STAT011 Statistical Methods I

Lecture 12 Hypothesis Testing

Lu Chen
Swarthmore College
2/28/2019

Review - Central Limit Theorem

- Population distribution is Normal, $X \sim N(\mu, \sigma)$,

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Population distribution is not Normal, $\mu_X = \mu$, $\sigma_X = \sigma$,

$$\bar{x} \overset{approx.}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

  - Population distribution is Bernoulli, $X \sim Bernoulli(p)$, $\mu_X = p$, $\sigma_X = \sqrt{p(1-p)}$,

$$\hat{p} \overset{approx.}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Review - Level C Confidence interval

- Statistical inference: *Confidence interval* and *hypothesis testing*
- Confidence interval
  - A simple simulation study
  - Margin of error $m = z^* \frac{\sigma}{\sqrt{n}}$ and critical points z^* and $-z^*$
  - Confidence interval for a population mean
- Level C confidence interval for
  - A population mean $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$
  - A population proportion $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Confidence level C	**0.9**	**0.95**	**0.99**
Critical point z^*	1.645	1.960	2.576

Review - Statistical inference

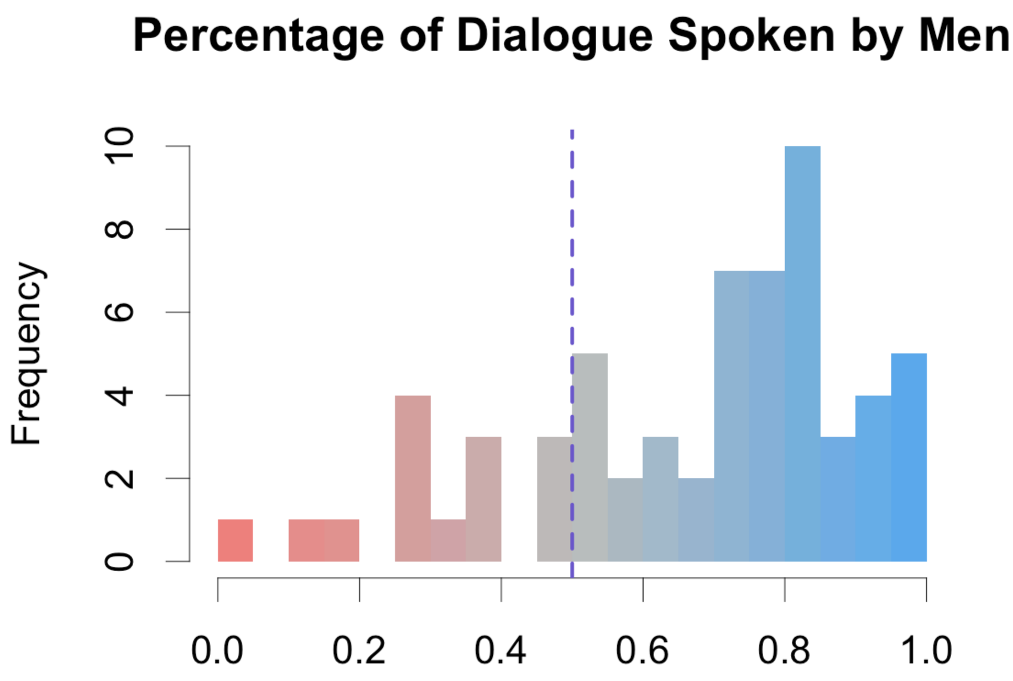
> **Statistical inference** uses a fact about a sample to estimate the truth about the whole population.

- **Confidence interval**
  - Assesses how well the sample statistic estimates the population parameter.
- **Hypothesis testing** (or **significance test**)
  - Assesses the evidence provided by the data in favor of **some claim about the population parameter**.

Outline

- Hypothesis testing
  1. Null and alternative hypotheses
  2. Test statistic
  3. P-value
  4. Statistical significance (conclusion)
- z test for a population mean
  - Definition
  - Examples

Data example - 62 screenplays in 2015



- $\bar{x} = 0.668$
- 95% confidence interval: $[0.619, 0.717]$
- We are 95% confident about the method that the interval $[0.619, 0.717]$ will contain the true population average of percentage of dialogue spoken by men in 2015 movies.
- It is extremely unlikely that the true population mean is 0.5. Therefore, statistically, men speak more in 2015 movies.

