Problem Set 1, Solutions
Math Camp 2025, UCSB
Instructors: Camilo Abbate and Sofía Olguín

1. **Question 1**

   Consider a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from a distribution with a mean μ and variance σ^2. Consider the variance estimator $\hat{\sigma}_n^2$ defined as:

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

   where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean.

   (a) Show that $\hat{\sigma}_n^2$ is a biased estimator of σ^2.
   We have shown in class that $\hat{\sigma}_n^2$ can be expressed as:

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

   Taking expectations, we can write:

   $$\begin{aligned}
   \mathbb{E}[\hat{\sigma}_n^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right] \\
   &= \mathbb{E}[X_i^2] - \mathbb{E}[\bar{X}_n^2] \\
   &= \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right) \\
   &= \sigma^2 - \frac{\sigma^2}{n} \\
   &= \left(1 - \frac{1}{n}\right)\sigma^2
   \end{aligned}$$

   (b) What is the bias?
   The bias is given by:

   $$\text{Bias}(\hat{\sigma}_n^2) = \mathbb{E}[\hat{\sigma}_n^2] - \sigma^2 = \left(1 - \frac{1}{n}\right)\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

   (c) How can you modify $\hat{\sigma}_n^2$ to make it unbiased?
   To make the estimator unbiased, we can multiply it by $\frac{n}{n-1}$:

   $$S_n^2 = \frac{n}{n-1}\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

This new estimator S_n^2 is unbiased because:

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

2. **Question 2**

   In class we saw that the MSE of an estimator $\hat{\theta}$ for a parameter θ can be decomposed as:

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

   Let's circle back to the comparison of the estimators

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

   and

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

   Recall from the Lecture notes that:

$$\begin{aligned} \mathrm{Var}(S_n^2) &= \frac{2}{n-1}\sigma^4 \qquad \text{and} \qquad \mathrm{Var}(\hat{\sigma}_n^2) = \frac{2(n-1)}{n^2}\sigma^4 \\ \mathrm{MSE}(S_n^2) &= \frac{2}{n-1}\sigma^4 \qquad \text{and} \qquad \mathrm{MSE}(\hat{\sigma}_n^2) = \frac{2n-1}{n^2}\sigma^4 \end{aligned}$$

(a) Find the expression for the ratio $\dfrac{Var(S_n^2)}{Var(\hat{\sigma}_n^2)}$.

$$\frac{Var(S_n^2)}{Var(\hat{\sigma}_n^2)} = \frac{\frac{2}{n-1}\sigma^4}{\frac{2(n-1)}{n^2}\sigma^4} = \frac{n^2}{(n-1)^2}$$

(b) Find the expression for the ratio $\dfrac{MSE(S_n)}{MSE(\hat{\sigma}_n^2)}$.

$$\frac{MSE(S_n^2)}{MSE(\hat{\sigma}_n^2)} = \frac{\frac{2}{n-1}\sigma^4}{\frac{2n-1}{n^2}\sigma^4} = \frac{2n^2}{(n-1)(2n-1)}$$

(c) Do a ggplot graph in R to visualize the two ratios above as a function of n (start your graphs with $n > 30$). What do you observe? (Report the graphs and code)

