X_{11} + X_{12} + X_{21} &= \frac{\text{Cov}(X_{11}, X_{12}) | X_{11} + X_{12} + X_{21}}{\sqrt{\text{Var}(X_{11} | X_{11} + X_{12} + X_{21}) \cdot \text{Var}(X_{12} | X_{11} + X_{12} + X_{21})}} \\
&= -\sqrt{\frac{p_{11} p_{12}}{(p_{12} + p_{21})(p_{11} + p_{21})}}.
\end{aligned}$$

(d) Notice that the conditional distribution of $(X_{11}, X_{12}) | X_{11} + X_{21}$ is identical to the conditional distribution of $(X_{11}, X_{12}) | (X_{11} + X_{21}, X_{12} + X_{22})$. By the results in Section 7.2 of Lecture notes, we know that X_{11} and X_{12} is conditionally independent given $(X_{11} + X_{21}, X_{12} + X_{22})$. Thus, the conditional distribution of $(X_{11}, X_{12}) | X_{11} + X_{21}$ is the product of two independent binomial distribution as:

$$(X_{11}, X_{12}) | X_{11} + X_{21} \sim \text{Binomial} \left(X_{11} + X_{12}, \frac{p_{11}}{p_{11} + p_{21}} \right) \otimes \text{Binomial} \left(X_{12} + X_{22}, \frac{p_{12}}{p_{12} + p_{22}} \right).$$

Finally, the conditional correlation $\text{Corr}(X_{11}, X_{12}) | X_{11} + X_{21}$ is zero. \square

Problem 2. Assume that we want to estimate θ from some data $\mathbf{X} = (X_1, \dots, X_n)$, where $X_i \sim P_\theta$ are independent and identically distributed. An estimator $\hat{\theta} = T(\mathbf{X})$ has been constructed and we quantify its performance via the squared loss function $L(T(\mathbf{X}), \theta) = (T(\mathbf{X}) - \theta)^2$, where T is some deterministic function. The risk function $R(T(\mathbf{X}), \theta)$ is defined as the expected value of the loss function as (see also Section 7.1 in [Hogg et al. 2012](#)):

$$R(T(\mathbf{X}), \theta) = \mathbb{E}[L(T(\mathbf{X}), \theta)] = \mathbb{E}[(T(\mathbf{X}) - \theta)^2]$$

Show that $R(T(\mathbf{X}), \theta) = \text{Bias}^2(T(\mathbf{X})) + \text{Var}(T(\mathbf{X}))$ where $\text{Bias}(T(\mathbf{X})) = \mathbb{E}[T(\mathbf{X})] - \theta$.

Proof. By direct calculations,

$$(T(\mathbf{X}) - \theta)^2 = (T(\mathbf{X}) - \mathbb{E}[T(\mathbf{X})] + \mathbb{E}[T(\mathbf{X})] - \theta)^2$$

$$= (T(\mathbf{X}) - \mathbb{E}[T(\mathbf{X})])^2 + (\mathbb{E}[T(\mathbf{X})] - \theta)^2 + 2(T(\mathbf{X}) - \mathbb{E}[T(\mathbf{X})]) (\mathbb{E}[T(\mathbf{X})] - \theta).$$

Now, taking the expectation yields that

$$\begin{aligned} R(T(\mathbf{X}), \theta) &= \mathbb{E}[(T(\mathbf{X}) - \theta)^2] \\ &= \mathbb{E}[(T(\mathbf{X}) - \mathbb{E}[T(\mathbf{X})])^2] + \mathbb{E}[(\mathbb{E}[T(\mathbf{X})] - \theta)^2] + \mathbb{E}[2(T(\mathbf{X}) - \mathbb{E}[T(\mathbf{X})]) (\mathbb{E}[T(\mathbf{X})] - \theta)] \\ &= \text{Var}(T(\mathbf{X})) + (\mathbb{E}[T(\mathbf{X})] - \theta)^2 + 2(\mathbb{E}[T(\mathbf{X})] - \theta) \cdot \underbrace{\mathbb{E}[T(\mathbf{X}) - \mathbb{E}[T(\mathbf{X})]]}_{=0} \\ &= \text{Var}(T(\mathbf{X})) + (\mathbb{E}[T(\mathbf{X})] - \theta)^2 \\ &= \text{Var}(T(\mathbf{X})) + \text{Bias}^2(T(\mathbf{X})) \end{aligned}$$

