Review for Quiz II

Wenhao Jiang

Department of Sociology
New York University

October 13, 2022

Point Estimate

Parameters and Estimates

- A **population parameter** describes characteristics of the population (p or μ), which we never observe.
- We use a point **estimate** ($\hat{p}$ or $\hat{\mu}$). calculate using the sample, to estimate the population parameter.

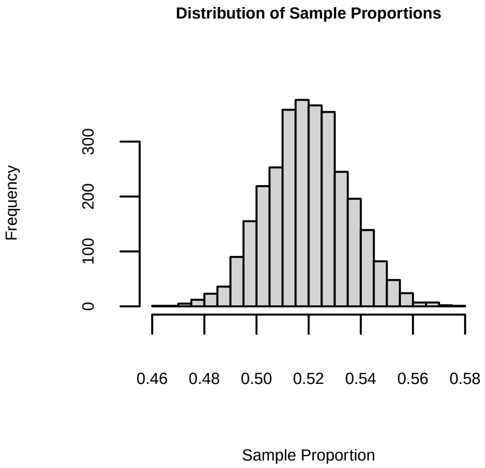
Estimates Uncertainty

- The point **estimate** always includes uncertainty.
- When we draw random samples from the population for a large number of times, the histogram of the sample estimates (e.g., $\hat{p}$s) will be normally distributed. The histogram is the **sampling distribution**.

Estimates Uncertainty

- The point **estimate** always includes uncertainty.
- When we draw random samples from the population for a large number of times, the histogram of the sample estimates (e.g., $\hat{p}$s) will be normally distributed. The histogram is the **sampling distribution**.

Estimates Uncertainty

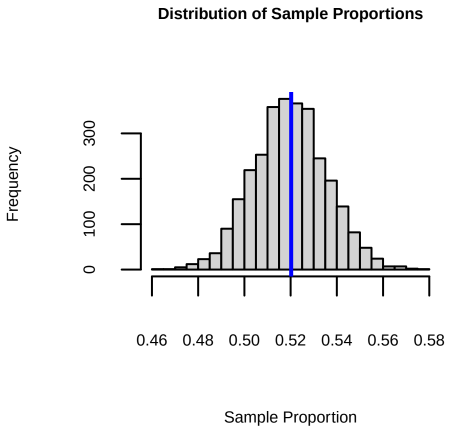


Exercise

- The statement is True or False?
  - In real world, we can normally observe the sampling distribution and see if it's a normal distribution.

Estimates Uncertainty

- The normal distribution of the sample proportions (or other sample statistics) are characterized by two important features
  - 1. The sampling distribution is centered around the **population parameter**



Estimates Uncertainty

- The normal distribution of the sample proportions (or other sample statistics) are characterized by two important features
  - 2. The standard deviation of the sample proportions depends on the population parameter p and sample size n

Estimates Uncertainty

- The normal distribution of the sample proportions (or other sample statistics) are characterized by two important features
  - 2. The standard deviation of the sample proportions depends on the population parameter p and sample size n
  - $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Estimates Uncertainty

- The normal distribution of the sample proportions (or other sample statistics) are characterized by two important features
  - 2. The standard deviation of the sample proportions depends on the population parameter p and sample size n
  - $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
  - We also call it **standard error** of $\hat{p}$

Estimates Uncertainty

- The normal distribution of the sample proportions (or other sample statistics) are characterized by two important features
  - 2. The standard deviation of the sample proportions depends on the population parameter p and sample size n
  - $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
  - We also call it **standard error** of $\hat{p}$
  - We do not observe p, so we use $\hat{p}$ to estimate $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Exercise

- The statement is True or False?
  - As the number of observations in a sample increases, the distribution of its values approaches normality.

Exercise

- The statement is True or False?
  - As the number of observations in a sample increases, the distribution of its values approaches normality.
  - As the number of observations in a sample increases, the (sampling) distribution of its sample means approaches normality.

Exercise

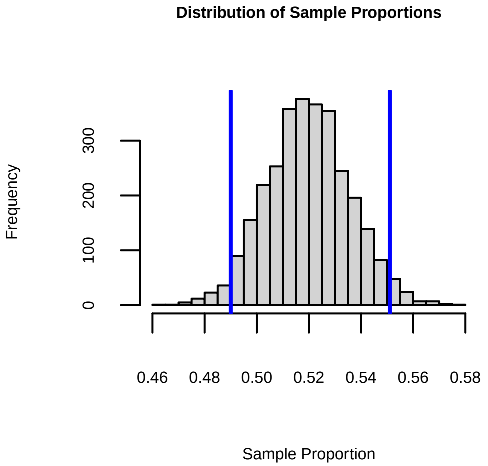
- In the US, a non-trivial proportion of individuals believe in god. Suppose there are only two types of people in the US population—theists and atheists. A social scientist wants to estimate the proportion of theists in the country. To do so, she samples 1225 individuals randomly from the population, and find that 639 believe in god.
- What is the point estimate?
  - $639/1225 = 0.52$
- What is the SE of the point estimate?