Null and alternative hypotheses

- Question of interest
  - **Statistically, do men truly speak more dialogue than women in 2015 movies?**
- Denote X as the amount of dialogue (in percentage) spoken by men in 2015 movies.
- Question of interest becomes
  - **Is mean of X greater than 0.5?**
- In hypothesis testing, we usually state a **null hypothesis** and an **alternative hypothesis** to test our question of interest.

Null and alternative hypotheses

Null hypothesis (H_0): the statement being tested in hypothesis testing. Hypothesis testing is designed to assess the strength of the evidence against the null hypothesis. Usually the null hypothesis is a statement of **no effect** or **no difference**.

- $H_0 : \mu = \mu_0$ (μ_0 is the hypothesized value)

Alternative hypothesis (H_a): the statement we hope or suspect is true in hypothesis testing.

- $H_a : \mu \neq \mu_0$
- $H_a : \mu > \mu_0$
- $H_a : \mu < \mu_0$

- Tests with $H_a : \mu \neq \mu_0$ are called **two-sided tests**. Tests with $H_a : \mu > \mu_0$ or $\mu < \mu_0$ are called **one-sided tests**.

Null and alternative hypotheses

Question of interest: Is mean of X greater than 0.5?

Null hypothesis

- $H_0 : \mu = 0.5$

Alternative hypothesis

- $H_a : \mu > 0.5$
- This is a one-sided test.

Note

- If H_0 is true, we expect the sample mean is close to the hypothesized value 0.5.
- The difference between the sample mean and the hypothesized value gives evidence against H_0.

Test statistic

The test will be based on the sample mean $\bar{x}$ that is used to estimate the population mean μ

- $H_0 : \mu = 0.5; H_a : \mu > 0.5$
- $\bar{x} = 0.668$ for the $n = 62$ screenplays in 2015.
- Assume SD $\sigma = 0.197$ is known. What is the distribution of $\bar{x}$ if H_0 is true?
- $$\bar{x} \overset{approx.}{\sim} N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right) = N\left(0.5, \frac{0.197}{\sqrt{62}}\right) = N(0.5, 0.025)$$
- The **test statistic** of hypothesis testing calculates how many standard deviations away the sample mean $\bar{x}$ is from the hypothesized value μ_0. If the distance is large, we reject H_0; otherwise, we cannot reject H_0.

Test statistic

- $H_0 : \mu = 0.5; H_a : \mu > 0.5; \bar{x} = 0.668, n = 62.$

The **test statistic** is defined as the difference between $\bar{x}$ and μ_0 divided by the SD of $\bar{x}$,

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{0.668 - 0.5}{0.025} = 6.72$$

- This is called a z **statistic**. It measures how many standard deviations $\bar{x}$ is away from μ_0.
- **What is the distribution of z?**
- Since $\bar{x} \overset{approx.}{\sim} N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$, $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \overset{approx.}{\sim} N(0, 1)$.
- The larger the z statistic, the stronger evidence against H_0. But how large is large?

P-value

- When z is 2 or 3, it means that $\bar{x}$ is 2 or 3 standard deviations away from μ_0, which is definitely large enough for us to say that the true μ is actually not μ_0.
  - In this example, $z = 6.72$.
- Another way to thinking about this: we can calculate the probability that z takes a value as extreme or more extreme than 6.72 and see whether this probability is too small.

> The probability, assuming H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the ***P*-value** of the test. **The smaller the P-value, the stronger the evidence against H_0 provided by the data**.

P-value

- $H_0 : \mu = 0.5; H_a : \mu > 0.5; n = 62; \bar{x} = 0.668.$
- $\bar{x} \overset{approx.}{\sim} N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right) = N(0.5, 0.025)$
- $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{0.668-0.5}{0.025} = 6.72 \overset{approx.}{\sim} N(0, 1)$
- We calculate P-value as the probability that $z \geq 6.72$

```r
1 - pnorm(6.72)
```

```
## [1] 9.086176e-12
```

- The P-value is 9×10^{-12}. It is the probability that $z \geq 6.72$, which is extremely small. It means that assuming H_0 is true ($\mu = 0.5$), it is almost impossible for us to observe data that results in z as extreme or more extreme than 6.72. Then it is most likely that H_0 is not true.
- Usually, when P-value is less than or equal to 0.1, or 0.05, or 0.01, we think it is small enough to reject H_0.