The result follows. \square

Problem 3 (MLE of simple linear regression; Exercises 7.19–7.21 in [Casella and Berger 2002](#)). Suppose that the random variables Y_1, \dots, Y_n satisfy

$$Y_i = \beta x_i + \epsilon_i, \quad i = 1, \dots, n$$

where x_1, \dots, x_n are fixed constants and $\epsilon_1, \dots, \epsilon_n \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$, $\sigma^2 > 0$ is unknown.

- (a) Find the maximum likelihood estimators (MLEs) of β and σ^2 . Show that the MLE $\hat{\beta}_{MLE}$ of β is unbiased. What is its variance?
- (b) Show that $\sum_{i=1}^n Y_i / \sum_{i=1}^n x_i$ is also an unbiased estimator of β . What is its variance? Show that it is larger than the variance in (a).
- (c) Show that $\frac{1}{n} \sum_{i=1}^n (Y_i / x_i)$ is also an unbiased estimator of β . What is its variance? Compare it to the variances in (a) and (b).

Solution. (a) The log-likelihood is given by

$$\begin{aligned} \log L(\beta, \sigma^2 | \mathbf{Y}) &= \sum_{i=1}^n \log p(Y_i | \beta, \sigma^2) \\ &= \sum_{i=1}^n \left[-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (Y_i - \beta x_i)^2 \right] \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta x_i)^2, \end{aligned}$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$. Taking the partial derivatives with respect to β and σ^2 yields that

$$\begin{aligned} \frac{\partial}{\partial \beta} \log L(\beta, \sigma^2 | \mathbf{Y}) &= \frac{1}{\sigma^2} \sum_{i=1}^n x_i (Y_i - \beta x_i), \\ \frac{\partial}{\partial \lambda} \log L(\beta, \sigma^2 | \mathbf{Y}) &= \frac{n}{2\lambda} - \frac{1}{2} \sum_{i=1}^n (Y_i - \beta x_i)^2, \end{aligned}$$

where we denote $\lambda = \frac{1}{\sigma^2}$. Given that $\frac{\partial^2}{\partial \beta^2} \log L(\beta, \sigma^2 | \mathbf{Y}) = -\frac{1}{\sigma^2} \sum_{i=1}^n x_i^2 < 0$, the log-likelihood $\log L(\beta, \sigma^2 | \mathbf{Y})$ is strictly concave with respect to β for any fixed $\sigma^2 > 0$. Hence, the solution $\beta^* = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$ to the equation

$\frac{\partial}{\partial \beta} \log L(\beta, \sigma^2 | \mathbf{Y}) = 0$ did maximize the log-likelihood $\frac{\partial}{\partial \beta} \log L(\beta, \sigma^2 | \mathbf{Y})$ for any fixed $\sigma^2 > 0$. The partial maximum is

$$\max_{\beta} \log L(\beta, \sigma^2 | \mathbf{Y}) = -\frac{n}{2} \log(2\pi) + \frac{n}{2} \log(\lambda) - \frac{\lambda}{2} \sum_{i=1}^n (Y_i - \beta^* x_i)^2,$$