Exercise

- In the US, a non-trivial proportion of individuals believe in god. Suppose there are only two types of people in the US population—theists and atheists. A social scientist wants to estimate the proportion of theists in the country. To do so, she samples 1225 individuals randomly from the population, and find that 639 believe in god.
- What is the point estimate?
  - $639/1225 = 0.52$
- What is the SE of the point estimate?
  - $\sqrt{\frac{0.52*(1-0.52)}{1225}} = 0.014$

Estimates Uncertainty

- In the normal distribution of sample means, 95% of the $\hat{p}$s will fall into



Estimates Uncertainty

- In the normal distribution of sample proportions, 95% of the $\hat{p}$s will fall into
- $p - 1.96 \times SE_{\hat{p}} \leq \hat{p} \leq p + 1.96 \times SE_{\hat{p}}$

Estimates Uncertainty

- In the normal distribution of sample proportions, 95% of the $\hat{p}$s will fall into
- $p - 1.96 \times SE_{\hat{p}} \leq \hat{p} \leq p + 1.96 \times SE_{\hat{p}}$
- Equivalently, $\hat{p} - 1.96 \times SE_{\hat{p}} \leq p \leq \hat{p} + 1.96 \times SE_{\hat{p}}$

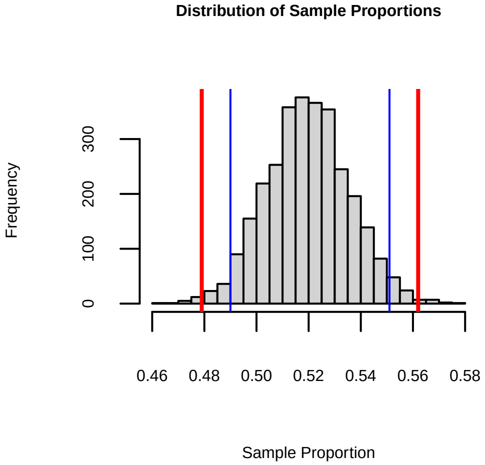
Estimates Uncertainty

- In the normal distribution of sample proportions, 95% of the $\hat{p}$s will fall into
- $p - 1.96 \times SE_{\hat{p}} \leq \hat{p} \leq p + 1.96 \times SE_{\hat{p}}$
- Equivalently, $\hat{p} - 1.96 \times SE_{\hat{p}} \leq p \leq \hat{p} + 1.96 \times SE_{\hat{p}}$
- This is the 95% **Confidence Interval** for the population p

(standard deviation)^2 = variance

Estimates Uncertainty

- We can adjust the 95% confidence interval to e.g., 99% confidence interval. Now, we want an interval where 99% of the $\hat{p}$s fall into



Estimates Uncertainty

- Looking at the z-score table
- $p - 2.575 \times SE_{\hat{p}} \leq \hat{p} \leq p + 2.575 \times SE_{\hat{p}}$

Estimates Uncertainty

- Looking at the z-score table
- $p - 2.575 \times SE_{\hat{p}} \leq \hat{p} \leq p + 2.575 \times SE_{\hat{p}}$
- Equivalently, $\hat{p} - 2.575 \times SE_{\hat{p}} \leq p \leq \hat{p} + 2.575 \times SE_{\hat{p}}$

Estimates Uncertainty

- Looking at the z-score table
- $p - 2.575 \times SE_{\hat{p}} \leq \hat{p} \leq p + 2.575 \times SE_{\hat{p}}$
- Equivalently, $\hat{p} - 2.575 \times SE_{\hat{p}} \leq p \leq \hat{p} + 2.575 \times SE_{\hat{p}}$
- This is the 99% **Confidence Interval** for the population p

Estimates Uncertainty

- We call $1.96 \times SE_{\hat{p}}$ or $2.575 \times SE_{\hat{p}}$ the **Margin of Error**

Estimates Uncertainty

- We call $1.96 \times SE_{\hat{p}}$ or $2.575 \times SE_{\hat{p}}$ the **Margin of Error**
- The statement is True or False?
  - At the same level of confidence (e.g., 95%), the width of the Confidence Interval is equal to twice of the Margin of Error

Exercise

- A researcher estimated a population parameter p by two samples, with sample size n_1 and n_2. She constructed 95% Confidence Interval for sample 1, and 99% CI for sample 2. She found that the width of the two CIs are the same.
- What is the relation between n_1 and n_2?
  - A. $n_1 > n_2$
  - B. $n_1 = n_2$
  - C. $n_1 < n_2$
  - D. Insufficient information

