Class 20

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-10-18

Hypothesis Testing Framework

- $\mathbf{H_0}$: The "Null" Hypothesis
  - Represents a position of skepticism, *nothing* is happening here
  - "There is *not* an association between process A and B"

Hypothesis Testing Framework

- $\mathbf{H_0}$: The "Null" Hypothesis
  - Represents a position of skepticism, *nothing* is happening here
  - "There is *not* an association between process A and B"
- $\mathbf{H_A}$: The "Alternative" Hypothesis
  - The complement of H_0, *something* is happening here
  - "There *is* an association between process A and B"

Conducting a hypothesis test

- Begin by *assuming* H_0 is the "true" state

Conducting a hypothesis test

- Begin by *assuming* H_0 is the “true” state
- Calculate *the probability that you would see results as extreme or more extreme* than what you saw in your study, assuming the distribution under H_0

Conducting a hypothesis test

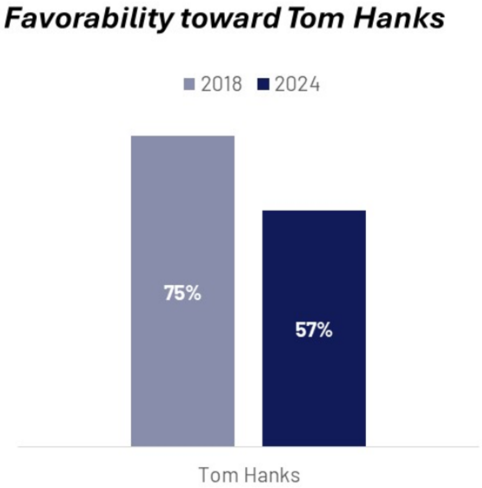
- Begin by *assuming* H_0 is the "true" state
- Calculate *the probability that you would see results as extreme or more extreme* than what you saw in your study, assuming the distribution under H_0
- The lower the probability, the less likely it is that we would see these results if H_0 was the "true" state of our population

Conducting a hypothesis test

- Begin by *assuming* H_0 is the "true" state
- Calculate *the probability that you would see results as extreme or more extreme* than what you saw in your study, assuming the distribution under H_0
- The lower the probability, the less likely it is that we would see these results if H_0 was the "true" state of our population
- If the probability is sufficiently low, we *reject* $\mathbf{H_0}$ and *accept* $\mathbf{H_A}$

Example: One Proportion

In a 2024 Ipsos survey of a representative sample of 2,027 Americans, 1,155 respondents (57.0%) reported that they had a favorable opinion of Tom Hanks, down from 75% in 2018.



Hypotheses

Hypotheses

- H_0: Americans do not have an opinion on Tom Hanks.

Hypotheses

- H_0: Americans do not have an opinion on Tom Hanks.

$$P(\text{Favorable}) = 0.5$$

Hypotheses

- H_0: Americans do not have an opinion on Tom Hanks.

$$P(\text{Favorable}) = 0.5$$

- H_A: Americans do have an opinion on Tom Hanks.

Hypotheses

- H_0: Americans do not have an opinion on Tom Hanks.

$$P(\text{Favorable}) = 0.5$$

- H_A: Americans do have an opinion on Tom Hanks.

$$P(\text{Favorable}) \neq 0.5$$

95% Confidence Interval

95% Confidence Interval

1. Calculate the standard error of the measurement for a proportion.

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

95% Confidence Interval

1. Calculate the standard error of the measurement for a proportion.

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

2. Find Z^* for $\alpha = 0.05$.

$$Z^* = Z_{1-\alpha/2}$$

95% Confidence Interval

1. Calculate the standard error of the measurement for a proportion.

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

2. Find Z^* for $\alpha = 0.05$.

$$Z^* = Z_{1-\alpha/2}$$

3. Construct confidence interval as $\text{point estimate} \pm Z^* \times SE$

Standard Error

$$
\begin{aligned}
SE &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\
&= \sqrt{\frac{0.57(1-0.57)}{2027}} \\
&= 0.011
\end{aligned}
$$