Statistical significance

Denote α as the **significance level**, which is the decisive value to reject the null hypothesis H_0. α is usually 0.05, sometimes 0.01 or 0.1.

If the P-value is as small or smaller than α, we say that the data are **statistically significant** at level α.

- In our example, P-value $= 9 \times 10^{-12} < 0.05$. The data show strong evidence that the sample mean $\bar{x} = 0.668$ is greater than 0.5. So we reject the null hypothesis.

Hypothesis testing

1. **Null and alternative hypotheses.**
   - $H_0 : \mu = 0.5; H_a : \mu > 0.5.$
2. **Test statistic.**
   - $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{0.668-0.5}{0.197/\sqrt{62}} = 6.72 \overset{approx.}{\sim} N(0, 1)$
3. ***P*-value.**
   - $P(Z \geq 6.72) = 9 \times 10^{-12}$ (Z denotes a standard Normal variable)
4. **Statistical significance (conclusion).**
   - P-value is less than $\alpha = 0.05$. We reject the null hypothesis that $\mu = 0.5$ at level $\alpha = 0.05$ and conclude that the true population mean percentage of dialogue spoken by men in 2015 movies is **statistically significantly larger than 0.5 at level 0.05**.

Hypothesis testing

1. **Null and alternative hypotheses**. State H_0 and H_a. The test is designed to assess the strength of the evidence against H_0; H_a is the statement that we will accept if the evidence enables us to reject H_0.
2. **Test statistic**. Calculate the value of the test statistic on which the test will be based. This statistic usually measures how far the data are from H_0.
3. ***P*-value**. Find the P-value, which is the probability, calculated assuming that H_0 is true, that the test statistic will weigh against H_0 at least as strongly as it does for these data.
4. **Statistical significance (conclusion)**. State a conclusion. If the P-value is less than or equal to the significance level α, the alternative hypothesis is true; if it is greater than α, the data do not provide sufficient evidence to reject H_0.

z Test for a population mean

To test the hypothesis $H_0 : \mu = \mu_0$ based on an SRS of size n from a population with unknown mean μ and known standard deviation σ, compute the **test statistic**

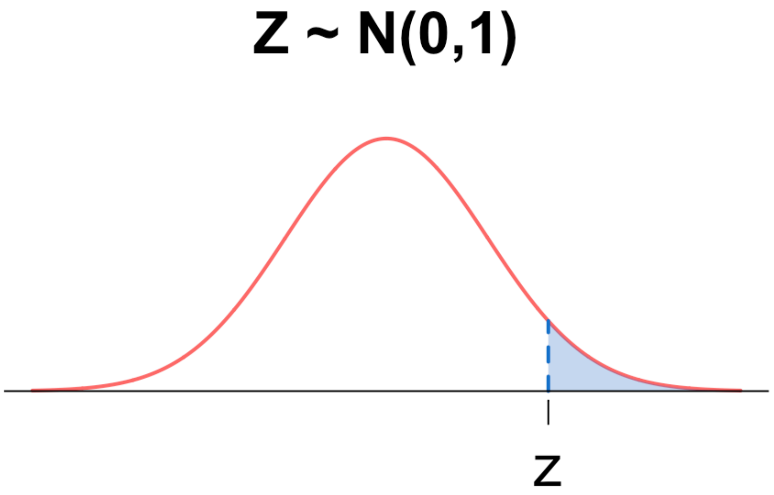
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

In terms of a standard Normal random variable Z, the P-value for a test of H_0 vs.

$$\begin{aligned} H_a : \mu > \mu_0 &\text{ is } P(Z \geq z) \\ H_a : \mu < \mu_0 &\text{ is } P(Z \leq z) \\ H_a : \mu \neq \mu_0 &\text{ is } 2P(Z \geq |z|) \end{aligned}$$