Estimates Uncertainty

- Note that the special formula $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ only applies to the point estimate of **proportion**. When the point estimate is about mean (e.g., mean income of a sample),
  - $SE_{\hat{\mu}} = \sqrt{\frac{var(Y)}{n}}$, where $var(Y)$ represent the variance of the population

Estimates Uncertainty

- Note that the special formula $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ only applies to the point estimate of **proportion**. When the point estimate is about mean (e.g., mean income of a sample),
  - $SE_{\hat{\mu}} = \sqrt{\frac{var(Y)}{n}}$, where $var(Y)$ represent the variance of the population
  - We do not observe $var(Y)$, so we use sample variance $\hat{var}(Y)$ to estimate $SE_{\hat{\mu}} = \sqrt{\frac{\hat{var}(Y)}{n}}$ sqrt(s^2/n)

Exercise

- A fisherman wants to know the mean weight of fish in his fish pond. Knowing that he cannot drain the pond and weigh all fish in a single time, he randomly catches 63 fish, finding that the mean weight of the fish he caught is 21 lbs, and the standard deviation is 2.6 lbs.
- What is the point estimate?

Exercise

- A fisherman wants to know the mean weight of fish in his fish pond. Knowing that he cannot drain the pond and weigh all fish in a single time, he randomly catches 63 fish, finding that the mean weight of the fish he caught is 21 lbs, and the standard deviation is 2.6 lbs.
- What is the point estimate?
- What is the SE of the point estimate?
  - $SE_{\hat{\mu}} = \sqrt{\frac{2.6^2}{63}} = 0.33$

Difference b/w two Samples

Difference between two proportions

- Suppose the proportion of men in the population supporting abortion is p_m, the proportion of women in the population supporting abortion is p_w, and the difference is $p_m - p_w$

Difference between two proportions

- Suppose the proportion of men in the population supporting abortion is p_m, the proportion of women in the population supporting abortion is p_w, and the difference is $p_m - p_w$
- We draw a random sample from the population with n_m men and n_w women ($n_m + n_w$ in total).
- We get point estimates $\hat{p}_m$, $\hat{p}_w$, and $\hat{p}_m - \hat{p}_w$

Difference between two proportions

- Suppose the proportion of men in the population supporting abortion is p_m, the proportion of women in the population supporting abortion is p_w, and the difference is $p_m - p_w$
- We draw a random sample from the population with n_m men and n_w women ($n_m + n_w$ in total).
- We get point estimates $\hat{p}_m$, $\hat{p}_w$, and $\hat{p}_m - \hat{p}_w$
- What is the standard error of $\hat{p}_m - \hat{p}_w$?
  - $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{p_m(1-p_m)}{n_m} + \frac{p_w(1-p_w)}{n_w}}$, and we estimate it using $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{\hat{p}_m(1-\hat{p}_m)}{n_m} + \frac{\hat{p}_w(1-\hat{p}_w)}{n_w}}$

Exercise

- The statement is True or False?
- When we sample from the population for many times and plot the histograms of $\hat{p}_m$, $\hat{p}_w$, and $\hat{p}_m - \hat{p}_w$ calculated from each sample, we will find that $\hat{p}_m$ and $\hat{p}_w$ form bell curves (normal distribution), but $\hat{p}_m - \hat{p}_w$ does not.

Exercise

- To study the gender difference in parental leave, a sociologist surveys 128 men and 254 women with children, finding that 11% men have taken paternity leave, while 65% women have.
- What is the point estimate of the gender difference in parental leave? (i.e., what is the proportion difference?)

Exercise

- To study the gender difference in parental leave, a sociologist surveys 128 men and 254 women with children, finding that 11% men have taken paternity leave, while 65% women have.
- What is the point estimate of the gender difference in parental leave? (i.e., what is the proportion difference?)
  - $0.65 - 0.11 = 0.54$

Exercise

- To study the gender difference in parental leave, a sociologist surveys 128 men and 254 women with children, finding that 11% men have taken paternity leave, while 65% women have.
- What is the point estimate of the gender difference in parental leave? (i.e., what is the proportion difference?)
  - $0.65 - 0.11 = 0.54$
- What is the SE of the point estimate?