Standard Error

$$
\begin{aligned}
SE &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\
&= \sqrt{\frac{0.57(1-0.57)}{2027}} \\
&= 0.011
\end{aligned}
$$

```r
sqrt((0.57 * (1 - 0.57)) / 2027)
```

```
[1] 0.01099625
```

Finding Z^*

$$\begin{aligned} Z^* &= Z_{1-\alpha/2} \\ &= Z_{1-0.025} \\ &= Z_{0.975} \end{aligned}$$

Finding Z^*

$$\begin{aligned} Z^* &= Z_{1-\alpha/2} \\ &= Z_{1-0.025} \\ &= Z_{0.975} \end{aligned}$$

```r
qnorm(0.975)
```

```
[1] 1.959964
```

Calculating the Margin of Error

$$\begin{aligned} \text{marginoferror} &= Z^* \times SE \\ &= 1.96 \times 0.011 \\ &= 0.022 \end{aligned}$$

Calculating the Margin of Error

$$
\begin{aligned}
\text{marginoferror} &= Z^* \times SE \\
&= 1.96 \times 0.011 \\
&= 0.022
\end{aligned}
$$

```
qnorm(0.975) * sqrt((0.57 * (1 - 0.57)) / 2027)
```

```
[1] 0.02155226
```

Finding the boundaries

```
(1155 / 2027) - qnorm(0.975) * sqrt((0.57 * (1 - 0.57)) / 2
```

```
[1] 0.5482553
```

```
(1155 / 2027) + qnorm(0.975) * sqrt((0.57 * (1 - 0.57)) / 2
```

```
[1] 0.5913599
```

Interpreting the Confidence Interval

- With 95% confidence, 54.8% to 59.1% of Americans have a favorable opinion of Tom Hanks.

Interpreting the Confidence Interval

- With 95% confidence, 54.8% to 59.1% of Americans have a favorable opinion of Tom Hanks.
- 57.0% of Americans have a favorable opinion of Tom Hanks (95% CI: 54.8-59.1%).

Interpreting the Confidence Interval

- With 95% confidence, 54.8% to 59.1% of Americans have a favorable opinion of Tom Hanks.
- 57.0% of Americans have a favorable opinion of Tom Hanks (95% CI: 54.8-59.1%).
- With 95% confidence, 57.0% $\pm$ 2.2% have a favorable opinion of Tom Hanks.

Testing the Hypothesis

1. Find the sampling distribution $N(p, SE)$ under the null hypothesis.

Testing the Hypothesis

1. Find the sampling distribution $N(p, SE)$ under the null hypothesis.
2. Calculate the test statistic for the null hypothesis.

$$Z = \frac{\hat{p} - p}{SE}$$

Testing the Hypothesis

1. Find the sampling distribution $N(p, SE)$ under the null hypothesis.
2. Calculate the test statistic for the null hypothesis.

$$Z = \frac{\hat{p} - p}{SE}$$

3. Find the probability of getting a test statistic as extreme or more extreme as this one, assuming the null hypothesis is true.

Testing the Hypothesis

1. Find the sampling distribution $N(p, SE)$ under the null hypothesis.
2. Calculate the test statistic for the null hypothesis.

$$Z = \frac{\hat{p} - p}{SE}$$

3. Find the probability of getting a test statistic as extreme or more extreme as this one, assuming the null hypothesis is true.
4. If the p-value is less than α, reject H_0 and accept H_A.