and $\frac{\partial^2}{\partial \lambda^2} \log L(\beta, \sigma^2 | \mathbf{Y}) = -\frac{n}{2\lambda^2} < 0$. It implies that $\max_{\beta} \log L(\beta, \sigma^2 | \mathbf{Y}) = \log L(\beta^*, \frac{1}{\lambda} | \mathbf{Y})$ is strictly concave with respect to λ and has its unique maximum at $\hat{\lambda} = \frac{n}{\sum_{i=1}^n (Y_i - \beta^* x_i)^2}$ by solving $\frac{\partial}{\partial \lambda} \log L(\beta^*, \frac{1}{\lambda} | \mathbf{Y}) = 0$. Therefore, $(\beta^*, \hat{\lambda})$ jointly maximizes the log-likelihood $\log L(\beta, \sigma^2 | \mathbf{Y})$, so it leads to the MLEs as:

$$\hat{\beta}_{MLE} = \beta^* = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \quad \text{and} \quad \hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \beta^* x_i)^2 = \frac{(\sum_{i=1}^n x_i^2) (\sum_{i=1}^n Y_i^2) - (\sum_{i=1}^n x_i Y_i)^2}{n (\sum_{i=1}^n x_i^2)}.$$

Finally, the expectation and variance of $\hat{\beta}_{MLE}$ are given by

$$\begin{aligned} \mathbb{E}(\hat{\beta}_{MLE}) &= \mathbb{E}\left(\frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2}\right) \\ &= \frac{1}{\sum_{i=1}^n x_i^2} \sum_{i=1}^n x_i \mathbb{E}(Y_i) \\ &= \frac{1}{\sum_{i=1}^n x_i^2} \sum_{i=1}^n \beta x_i^2 \\ &= \beta, \end{aligned}$$

showing that $\hat{\beta}_{MLE}$ is an unbiased estimator of β , and

$$\begin{aligned} \text{Var}(\hat{\beta}_{MLE}) &= \text{Var}\left(\frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2}\right) \\ &= \frac{1}{(\sum_{i=1}^n x_i^2)^2} \sum_{i=1}^n x_i^2 \text{Var}(Y_i) \\ &= \frac{\sigma^2}{\sum_{i=1}^n x_i^2}. \end{aligned}$$

Notes: Indeed, the distribution of $\hat{\beta}_{MLE}$ is $N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right)$.

(b) By direct calculation, we have that

$$\begin{aligned} \mathbb{E}\left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right) &= \frac{1}{\sum_{i=1}^n x_i} \sum_{i=1}^n \mathbb{E}(Y_i) \\ &= \frac{1}{\sum_{i=1}^n x_i} \sum_{i=1}^n \beta x_i \\ &= \beta \end{aligned}$$

showing that $\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}$ is also an unbiased estimator of β , and

$$\begin{aligned} \text{Var}\left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right) &= \frac{1}{(\sum_{i=1}^n x_i)^2} \sum_{i=1}^n \text{Var}(Y_i) \\ &= \frac{\sigma^2}{(\sum_{i=1}^n x_i)^2 / n} \end{aligned}$$

The denominator here is $\frac{(\sum_{i=1}^n x_i)^2}{n} = n(\bar{x}_n)^2$ where $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$. Given that

$$0 \leq \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \sum_{i=1}^n x_i^2 - n(\bar{x}_n)^2,$$

it implies that

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

and thus, $\text{Var}(\hat{\beta}_{\text{MLE}}) \leq \text{Var}\left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right)$. Here, the equality holds only when $x_1 = \dots = x_n$.