Exercise

- To study the gender difference in parental leave, a sociologist surveys 128 men and 254 women with children, finding that 11% men have taken paternity leave, while 65% women have.
- What is the point estimate of the gender difference in parental leave? (i.e., what is the proportion difference?)
  - $0.65 - 0.11 = 0.54$
- What is the SE of the point estimate?
  - $\sqrt{\frac{0.11*(1-0.11)}{128} + \frac{0.65*(1-0.65)}{254}} = 0.04$

Difference between two Sample Proportions - A Special Case

- What is the standard error of $\hat{p}_m - \hat{p}_w$ **under the null hypothesis** $H_0 : p_m = p_w = p$?
- In this **special case** of null hypothesis $H_0 : p_m = p_w = p$, we will replace p_m and p_w by a uniform p.
- $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{p(1-p)}{n_m} + \frac{p(1-p)}{n_w}}$

Difference between two Sample Proportions - A Special Case

- What is the standard error of $\hat{p}_m - \hat{p}_w$ **under the null hypothesis** $H_0 : p_m = p_w = p$?
- In this **special case** of null hypothesis $H_0 : p_m = p_w = p$, we will replace p_m and p_w by a uniform p.
- $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{p(1-p)}{n_m} + \frac{p(1-p)}{n_w}}$
- We do not observe p. We estimate p using the sample we drew.
- We denote this $\hat{p}$ in the case as $\hat{p}_{pooled}$, and $\hat{p}_{pooled} = \frac{\hat{p}_m \times n_m + \hat{p}_w \times n_w}{n_m + n_w}$

Difference between two Sample Proportions - A Special Case

z = (point estimate - null hypothesis value)/SE = (0.04 - 0)/0.068 = 0.588

p-value = 0.2776*2=0.5552 > 0.05

we conclude that we do not reject the null hypothesis

- What is the standard error of $\hat{p}_m - \hat{p}_w$ **under the null hypothesis** $H_0 : p_m = p_w = p$?
- In this **special case** of null hypothesis $H_0 : p_m = p_w = p$, we will replace p_m and p_w by a uniform p.
- $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{p(1-p)}{n_m} + \frac{p(1-p)}{n_w}}$
- We do not observe p. We estimate p using the sample we drew.
- We denote this $\hat{p}$ in the case as $\hat{p}_{pooled}$, and $\hat{p}_{pooled} = \frac{\hat{p}_m \times n_m + \hat{p}_w \times n_w}{n_m + n_w}$
- $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_m} + \frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_w}}$

(0.52*120 + 0.56*100)/(100+120) = p_pooled <- the proportion of people who support abortion in the pooled sample (i.e., men and women together)

= 0.538

SE = 0.068 of the point estimate

Difference between two Sample Means

- In many cases, instead of proportion, we are interested in the mean values of two populations and their difference.

Difference between two Sample Means

- In many cases, instead of proportion, we are interested in the mean values of two populations and their difference.
- For example, μ_1 could be mean hours worked among men in household production and μ_2 could be mean hours worked among women. We want to know $\mu_1 - \mu_2$

Difference between two Sample Means

- In many cases, instead of proportion, we are interested in the mean values of two populations and their difference.
- For example, μ_1 could be mean hours worked among men in household production and μ_2 could be mean hours worked among women. We want to know $\mu_1 - \mu_2$
- We never observe it at the population level. Instead, we gauge it using a male sample and a female sample. The point estimate for the difference of the two sample means is $\bar{Y}_1 - \bar{Y}_2$

Difference between two Sample Means

- The standard error of the point estimate $\bar{Y}_1 - \bar{Y}_2$, $SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}$
- σ_1 is the standard deviation of population 1 (e.g., standard deviation of hours worked of the male population), and σ_2 is the standard deviation of population 2 (e.g., standard deviation of hours worked of the female population)

Difference between two Sample Means

- The standard error of the point estimate $\bar{Y}_1 - \bar{Y}_2$, $SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}$
- σ_1 is the standard deviation of population 1 (e.g., standard deviation of hours worked of the male population), and σ_2 is the standard deviation of population 2 (e.g., standard deviation of hours worked of the female population)
- We never observe σ_1 and σ_2

Difference between two Sample Means

- The standard error of the point estimate $\bar{Y}_1 - \bar{Y}_2$, $SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}$
- σ_1 is the standard deviation of population 1 (e.g., standard deviation of hours worked of the male population), and σ_2 is the standard deviation of population 2 (e.g., standard deviation of hours worked of the female population)
- We never observe σ_1 and σ_2
- We estimate it by $SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$
- s_1 is the standard deviation of sample 1 (e.g., standard deviation of hours worked of the male sample), and s_2 is the standard deviation of sample 2 (e.g., standard deviation of hours worked of the female sample)

Exercise

Gender	Sample size	Mean of hours	Std Dev
Men	250	1.8	0.2
Women	300	4.5	0.3

- What is the point estimate of the gender difference in mean of hours?