Finding the Null Distribution

To find the null distribution, replace the sample statistic $\hat{p} = 0.57$ with the population parameter $p = 0.5$.

$$
\begin{aligned}
SE &= \sqrt{\frac{p(1-p)}{n}} \\
&= \sqrt{\frac{0.5(1-0.5)}{2027}} \\
&= 0.011
\end{aligned}
$$

Finding the Null Distribution

To find the null distribution, replace the sample statistic $\hat{p} = 0.57$ with the population parameter $p = 0.5$.

$$
\begin{aligned}
SE &= \sqrt{\frac{p(1-p)}{n}} \\
&= \sqrt{\frac{0.5(1-0.5)}{2027}} \\
&= 0.011
\end{aligned}
$$

The null hypothesis is that our observed sample statistic $\hat{p} = 0.57$ comes from the sampling distribution $N(\mu = 0.5, \sigma = 0.011)$.

Calculating the test statistic

If the sampling distribution under H_0 is $\hat{p} \sim N(0.5, 0.011)$, then the test statistic for $\hat{p} = 0.57$ is...

$$\begin{aligned} Z &= \frac{\hat{p} - p}{SE} \\ &= \frac{0.57 - 0.5}{0.011} \\ &= 6.36 \end{aligned}$$

Calculating the test statistic

If the sampling distribution under H_0 is $\hat{p} \sim N(0.5, 0.011)$, then the test statistic for $\hat{p} = 0.57$ is...

$$
\begin{aligned}
Z &= \frac{\hat{p} - p}{SE} \\
&= \frac{0.57 - 0.5}{0.011} \\
&= 6.36
\end{aligned}
$$

$\hat{p}$ is 6.36 standard errors greater than p under the null hypothesis.

Get the p-value

- Use the test statistic Z to find the two-sided probability $P(Z \geq 6.36 \text{ or } Z \leq -6.36)$.

```
pnorm(-6.36) + pnorm(6.36, lower.tail = F)
```

```
[1] 2.017537e-10
```

Get the p-value

- Use the test statistic Z to find the two-sided probability $P(Z \geq 6.36 \,\mathrm{or}\, Z \leq -6.36)$.

```
pnorm(-6.36) + pnorm(6.36, lower.tail = F)
```

```
[1] 2.017537e-10
```

- If H_0:$P(\mathrm{Favorable}) = 0.5$ were true, then the probability that we would see a sample proportion as different from $p = 0.5$ as $\hat{p} = 0.57$ is very low.

Get the p-value

- Use the test statistic Z to find the two-sided probability $P(Z \geq 6.36 \,\text{or}\, Z \leq -6.36)$.

```r
pnorm(-6.36) + pnorm(6.36, lower.tail = F)
```

```
[1] 2.017537e-10
```

- If H_0:$P(\text{Favorable}) = 0.5$ were true, then the probability that we would see a sample proportion as different from $p = 0.5$ as $\hat{p} = 0.57$ is very low.
- $P(|Z| \geq 6.36) < 0.05$, so we *reject* H_0 and *accept* H_A.

Difference Between 2 Proportions

- In 2018, Ipsos surveyed 1,005 Americans using the same questions, and 754 (75.0%) had a favorable opinion of Tom Hanks.

Difference Between 2 Proportions

- In 2018, Ipsos surveyed 1,005 Americans using the same questions, and 754 (75.0%) had a favorable opinion of Tom Hanks.
- Research Question: Has Tom Hanks' favorability dropped between 2018 and 2024?

Difference Between 2 Proportions

- In 2018, Ipsos surveyed 1,005 Americans using the same questions, and 754 (75.0%) had a favorable opinion of Tom Hanks.
- Research Question: Has Tom Hanks' favorability dropped between 2018 and 2024?
- Does $\hat{p}_{2018} = \hat{p}_{2024}$? Does $\hat{p}_{2018} - \hat{p}_{2024} = 0$?

Sampling Distribution

- Sample statistic is the difference between 2 sample proportions $\hat{p}_1 - \hat{p}_2$

Sampling Distribution

- Sample statistic is the difference between 2 sample proportions $\hat{p}_1 - \hat{p}_2$
- When assumptions met, $\hat{p}_1 - \hat{p}_2 \sim N(p_1 - p_2, SE_{p_1 - p_2})$

Standard Error

The standard error for the difference between 2 proportions requires the population proportion p and sample size n for each group.