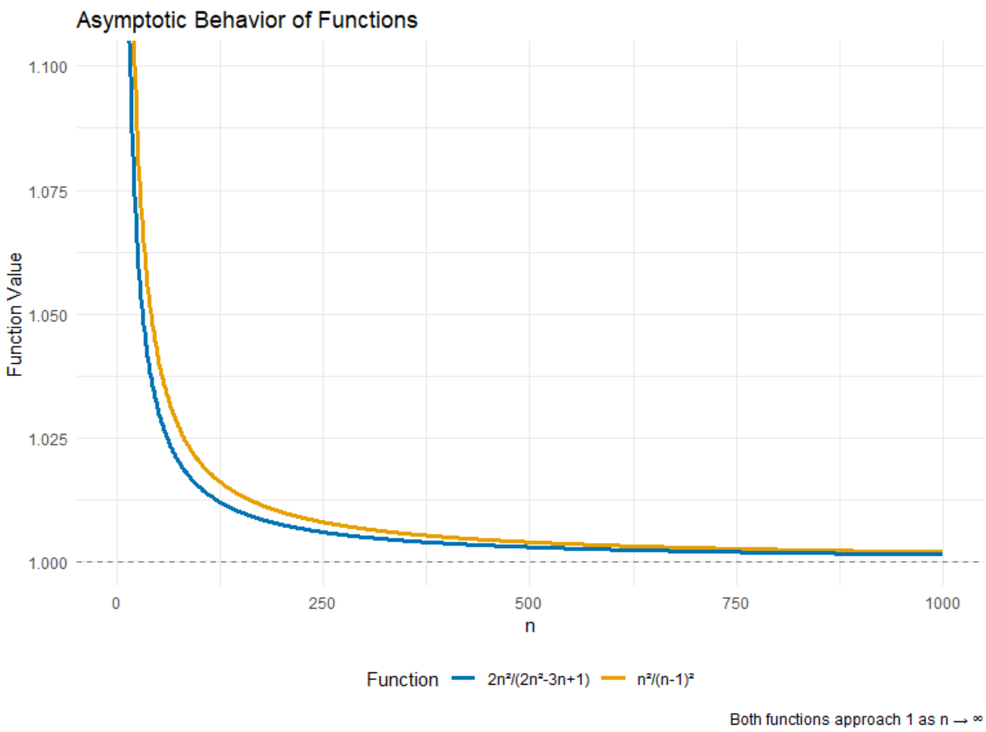

Figure 1: Ratios of Variances and MSEs

3. **Question 3. The Chi-Squared Distribution**

   The chi-squared distribution with k degrees of freedom is defined as the distribution of the sum of squares of k independent standard normal random variables:

$$Q = \sum_{i=1}^{k} Z_i^2 \sim \chi^2(k)$$

   where $Z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ for $i = 1, 2, \ldots, k$.

   In this exercise, you will visualize the shape of the $\chi^2(k)$ distribution for different ks using simulations.

   (a) Generate $R = 1,000$ random samples from the $\chi^2(5)$ distribution as follows:

      i. For each replication $r = 1, 2, \ldots, R$:

         - Generate a random vector $\mathbf{z}^{(r)} = (z_1^{(r)}, z_2^{(r)}, \ldots, z_5^{(r)})$ where each $z_i^{(r)} \sim N(0,1)$
         - Compute $Q^{(r)} = \sum_{i=1}^{5} (z_i^{(r)})^2$

      ii. Store all $Q^{(r)}$ values in a vector $\mathbf{Q}$

   (b) Create a histogram of the simulated values $\mathbf{Q}$ with 60 bins, scaled to have unit area (i.e., a density histogram).

   (c) On the same graph, plot the theoretical probability density function of the $\chi^2(5)$ distribution.