Exercise

Gender	Sample size	Mean of hours	Std Dev
Men	250	1.8	0.2
Women	300	4.5	0.3

- What is the point estimate of the gender difference in mean of hours?
  - $\bar{Y}_1 - \bar{Y}_2 = 4.5 - 1.8 = 2.7$

Exercise

Gender	Sample size	Mean of hours	Std Dev
Men	250	1.8	0.2
Women	300	4.5	0.3

- What is the point estimate of the gender difference in mean of hours?
  - $\bar{Y}_1 - \bar{Y}_2 = 4.5 - 1.8 = 2.7$
- What is $SE_{\bar{Y}_1 - \bar{Y}_2}$?

Exercise

Gender	Sample size	Mean of hours	Std Dev
Men	250	1.8	0.2
Women	300	4.5	0.3

- What is the point estimate of the gender difference in mean of hours?
  - $\bar{Y}_1 - \bar{Y}_2 = 4.5 - 1.8 = 2.7$
- What is $SE_{\bar{Y}_1 - \bar{Y}_2}$?
  - $SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} = \sqrt{\frac{0.2^2}{250} + \frac{0.3^2}{300}} = 0.02$

Exercise

Gender	Sample size	Mean of hours	Std Dev
Men	250	1.8	0.2
Women	300	4.5	0.3

- What is the point estimate of the gender difference in mean of hours?
  - $\bar{Y}_1 - \bar{Y}_2 = 4.5 - 1.8 = 2.7$
- What is $SE_{\bar{Y}_1 - \bar{Y}_2}$?
  - $SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} = \sqrt{\frac{0.2^2}{250} + \frac{0.3^2}{300}} = 0.02$
- What is the 95% Confidence Interval of $Y_1 - Y_2$?

Exercise

standard error?

Gender	Sample size	Mean of hours	Std Dev
Men	250	1.8	0.2
Women	300	4.5	0.3

- What is the point estimate of the gender difference in mean of hours?
  - $\bar{Y}_1 - \bar{Y}_2 = 4.5 - 1.8 = 2.7$
- What is $SE_{\bar{Y}_1 - \bar{Y}_2}$?
  - $SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} = \sqrt{\frac{0.2^2}{250} + \frac{0.3^2}{300}} = 0.02$
- What is the 95% Confidence Interval of $Y_1 - Y_2$?
  - $\bar{Y}_1 - \bar{Y}_2 - 1.96 \times SE_{\bar{Y}_1 - \bar{Y}_2} \leq Y_1 - Y_2 \leq \bar{Y}_1 - \bar{Y}_2 + 1.96 \times SE_{\bar{Y}_1 - \bar{Y}_2}$, i.e., $2.66 \leq Y_1 - Y_2 \leq 2.74$

Hypothesis Testing

Terminology

She wants to know whether there is gender difference in the income trajectory after childbirth. What will be the null hypothesis? No gender difference

- The **null hypothesis** H_0 often states “nothing is going on here”
- The **alternative hypothesis** H_A represents what we really think is going on—a substantive alternative to the null

Terminology

- The **null hypothesis** H_0 often states "nothing is going on here"
- The **alternative hypothesis** H_A represents what we really think is going on—a substantive alternative to the null
- For example, we want to know the gender difference in the hours devoted to housework
  - H_0: there is no gender difference **in the population**
  - H_A: there is gender difference **in the population**

Terminology

- We use a sample to estimate a population property, and decide whether to reject H_0 or not.
- Point estimate always involves uncertainties, so one may falsely reject the null hypothesis or fail to reject the null hypothesis

Truth	Do not reject H_0	Reject H_0 in favor of H_A
H_0 is true	Right decision	Type 1 error
H_A is true	Type 2 error	Right decision

H0: There is no gender difference in income trajectory after childbirth
You conclude from your sample that there is gender difference
However, in the population, there is actually no gender difference
Type 1 error

Hypothesis Testing

- Hypothesis testing focuses on H_0 and the probability of **Type 1 error**

Hypothesis Testing

- Hypothesis testing focuses on H_0 and the probability of **Type 1 error**
- In a more plain language, it means that, **given that the null hypothesis is true**, what is the probability of getting the test statistics. (i.e., sample point estimate)
  - E.g., if H_0: there is no gender differences in yearly income, yet we observe from our sample that $\bar{Y}_1 - \bar{Y}_2 = 10,000$, it is reasonable to speculate that H_0 is incorrect (i.e., we reject H_0 and accept H_A: there are gender differences in yearly income at the level of population).

Hypothesis Testing

- Hypothesis testing focuses on H_0 and the probability of **Type 1 error**
- In a more plain language, it means that, **given that the null hypothesis is true**, what is the probability of getting the test statistics.
  - E.g., if H_0: there is no gender differences in yearly income, yet we observe from our sample that $\bar{Y}_1 - \bar{Y}_2 = 10,000$, it is reasonable to speculate that H_0 is incorrect (i.e., we reject H_0 and accept H_A: there are gender differences in yearly income at the level of population).
  - The lower the probability, the less likely the null hypothesis is true.
  - The higher the probability, the more likely the null hypothesis is true.