(c) By direct calculations, we have that

$$\begin{aligned} \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n \frac{Y_i}{x_i}\right) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left(\frac{Y_i}{x_i}\right) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\beta x_i}{x_i} \\ &= \beta \end{aligned}$$

showing that $\frac{1}{n} \sum_{i=1}^n \frac{Y_i}{x_i}$ is also an unbiased estimator of β , and

$$\begin{aligned} \text{Var}\left(\frac{1}{n} \sum_{i=1}^n \frac{Y_i}{x_i}\right) &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}\left(\frac{Y_i}{x_i}\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \frac{1}{x_i^2} \text{Var}(Y_i) \\ &= \frac{\sigma^2}{n^2} \sum_{i=1}^n \frac{1}{x_i^2} \end{aligned}$$

By the Cauchy-Schwarz inequality, we know that

$$\left(\sum_{i=1}^n \frac{1}{x_i^2}\right) \left(\sum_{i=1}^n x_i^2\right) \geq \left(\sum_{i=1}^n x_i \cdot \frac{1}{x_i}\right)^2 = n^2,$$

and thus,

$$\frac{\sigma^2}{n^2} \left(\sum_{i=1}^n \frac{1}{x_i^2}\right) \geq \frac{\sigma^2}{\sum_{i=1}^n x_i^2},$$

i.e., $\text{Var}\left(\frac{1}{n} \sum_{i=1}^n \frac{Y_i}{x_i}\right) \geq \text{Var}(\hat{\beta}_{MLE})$.

Notice, however, that the variances of estimators in (b) and (c) are *not comparable*. That is, we cannot conclude whether $\frac{n\sigma^2}{(\sum_{i=1}^n x_i)^2}$ or $\frac{\sigma^2}{n^2} \sum_{i=1}^n \frac{1}{x_i^2}$ is bigger without further information on $x_i, i = 1, \dots, n$. Consider the following two cases:

- *Case 1:* Take $x_i = \frac{(-1)^i}{\sqrt{n}}$ for $i = 1, \dots, n$ so that $|\sum_{i=1}^n x_i| \leq \frac{1}{\sqrt{n}} \rightarrow 0$ as $n \rightarrow \infty$. It implies that

$$\text{Var}\left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right) = \frac{n\sigma^2}{(\sum_{i=1}^n x_i)^2} \geq n^3 \sigma^2 \rightarrow \infty \quad \text{as } n \rightarrow \infty,$$

while

$$\text{Var} \left(\frac{1}{n} \sum_{i=1}^n \frac{Y_i}{x_i} \right) = \frac{\sigma^2}{n^2} \sum_{i=1}^n \frac{1}{x_i^2} = \sigma^2 < \infty.$$

In this case,

$$\text{Var} \left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i} \right) = \frac{n\sigma^2}{(\sum_{i=1}^n x_i)^2} > \frac{\sigma^2}{n^2} \sum_{i=1}^n \frac{1}{x_i^2} = \text{Var} \left(\frac{1}{n} \sum_{i=1}^n \frac{Y_i}{x_i} \right).$$

- *Case 2:* When $x_i, i = 1, \dots, n$ are all positive and not identical, we can apply the Jensen's inequality to the convex function $f(u) = \frac{1}{u^2}$ for $u > 0$ to obtain that

$$\frac{1}{\left(\frac{1}{n} \sum_{i=1}^n x_i\right)^2} \leq \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{x_i^2}\right),$$

which in turn shows that $\frac{n\sigma^2}{(\sum_{i=1}^n x_i)^2} \leq \frac{\sigma^2}{n^2} \sum_{i=1}^n \frac{1}{x_i^2}$. Thus, in this case, $\text{Var} \left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i} \right) \leq \text{Var} \left(\frac{1}{n} \sum_{i=1}^n \frac{Y_i}{x_i} \right)$ and the equality does not hold when $x_i, i = 1, \dots, n$ are not the same.