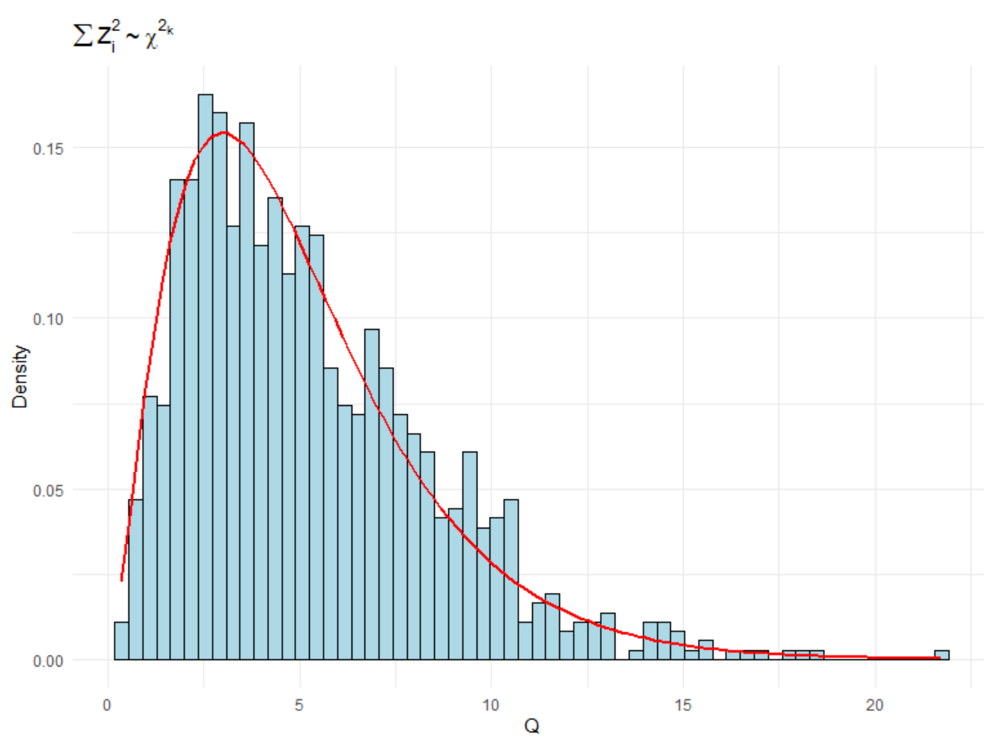

Figure 2: Histogram of Simulated Chi-Squared Values with Theoretical PDF Overlay

4. **Question 4. Randomization Inference**

Use R to solve the following problem and report your code.

Consider the following extremely small sample from the GAIN Experiment.

Table 1: Six Observations from the GAIN Experiment

Individual	Potential Outcome $\mathbf{Y_i(0)}$	Potential Outcome $\mathbf{Y_i(1)}$	Treatment $\mathbf{D_i}$	Observed Outcome $\mathbf{Y_i}^{Obs}$
1	66	?	0	66
2	0	?	0	0
3	0	?	0	0
4	?	0	1	0
5	?	607	1	607
6	?	436	1	436

The fundamental problem of causal inference is that we can never observe both potential outcomes for any single individual. We only observe:

$$Y_i^{Obs} = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0)$$

A common estimator for the Average Treatment Effect (ATE) is the simple difference in means from the sample:

$$\hat{\tau} = \frac{1}{N_1} \sum_{i:D_i=1} Y_i - \frac{1}{N_0} \sum_{i:D_i=0} Y_i$$

where N_1 and N_0 are the number of treated and control units, respectively.

Calculating $\hat{\tau}$

(a) Calculate the mean outcome for the treatment group ($D_i = 1$).

$$\text{Mean}_{D=1} = \frac{607 + 436 + 0}{3} = 347.67$$

(b) Calculate the mean outcome for the control group ($D_i = 0$).

$$\text{Mean}_{D=0} = \frac{66 + 0 + 0}{3} = 22$$

(c) Calculate the simple difference in means ($\hat{\tau}$). Based **only** on this result, what would you conclude about the effect of the program?

$$\hat{\tau} = 347.67 - 22 = 325.67$$

One would naïvely conclude the program has a large, positive effect on earnings.

Fisher's Exact Test

We will now use Randomization Inference to test the sharp null hypothesis of no treatment effect for any individual, formalized as:

$$H_0 : Y_i(1) = Y_i(0) \quad \forall i$$

Under this null hypothesis, all potential outcomes are known and fixed. The observed differences are solely due to the random assignment of treatment D_i. Notice that if we assume the null is true, table 1 becomes:

Table 2: Six Observations from the GAIN Experiment (Under Null Hypothesis)