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Assumptions

- Observations within each sample are independent of each other.

Assumptions

- Observations within each sample are independent of each other.
- There are 10+ successes and 10+ failures in each sample.

Assumptions

- Observations within each sample are independent of each other.
- There are 10+ successes and 10+ failures in each sample.
- Sample 1 is independent of Sample 2.

Hypotheses

Hypotheses

- H_0: The proportion of Americans who view Tom Hanks favorably did not change between 2018 and 2024.

Hypotheses

- H_0: The proportion of Americans who view Tom Hanks favorably did not change between 2018 and 2024.

$$P(\mathrm{Favorable in 2018}) = P(\mathrm{Favorable in 2024})$$

Hypotheses

- H_0: The proportion of Americans who view Tom Hanks favorably did not change between 2018 and 2024.

$$P(\mathrm{Favorable in 2018}) = P(\mathrm{Favorable in 2024})$$

- H_A: The proportion of Americans who view Tom Hanks favorably changed between 2018 and 2024.

Hypotheses

- H_0: The proportion of Americans who view Tom Hanks favorably did not change between 2018 and 2024.

$$P(\text{Favorablein2018}) = P(\text{Favorablein2024})$$

- H_A: The proportion of Americans who view Tom Hanks favorably changed between 2018 and 2024.

$$P(\text{Favorablein2018}) \neq P(\text{Favorablein2024})$$

Hypotheses

- H_0: The proportion of Americans who view Tom Hanks favorably did not change between 2018 and 2024.

$$P(\text{Favorablein2018}) - P(\text{Favorablein2024}) = 0$$

- H_A: The proportion of Americans who view Tom Hanks favorably changed between 2018 and 2024.

$$P(\text{Favorablein2018}) - P(\text{Favorablein2024}) \neq 0$$

95% Confidence Interval

1. Find the point estimate for the difference in proportions $\hat{p}_1 - \hat{p}_2$.

95% Confidence Interval

1. Find the point estimate for the difference in proportions $\hat{p}_1 - \hat{p}_2$.
2. Calculate the standard error for the difference between 2 proportions.

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

95% Confidence Interval

1. Find the point estimate for the difference in proportions $\hat{p}_1 - \hat{p}_2$.
2. Calculate the standard error for the difference between 2 proportions.

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

2. Find Z^* for $\alpha = 0.05$.

$$Z^* = Z_{1-\alpha/2}$$

95% Confidence Interval

1. Find the point estimate for the difference in proportions $\hat{p}_1 - \hat{p}_2$.
2. Calculate the standard error for the difference between 2 proportions.

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

2. Find Z^* for $\alpha = 0.05$.

$$Z^* = Z_{1-\alpha/2}$$

3. Construct confidence interval as $\hat{p}_1 - \hat{p}_2 \pm Z^* \times SE$

Point Estimate

$$\hat{p}_1 - \hat{p}_2 = \frac{754}{1005} - \frac{1155}{2027}$$
$$= 0.180$$

Standard Error

When constructing a confidence interval for $\hat{p}_1 - \hat{p}_2$, we use the sample proportions $\hat{p}_1$ and $\hat{p}_2$ as estimates for the population parameters p_1 and p_2.