Notes: It is WRONG to apply the Jensen's equality to the function $f(u) = \frac{1}{u^2}$ without restricting to \mathbb{R}^+ , because $f(u)$ is not convex in \mathbb{R} ; see Figure 1 below. \square

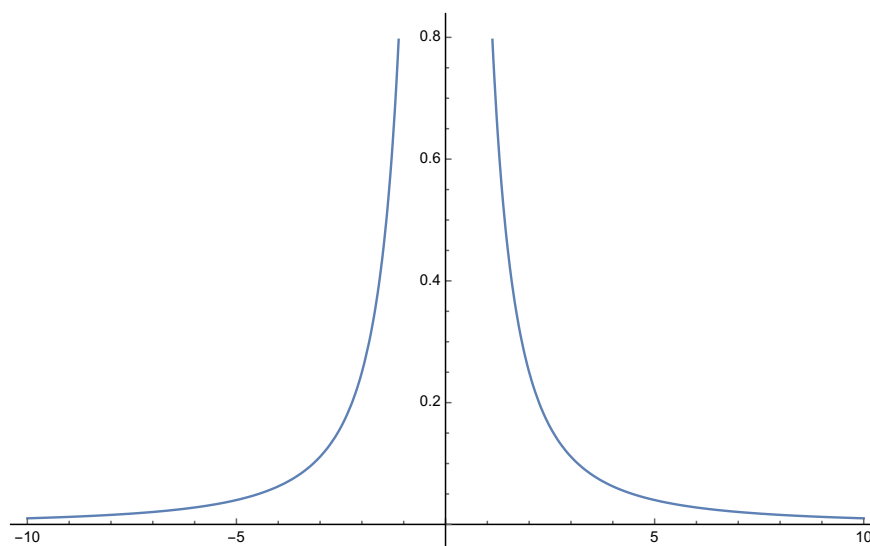


Figure 1: Plot of $\frac{1}{x^2}$ on $x \in [-10, 10]$.

Remark 1. There are two different arguments about why $\hat{\beta}_{MLE}$ must attain the minimum variance among all the unbiased estimators:

- On the one hand, by maximizing the log-likelihood under Gaussian and homoscedastic assumptions on the errors $\epsilon_i, i = 1, \dots, n$, $\hat{\beta}_{MLE}$ coincides with the ordinary least square solution. According to Gauss-Markov theorem¹, $\hat{\beta}_{MLE}$ is the best linear unbiased estimator (BLUP) of β under the mean square error criterion (see Problem 2).
- On the other hand, $\hat{\beta}_{MLE}$ is a function of the two-dimensional complete and sufficient statistic for (β, σ^2) as:

$$\left(\sum_{i=1}^n Y_i^2, \sum_{i=1}^n x_i Y_i \right),$$

¹See https://en.wikipedia.org/wiki/Gauss-Markov_theorem.

so it is also the uniformly minimum-variance unbiased estimator (UMVUE); see Section 12.3 in [Perlman \[2020b\]](#).

A statistic $T(\mathbf{X})$ is called sufficient for θ if the conditional distribution of the sample \mathbf{X} given the value of $T(\mathbf{X})$ does not depend on θ ; see Definition 6.2.1 in [Casella and Berger \[2002\]](#). Also, $T(\mathbf{X})$ is called complete if $\mathbb{E}_\theta g(T) = 0$ for all θ implies $\mathbb{P}_\theta(g(T) = 0) = 1$ for all θ , where \mathbb{E}_θ and \mathbb{P}_θ are taken with respect to the distribution of $T(\mathbf{X})$; see Definition 6.2.21 in [Casella and Berger \[2002\]](#). These concepts will be discussed in detail during STAT 513.

References

- G. Casella and R. Berger. *Statistical Inference*. Duxbury advanced series. Thomson Learning, 2nd ed. edition, 2002.
- R. Hogg, J. McKean, and A. Craig. *Introduction to Mathematical Statistics*. Pearson Education, 7th edition, 2012.
- M. Perlman. Probability and Mathematical Statistics I (STAT 512 Lecture Notes), 2020a. URL <https://sites.stat.washington.edu/people/mdperlma/STAT%20512%20MDP%20Notes.pdf>.
- M. Perlman. Probability and Mathematical Statistics I (STAT 512 Lecture Notes), 2020b. URL <https://sites.stat.washington.edu/people/mdperlma/STAT%20513%20MDP%20Notes.pdf>.