Individual	**Potential Outcome $\mathbf{Y_i(0)}$**	**Potential Outcome $\mathbf{Y_i(1)}$**	**Treatment $\mathbf{D_i}$**	**Observed Outcome $\mathbf{Y_i}^{Obs}$**
1	66	66	0	66
2	0	0	0	0
3	0	0	0	0
4	0	0	1	0
5	607	607	1	607
6	436	436	1	436

And we can derive the **randomization distribution** of any test statistic: a function of the assignment vector $\mathbf{D}$ and the observed outcomes $\mathbf{Y}^{\text{obs}}$. This statistic varies only because of the random assignment of treatment ($\mathbf{D}$), not because of any uncertainty about the potential outcomes.[1] Consider the test statistic $T(\mathbf{D}, \mathbf{Y}^{obs})$, basically $\hat{\tau}$:

$$T = T(\mathbf{D}, \mathbf{Y}^{\text{obs}}) = \frac{1}{3}\sum_{i=1}^{6} D_i \cdot Y_i^{\text{obs}} - \frac{1}{3}\sum_{i=1}^{6}(1 - D_i) \cdot Y_i^{\text{obs}}.$$

(a) For the observed assignment vector $\mathbf{D}^{obs} = (0,0,0,1,1,1)$, calculate the value of the test statistic, T. You can call this T^{obs} (Repeating (c) from last question.)

$$T^{obs} = \frac{1}{3}(0 + 607 + 436) - \frac{1}{3}(66 + 0 + 0) = 325.67$$

(b) How many possible $\mathbf{D}$ vectors are there for assigning the three treatment assignments ($D_i = 1$) to these six individuals? This is the number of possible permutations of the assignment vector. There are $\binom{6}{3} = 20$ possible ways to assign three treatment statuses.

(c) The **randomization distribution** of the test statistic is generated by calculating $T(\mathbf{D}, \mathbf{Y}^{obs})$ for **every possible assignment vector D**, holding the vector of outcomes fixed. For each possible assignment, calculate the T.

[1] There is no missing data issue.

(d) Plot a histogram of T calculated under all possible random assignments. This histogram represents the exact distribution of the test statistic under the sharp null hypothesis.

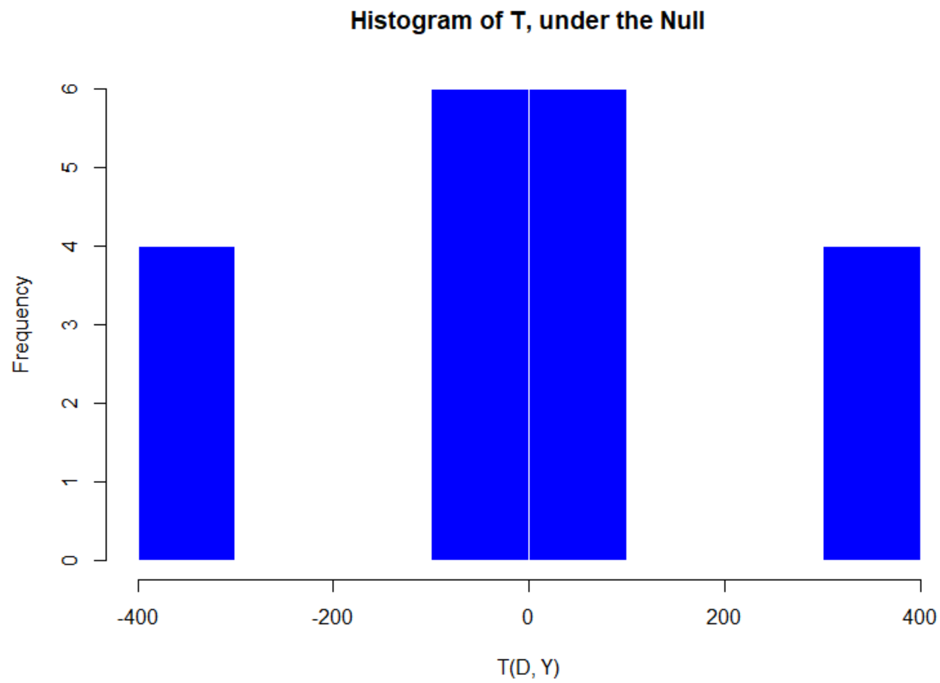

Figure 3: Randomization Distribution of Test Statistic T

(e) On the histogram, mark the value of T^{Obs} calculated in (a).