Hypothesis Testing

- Hypothesis testing focuses on H_0 and the probability of **Type 1 error**
- In a more plain language, it means that, **given that the null hypothesis is true**, what is the probability of getting the test statistics.
  - E.g., if H_0: there is no gender differences in yearly income, yet we observe from our sample that $\bar{Y}_1 - \bar{Y}_2 = 10,000$, it is reasonable to speculate that H_0 is incorrect (i.e., we reject H_0 and accept H_A: there are gender differences in yearly income at the level of population).
  - The lower the probability, the less likely the null hypothesis is true.
  - The higher the probability, the more likely the null hypothesis is true.
- How low should this probability to be, such that we are "confident" enough to reject the null hypothesis?

Hypothesis Testing

- Hypothesis testing focuses on H_0 and the probability of **Type 1 error**
- In a more plain language, it means that, **given that the null hypothesis is true**, what is the probability of getting the test statistics.
  - E.g., if H_0: there is no gender differences in yearly income, yet we observe from our sample that $\bar{Y}_1 - \bar{Y}_2 = 10,000$, it is reasonable to speculate that H_0 is incorrect (i.e., we reject H_0 and accept H_A: there are gender differences in yearly income at the level of population).
  - The lower the probability, the less likely the null hypothesis is true.
  - The higher the probability, the more likely the null hypothesis is true.
- How low should this probability to be, such that we are "confident" enough to reject the null hypothesis?
  - This is called **Significance level** (α), usually pre-set at 0.05 (5%). When the probability is lower than 0.05, we are confident to reject the null hypothesis.

Hypothesis Testing

- What does it mean to say, when the probability is lower than 0.05, we are confident to reject the null hypothesis?

Hypothesis Testing

- What does it mean to say, when the probability is lower than 0.05, we are confident to reject the null hypothesis?
- We sample many times (K times) from the population, assuming that there is no gender difference in yearly income. In 95% of the K point estimates, $\bar{Y}_1 - \bar{Y}_2$ falls within [*lower bar*, *higher bar*], e.g., $[-1000, 1000]$
- The point estimate of the single sample that we draw, however, reports that $\bar{Y}_1 - \bar{Y}_2 = 10,000$. This is very different from 0 and the times of getting $\bar{Y}_1 - \bar{Y}_2 = 10,000$ are much smaller than $0.05K$. We therefore reject H_0 at the Significance level $\alpha = 0.05$.

Hypothesis Testing

- How do we calculate the probability of getting the sample statistics, given the null hypothesis?
- This is where z-score, the z-score table, and p-value step in

Hypothesis Testing

- How do we calculate the probability of getting the sample statistics, given the null hypothesis?
- This is where z-score, the z-score table, and p-value step in
- Suppose we are interested in the gender difference in yearly income. We sample 1200 men and 1000 women, finding that $\bar{Y}_{men} = 50000$ and $\bar{Y}_{women} = 49000$. We also find that $s_{men} = 8000$ and $s_{women} = 7000$

Hypothesis Testing

- How do we calculate the probability of getting the sample statistics, given the null hypothesis?
- This is where z-score, the z-score table, and p-value step in
- Suppose we are interested in the gender difference in yearly income. We sample 1200 men and 1000 women, finding that $\bar{Y}_{men} = 50000$ and $\bar{Y}_{women} = 49000$. We also find that $s_{men} = 8000$ and $s_{women} = 7000$
    - 0. Calculate the point estimate $\bar{Y}_{men} - \bar{Y}_{women} = 50000 - 49000 = 1000$, and $SE_{\bar{Y}_{men}-\bar{Y}_{women}} = \sqrt{\frac{8000^2}{1200} + \frac{7000^2}{1000}} = 319.9$

Hypothesis Testing

- How do we calculate the probability of getting the sample statistics, given the null hypothesis?
- This is where z-score, the z-score table, and p-value step in
- Suppose we are interested in the gender difference in yearly income. We sample 1200 men and 1000 women, finding that $\bar{Y}_{men} = 50000$ and $\bar{Y}_{women} = 49000$. We also find that $s_{men} = 8000$ and $s_{women} = 7000$
  - 0. Calculate the point estimate $\bar{Y}_{men} - \bar{Y}_{women} = 50000 - 49000 = 1000$, and $SE_{\bar{Y}_{men} - \bar{Y}_{women}} = \sqrt{\frac{8000^2}{1200} + \frac{7000^2}{1000}} = 319.9$
  - 1. Look at the null hypothesis. In this case, H_0: $\bar{Y}_{men} - \bar{Y}_{women} = 0$.