Standard Error

When constructing a confidence interval for $\hat{p}_1 - \hat{p}_2$, we use the sample proportions $\hat{p}_1$ and $\hat{p}_2$ as estimates for the population parameters p_1 and p_2.

$$
\begin{aligned}
SE_{(\hat{p}_1 - \hat{p}_2)} &= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\
&= \sqrt{\frac{0.75(1-0.75)}{1005} + \frac{0.57(1-0.57)}{2027}} \\
&= 0.018
\end{aligned}
$$

Finding Z^*

$$
\begin{aligned}
Z^* &= Z_{1-\alpha/2} \\
&= Z_{1-0.025} \\
&= Z_{0.975}
\end{aligned}
$$

Finding Z^*

$$
\begin{aligned}
Z^* &= Z_{1-\alpha/2} \\
&= Z_{1-0.025} \\
&= Z_{0.975}
\end{aligned}
$$

```
qnorm(0.975)
```

```
[1] 1.959964
```

Calculating the Margin of Error

$$\begin{aligned} \text{margin of error} &= Z^* \times SE \\ &= 1.96 \times 0.018 \\ &= 0.034 \end{aligned}$$

Calculating the Margin of Error

$$
\begin{aligned}
\text{marginoferror} &= Z^* \times SE \\
&= 1.96 \times 0.018 \\
&= 0.034
\end{aligned}
$$

```
qnorm(0.975) * sqrt((0.75*(1-0.75))/1005 + (0.57*(1-0.57))/
```

```
[1] 0.03436845
```

Finding the boundaries

```
diff <- ((754 / 1005) - (1155 / 2027))

margin <- qnorm(0.975) * sqrt((0.75*(1-0.75))/1005 + (0.57*

diff - margin
```

```
[1] 0.1460727
```

```
diff + margin
```

```
[1] 0.2148096
```

Interpreting the Confidence Interval

- With 95% confidence, 14.6% to 21.5% more Americans had a favorable opinion of Tom Hanks in 2018 than in 2024.

Interpreting the Confidence Interval

- With 95% confidence, 14.6% to 21.5% more Americans had a favorable opinion of Tom Hanks in 2018 than in 2024.
- 18.0% more Americans had a favorable opinion of Tom Hanks in 2018 than in 2024 (95% CI: 14.6-21.5%).

Interpreting the Confidence Interval

- With 95% confidence, 14.6% to 21.5% more Americans had a favorable opinion of Tom Hanks in 2018 than in 2024.
- 18.0% more Americans had a favorable opinion of Tom Hanks in 2018 than in 2024 (95% CI: 14.6-21.5%).
- With 95% confidence, 18.0% $\pm$ 3.4% more Americans had a favorable opinion of Tom Hanks in 2018 than 2024.

Testing the Hypothesis

1. Find the sampling distribution $N(0, SE_{p_1 - p_2})$ under the null hypothesis.

Testing the Hypothesis

1. Find the sampling distribution $N(0, SE_{p_1 - p_2})$ under the null hypothesis.
2. Calculate the test statistic for the null hypothesis $p_1 = p_2$.

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{p_1 = p_2}}$$

Testing the Hypothesis

1. Find the sampling distribution $N(0, SE_{p_1 - p_2})$ under the null hypothesis.
2. Calculate the test statistic for the null hypothesis $p_1 = p_2$.

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{p_1 = p_2}}$$

3. Find the probability of getting a test statistic as extreme or more extreme as this one, assuming the null hypothesis is true.

Testing the Hypothesis

1. Find the sampling distribution $N(0, SE_{p_1 - p_2})$ under the null hypothesis.
2. Calculate the test statistic for the null hypothesis $p_1 = p_2$.

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{p_1 = p_2}}$$

3. Find the probability of getting a test statistic as extreme or more extreme as this one, assuming the null hypothesis is true.
4. If the p-value is less than α, reject H_0 and accept H_A.