These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

P-value



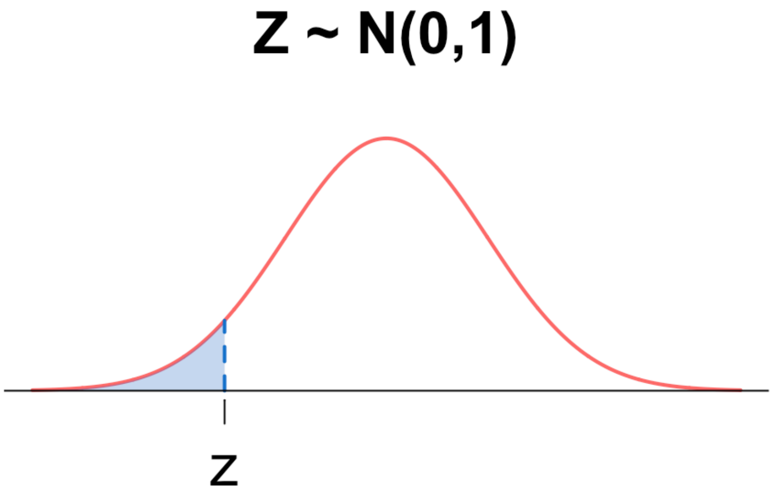
$H_0 : \mu = \mu_0$; $\boldsymbol{H_a : \mu > \mu_0}$

$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \overset{approx.}{\sim} N(0, 1)$

P-value $= P(Z \geq z)$

In H_a, we suspect μ is **greater** than μ_0. Then we expect a sample mean larger than μ_0. Therefore, for P-value, we calculate the probability that the standard Normal variable Z takes a value that is **greater** than z.

P-value



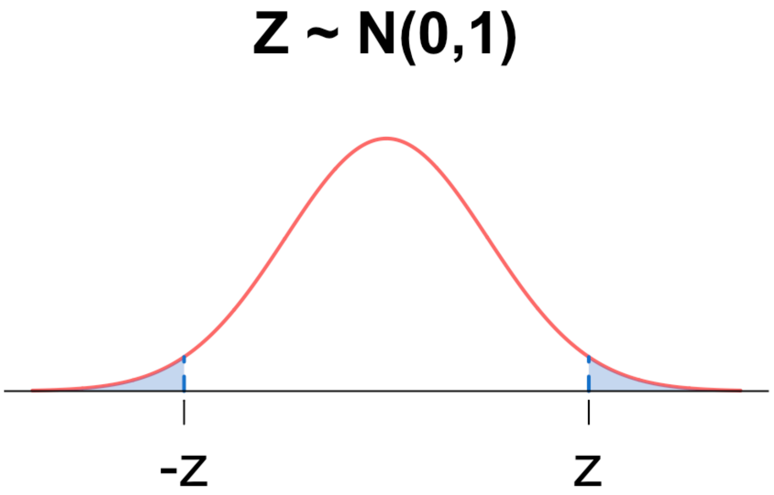
$H_0 : \mu = \mu_0$; $\boldsymbol{H_a : \mu < \mu_0}$

$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \overset{approx.}{\sim} N(0, 1)$

P-value $= P(Z \leq z)$

In H_a, we suspect μ is **smaller** than μ_0. Then we expect a sample mean smaller than μ_0. Therefore, for P-value, we calculate the probability that the standard Normal variable Z takes a value that is **smaller** than z.

P-value



$H_0 : \mu = \mu_0$; $\boldsymbol{H_a : \mu \neq \mu_0}$

$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \overset{approx.}{\sim} N(0, 1)$

P-value

$= P(Z \geq z) + P(Z \leq -z) = 2P(Z \geq |z|)$

In H_a, we suspect μ is either **greater or smaller** than μ_0. Then we expect a sample mean either greater or smaller than μ_0. Therefore, for P-value, we calculate the probability that the standard Normal variable Z takes a value that is **greater or smaller** than z.

Example 1

Do men and women have similar amount of dialogue in 2015 movies?