Hypothesis Testing

- How do we calculate the probability of getting the sample statistics, given the null hypothesis?
- This is where z-score, the z-score table, and p-value step in
- Suppose we are interested in the gender difference in yearly income. We sample 1200 men and 1000 women, finding that $\bar{Y}_{men} = 50000$ and $\bar{Y}_{women} = 49000$. We also find that $s_{men} = 8000$ and $s_{women} = 7000$
  - 0. Calculate the point estimate $\bar{Y}_{men} - \bar{Y}_{women} = 50000 - 49000 = 1000$, and $SE_{\bar{Y}_{men} - \bar{Y}_{women}} = \sqrt{\frac{8000^2}{1200} + \frac{7000^2}{1000}} = 319.9$
  - 1. Look at the null hypothesis. In this case, H_0: $\bar{Y}_{men} - \bar{Y}_{women} = 0$.
  - 2. Calculate the z-score: $z = \frac{point\ estimate - null\ value}{SE} = \frac{\bar{Y}_{men} - \bar{Y}_{women} - 0}{SE_{\bar{Y}_{men} - \bar{Y}_{women}}} = \frac{1000}{319.9} = 3.13$

Hypothesis Testing

- How do we calculate the probability of getting the sample statistics, given the null hypothesis?
- This is where z-score, the z-score table, and p-value step in
- Suppose we are interested in the gender difference in yearly income. We sample 1200 men and 1000 women, finding that $\bar{Y}_{men} = 50000$ and $\bar{Y}_{women} = 49000$. We also find that $s_{men} = 8000$ and $s_{women} = 7000$
    - 0. Calculate the point estimate $\bar{Y}_{men} - \bar{Y}_{women} = 50000 - 49000 = 1000$, and $SE_{\bar{Y}_{men} - \bar{Y}_{women}} = \sqrt{\frac{8000^2}{1200} + \frac{7000^2}{1000}} = 319.9$
    - 1. Look at the null hypothesis. In this case, H_0: $\bar{Y}_{men} - \bar{Y}_{women} = 0$.
    - 2. Calculate the z-score: $z = \frac{point\ estimate - null\ value}{SE} = \frac{\bar{Y}_{men} - \bar{Y}_{women} - 0}{SE_{\bar{Y}_{men} - \bar{Y}_{women}}} = \frac{1000}{319.9} = 3.13$
    - 3. Check the p-value associated with the z-score. We find 0.0009, meaning that if there is no gender difference in the population, the chance that we get the above sample statistics is 0.0009.

Hypothesis Testing

- How do we calculate the probability of getting the sample statistics, given the null hypothesis?
- This is where z-score, the z-score table, and p-value step in
- Suppose we are interested in the gender difference in yearly income. We sample 1200 men and 1000 women, finding that $\bar{Y}_{men} = 50000$ and $\bar{Y}_{women} = 49000$. We also find that $s_{men} = 8000$ and $s_{women} = 7000$
  - 0. Calculate the point estimate $\bar{Y}_{men} - \bar{Y}_{women} = 50000 - 49000 = 1000$, and $SE_{\bar{Y}_{men} - \bar{Y}_{women}} = \sqrt{\frac{8000^2}{1200} + \frac{7000^2}{1000}} = 319.9$
  - 1. Look at the null hypothesis. In this case, H_0: $\bar{Y}_{men} - \bar{Y}_{women} = 0$.
  - 2. Calculate the z-score: $z = \frac{point\ estimate - null\ value}{SE} = \frac{\bar{Y}_{men} - \bar{Y}_{women} - 0}{SE_{\bar{Y}_{men} - \bar{Y}_{women}}} = \frac{1000}{319.9} = 3.13$
  - 3. Check the p-value associated with the z-score. We find 0.0009, meaning that if there is no gender difference in the population, the chance that we get the above sample statistics is 0.0009.
  - 4. Note that to get a difference of 1000, it is also possible that women's income lead men's. The possibility to get such a deviation (1000) given the null hypothesis $Y_{men} = Y_{women}$ is therefore $0.0009 \times 2 = 0.0018$. This is the p-value $= 0.0018$.