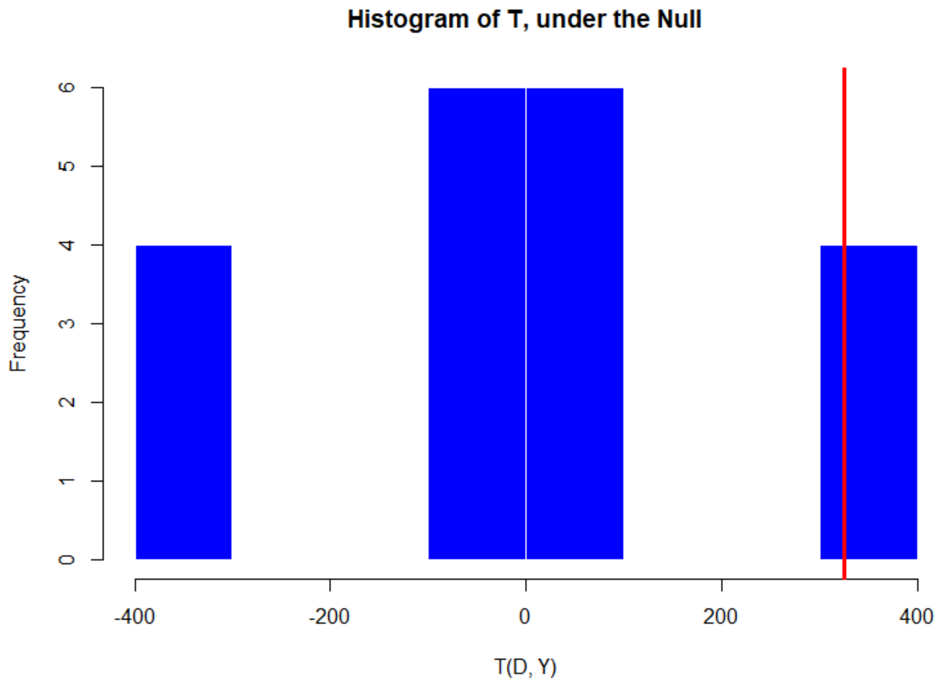

Figure 4: Randomization Distribution of Test Statistic T

(f) Find the 2.5th percentile and the 97.5th percentile of this randomization distribution. These form a 95% reference interval for the test statistic under the null hypothesis. Assuming a continuous distribution and using linear interpolation (type = 7 for the quantile() command in R) for the quantiles, the 2.5th percentile is -348.7667 and the 97.5th percentile is 348.7667.

(g) Is the observed test statistic T^{obs} from part (a) inside or outside this 95% interval? Based on this, would you reject or fail to reject the null hypothesis H_0 at the 5% significance level? The observed test statistic $T^{obs} = 325.67$ is inside the 95% interval (-348.7667, 348.7667). Therefore, we fail to reject the sharp null hypothesis of no effect for any individual at the 5% significance level.

(h) **(Optional)** Calculate the two-sided **p-value**. This is the probability, under the null hypothesis, of observing a value of the test statistic that is as extreme in absolute value than the one actually observed. Formally, it is the proportion of assignment vectors for which $|T(\mathbf{D}, \mathbf{Y}^{obs})| \geq |T^{obs}|$.

(i) **(Optional)** Based on this p-value, what do you conclude about the null hypothesis?

Naturally, for a larger dataset, performing the Fisher exact test gets computationally intensive, so it's rarely seen in practice. However, this is a pretty cool paper that uses the randomization inference framework to re-evaluate the results of several high-profile RCTs.