- $H_0 : \mu = 0.5$, men and women have similar amount of dialogue.
- $H_a : \mu \neq 0.5$, men and women have different amount of dialogue.
- Test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{0.668 - 0.5}{0.197/\sqrt{62}} = 6.72 \overset{approx.}{\sim} N(0, 1)$$

- If not mentioned, use $\alpha = 0.05$.
- P-value: $P = 2P(Z \geq 6.72) = 1.8 \times 10^{-11} < 0.05$(`2*(1-pnorm(6.72))`)
- Conclusion: P-value is less than $\alpha = 0.05$. We reject the null hypothesis that $\mu = 0.5$ at level $\alpha = 0.05$ and conclude that men and women have **significantly different** amount of dialogue in 2015 movies.

Example 2

Suppose a study was done to determine if the average amount of sleep that students get the day before an exam is less than 6 hours. An SRS of 100 students from a university was taken and a mean of 5.5 hours was computed from the sample. Conduct a test at level 0.01 with population standard deviation 2.5 hours.

- $\mu_0 = 6, n = 100, \bar{x} = 5.5, \sigma = 2, \alpha = 0.01$
- $H_0: \mu = 6$, the amount of sleep before exam is 6 hours.
- $H_a: \mu < 6$, the amount of sleep before exam is less than 6 hours.
- $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{5.5-6}{2.5/\sqrt{100}} = -2 \overset{approx.}{\sim} N(0, 1)$
- $P = P(Z \leq -2) = 0.023 > 0.01$(`pnorm(-2)`)
- We cannot reject the null hypothesis that $\mu = 6$ at level $\alpha = 0.01$. The amount of sleep before exam is not significantly less than 6 hours.
- Note: for this question, if $\alpha = 0.05$, what is the conclusion?

Example 3

The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures the motivation, attitude, and study habits of college students. Scores range from 0 to 200 and follow (approximately) a Normal distribution with a mean of 115 and a standard deviation of 25. You suspect that incoming freshmen at Swarthmore have a mean greater than 115 because they are often excited about entering Swarthmore. To test your suspicion, you give the SSHA to 25 incoming freshmen and find their mean score to be 130.

- $\mu_0 = 115, n = 25, \bar{x} = 130, \sigma = 25, \alpha = 0.05$
- $H_0 : \mu = 115$, mean SSHA score of Swarthmore freshmen is 115.
- $H_a : \mu > 115$, mean SSHA score of Swarthmore freshmen is greater than 115.
- $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{130-115}{25/\sqrt{25}} = 3 \overset{approx.}{\sim} N(0, 1)$
- $P = P(Z \geq 3) = 0.001 < 0.05$(`1-pnorm(3)`)
- We reject the null hypothesis that $\mu = 115$ at level $\alpha = 0.05$. The mean SSHA score of Swarthmore freshmen is significantly greater than 115.

Example 4

A student tossed a coin 20 times and got 7 heads. Is it a fair coin?

- $p_0 = 0.5\,, n = 20, \hat{p} = 0.35, \alpha = 0.05$
- $H_0 : p = 0.5$, this is a fair coin.
- $H_a : p \neq 0.5$, this is an unfair coin.
- $z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.35-0.5}{\sqrt{0.5\times0.5/20}} = -1.34 \overset{approx.}{\sim} N(0,1)$
- $P = 2P(Z \geq |-1.34|) = 0.18 > 0.05$(`2*(1-pnorm(1.34))`)
- We cannot reject the null hypothesis that $p = 0.5$ at level $\alpha = 0.05$. The data do not provide enough evidence to show that this is an unfair coin.

Summary

z Test			
H_0	$\mu = \mu_0$		
H_a	$\mu > \mu_0$	$\mu < \mu_0$	$\mu \neq \mu_0$
Test statistic	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \overset{approx.}{\sim} N(0, 1)$		
P**-value**	$P(Z \geq z)$ `1-pnorm(z)`	$P(Z \leq z)$ `pnorm(z)`	$2P(Z \geq \|z\|)$ `2*(1-pnorm(abs(z)))`
Conclusion	$P > \alpha$: statistically significant at level α; cannot reject H_0 $P \leq \alpha$: statistically insignificant at level α; reject H_0		

Midterm I

- **Thursday 3/7 during class time**
  - Practice problems: available on Friday 3/1
  - Review class on Tuesday 3/5
- Homework 6: due on Monday 3/4 11:59 pm
  - Solutions available on Tuesday 3/5
- DataCamp assignment: finish it by tonight and get partial credits.