Hypothesis Testing

- How do we calculate the probability of getting the sample statistics, given the null hypothesis?
- This is where z-score, the z-score table, and p-value step in
- Suppose we are interested in the gender difference in yearly income. We sample 1200 men and 1000 women, finding that $\bar{Y}_{men} = 50000$ and $\bar{Y}_{women} = 49000$. We also find that $s_{men} = 8000$ and $s_{women} = 7000$
  - 0. Calculate the point estimate $\bar{Y}_{men} - \bar{Y}_{women} = 50000 - 49000 = 1000$, and $SE_{\bar{Y}_{men} - \bar{Y}_{women}} = \sqrt{\frac{8000^2}{1200} + \frac{7000^2}{1000}} = 319.9$
  - 1. Look at the null hypothesis. In this case, H_0: $\bar{Y}_{men} - \bar{Y}_{women} = 0$.
  - 2. Calculate the z-score: $z = \frac{point\ estimate - null\ value}{SE} = \frac{\bar{Y}_{men} - \bar{Y}_{women} - 0}{SE_{\bar{Y}_{men} - \bar{Y}_{women}}} = \frac{1000}{319.9} = 3.13$
  - 3. Check the p-value associated with the z-score. We find 0.0009, meaning that if there is no gender difference in the population, the chance that we get the above sample statistics is 0.0009.
  - 4. Note that to get a difference of 1000, it is also possible that women's income lead men's. The possibility to get such a deviation (1000) given the null hypothesis $Y_{men} = Y_{women}$ is therefore $0.0009 \times 2 = 0.0018$. This is the p-value $= 0.0018$.

Hypothesis Testing

- p-value $= 0.0018$ is smaller than $\alpha = 0.05$. We reject the null hypothesis at the 0.05 significance level.
- Equivalently, the 95% confidence interval of $Y_{men} - Y_{women}$ does not contain 1000 under the null hypothesis.

Exercise

- Different countries in the world have different sex ratios at birth. A scholar wants to know whether country C has a skewed sex ratio at birth or not. He samples 1200 infants from the country and finds that 53% of the infants in the sample are boys and 47% are girls.
- What is the null hypothesis?

Exercise

- Different countries in the world have different sex ratios at birth. A scholar wants to know whether country C has a skewed sex ratio at birth or not. He samples 1200 infants from the country and finds that 53% of the infants in the sample are boys and 47% are girls.
- What is the null hypothesis?
- What is the point estimate of the boy proportion?

Exercise

- Different countries in the world have different sex ratios at birth. A scholar wants to know whether country C has a skewed sex ratio at birth or not. He samples 1200 infants from the country and finds that 53% of the infants in the sample are boys and 47% are girls.
- What is the null hypothesis? p_boy = 0.5
- What is the point estimate of the boy proportion? 0.53
- What is the SE of the point estimate? sqrt(0.53*(1-0.53)/1200) = 0.0144

Exercise

- Different countries in the world have different sex ratios at birth. A scholar wants to know whether country C has a skewed sex ratio at birth or not. He samples 1200 infants from the country and finds that 53% of the infants in the sample are boys and 47% are girls.
- What is the null hypothesis?
- What is the point estimate of the boy proportion?
- What is the SE of the point estimate?
- What is the z-score of the point estimate under the null hypothesis?

Exercise

- Different countries in the world have different sex ratios at birth. A scholar wants to know whether country C has a skewed sex ratio at birth or not. He samples 1200 infants from the country and finds that 53% of the infants in the sample are boys and 47% are girls.
- What is the null hypothesis?
- What is the point estimate of the boy proportion?
- What is the SE of the point estimate?
- What is the z-score of the point estimate under the null hypothesis?
- What is the p-value associated with the z-score?

Exercise

- Different countries in the world have different sex ratios at birth. A scholar wants to know whether country C has a skewed sex ratio at birth or not. He samples 1200 infants from the country and finds that 53% of the infants in the sample are boys and 47% are girls.
- What is the null hypothesis?
- What is the point estimate of the boy proportion?
- What is the SE of the point estimate?
- What is the z-score of the point estimate under the null hypothesis?
- What is the p-value associated with the z-score?
- Can we reject the null hypothesis at $\alpha = 0.01$?

Exercise

- Different countries in the world have different sex ratios at birth. A scholar wants to know whether country C has a skewed sex ratio at birth or not. He samples 1200 infants from the country and finds that 53% of the infants in the sample are boys and 47% are girls.
- What is the null hypothesis? p_boy = 0.5
- What is the point estimate of the boy proportion? p_boy = 0.53
- What is the SE of the point estimate? SE = 0.0144
- What is the z-score of the point estimate under the null hypothesis? z = (0.53-0.5)/0.0144 = 2.08
- What is the p-value associated with the z-score? p=0.0188*2=0.0376
- Can we reject the null hypothesis at $\alpha = 0.01$? cannot reject at 0.01 significant level
- Does the 95% confidence interval of p under the null hypothesis include $\hat{p}$? What about 99% CI?

95% does not include p hat -> we reject null hypothesis at 0.05 significance level

99% includes p hat -> we do not reject null hypothesis at 0.01 significance level