Pooled Population Proportion

When the null hypothesis is that $p_1 = p_2$ or $p_1 - p_2 = 0$, we use the pooled population parameter $\hat{p}$ to calculate the standard error.

$$\begin{aligned}\hat{p}_{\text{pooled}} &= \frac{\text{count}_1 + \text{count}_2}{n_1 + n_2} \\ &= \frac{754 + 1155}{1005 + 2027} \\ &= 0.630\end{aligned}$$

Pooled Population Proportion

When the null hypothesis is that $p_1 = p_2$ or $p_1 - p_2 = 0$, we use the pooled population parameter $\hat{p}$ to calculate the standard error.

$$\begin{aligned}\hat{p}_{\text{pooled}} &= \frac{\text{count}_1 + \text{count}_2}{n_1 + n_2} \\ &= \frac{754 + 1155}{1005 + 2027} \\ &= 0.630\end{aligned}$$

When H_0: $p_1 - p_2 \neq 0$, then you use $\hat{p}_1$ and $\hat{p}_2$ when calculating the standard error for the hypothesis test.

Finding the Null Distribution

To find the null distribution for $p_1 = p_2$, replace the sample statistics $\hat{p}_1$ and $\hat{p}_2$ with the pooled population proportion $\hat{p}$.

$$
\begin{aligned}
SE_{(p_1=p_2)} &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}} \\
&= \sqrt{\frac{0.630(1-0.630)}{1005} + \frac{0.630(1-0.630)}{2027}} \\
&= 0.019
\end{aligned}
$$

Finding the Null Distribution

To find the null distribution for $p_1 = p_2$, replace the sample statistics $\hat{p}_1$ and $\hat{p}_2$ with the pooled population proportion $\hat{p}$.

$$
\begin{aligned}
SE_{(p_1=p_2)} &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}} \\
&= \sqrt{\frac{0.630(1-0.630)}{1005} + \frac{0.630(1-0.630)}{2027}} \\
&= 0.019
\end{aligned}
$$

The null hypothesis is that our observed sample statistic $\hat{p}_1 - \hat{p}_2$ comes from the sampling distribution $N(\mu = 0, \sigma = 0.019)$.

Calculating the test statistic

If the sampling distribution under H_0 is $\hat{p}_1 - \hat{p}_2 \sim N(0, 0.019)$, then the test statistic for $\hat{p}_1 - \hat{p}_2 =$ is...

$$
\begin{aligned}
Z &= \frac{\hat{p} - p}{SE} \\
&= \frac{0.750 - 0.570}{0.019} \\
&= 9.47
\end{aligned}
$$

Calculating the test statistic

If the sampling distribution under H_0 is $\hat{p}_1 - \hat{p}_2 \sim N(0, 0.019)$, then the test statistic for $\hat{p}_1 - \hat{p}_2 =$ is...

$$\begin{aligned} Z &= \frac{\hat{p} - p}{SE} \\ &= \frac{0.750 - 0.570}{0.019} \\ &= 9.47 \end{aligned}$$

$\hat{p}_1 - \hat{p}_2$ is 9.47 standard errors greater than $\hat{p}_1 - \hat{p}_2 = 0$ under the null hypothesis.

Get the p-value

- Use the test statistic Z to find the two-sided probability $P(Z \geq 9.47 \text{ or } Z \leq -9.47)$.

```
pnorm(-9.47) + pnorm(9.47, lower.tail = F)
```

```
[1] 2.798437e-21
```

Get the p-value

- Use the test statistic Z to find the two-sided probability $P(Z \geq 9.47 \text{ or } Z \leq -9.47)$.

```
pnorm(-9.47) + pnorm(9.47, lower.tail = F)
```

```
[1] 2.798437e-21
```

- If H_0:$p_{2018} = p_{2024}$ were true, then the probability that we would see a difference as large as 18.0% is very small.

Get the p-value

- Use the test statistic Z to find the two-sided probability $P(Z \geq 9.47 \text{ or } Z \leq -9.47)$.

```
pnorm(-9.47) + pnorm(9.47, lower.tail = F)
```

```
[1] 2.798437e-21
```

- If H_0:$p_{2018} = p_{2024}$ were true, then the probability that we would see a difference as large as 18.0% is very small.
- $P(|Z| \geq 9.47) < 0.05$, so we *reject* H_0 and *accept* H_A.