Chapter 10: Inferences based on two samples

MATH 450

November 16^{th}, 2017

Overview

Week 1	Chapter 1: Descriptive statistics
Week 2	Chapter 6: Statistics and Sampling Distributions
Week 4	Chapter 7: Point Estimation
Week 7	Chapter 8: Confidence Intervals
Week 10	Chapter 9: Tests of Hypotheses
Week 12	**Chapter 10: Two-sample inference**

Overview

Two-sample inference: example

Example

Let μ_1 and μ_2 denote true average decrease in cholesterol for two drugs. From two independent samples $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$, we want to test:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

Settings

- This week: independent samples

Assumption

1. *$X_1, X_2, \ldots, X_m$ is a random sample from a population with mean μ_1 and variance σ_1^2.*
2. *$Y_1, Y_2, \ldots, Y_n$ is a random sample from a population with mean μ_2 and variance σ_2^2.*
3. *The X and Y samples are independent of each other.*

- Next week: paired-sample test

Problem

Assume that

- $X_1, X_2, \ldots, X_m$ *is a random sample from a population with mean* μ_1 *and variance* σ_1^2*.*
- $Y_1, Y_2, \ldots, Y_n$ *is a random sample from a population with mean* μ_2 *and variance* σ_2^2*.*
- *The* X *and* Y *samples are independent of each other.*

Compute (in terms of $\mu_1, \mu_2, \sigma_1, \sigma_2, m, n$*)*

(a) $E[\bar{X} - \bar{Y}]$

(b) $Var[\bar{X} - \bar{Y}]$ *and* $\sigma_{\bar{X}-\bar{Y}}$

Properties of $\bar{X} - \bar{Y}$

Proposition

The expected value of $\overline{X} - \overline{Y}$ is $\mu_1 - \mu_2$, so $\overline{X} - \overline{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$. The standard deviation of $\overline{X} - \overline{Y}$ is

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Normal distributions with known variances

Confidence intervals

When both population distributions are normal, standardizing $\overline{X} - \overline{Y}$ gives a random variable Z with a standard normal distribution. Since the area under the z curve between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is $1 - \alpha$, it follows that

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} < z_{\alpha/2}\right) = 1 - \alpha$$

Manipulation of the inequalities inside the parentheses to isolate $\mu_1 - \mu_2$ yields the equivalent probability statement

$$P\left(\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} < \mu_1 - \mu_2 < \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right) = 1 - \alpha$$

Testing the difference between two population means

- Setting: independent normal random samples $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ with known values of σ_1 and σ_2. Constant Δ_0.
- Null hypothesis:

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

- Alternative hypothesis:
  (a) $H_a : \mu_1 - \mu_2 > \Delta_0$
  (b) $H_a : \mu_1 - \mu_2 < \Delta_0$
  (c) $H_a : \mu_1 - \mu_2 \neq \Delta_0$
- When $\Delta = 0$, the test (c) becomes

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

Proposition

Null hypothesis: H_0: $\mu_1 - \mu_2 = \Delta_0$

Test statistic value: $z = \dfrac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}}$

Alternative Hypothesis	**Rejection Region for Level α Test**
H_a: $\mu_1 - \mu_2 > \Delta_0$	$z \geq z_\alpha$ (upper-tailed test)
H_a: $\mu_1 - \mu_2 < \Delta_0$	$z \leq -z_\alpha$ (lower-tailed test)
H_a: $\mu_1 - \mu_2 \neq \Delta_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed test)

Large-sample tests/confidence intervals (with unknown σ)

Principles

- Central Limit Theorem: $\bar{X}$ and $\bar{Y}$ are approximately normal when $n > 30 \rightarrow$ so is $\bar{X} - \bar{Y}$. Thus

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

  is approximately standard normal
- When n is sufficiently large $S_1 \approx \sigma_1$ and $S_2 \approx \sigma_2$
- Conclusion:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

  is approximately standard normal when n is sufficiently large

If $m, n > 40$, we can ignore the normal assumption and replace σ by S

Proposition

Use of the test statistic value

$$z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

along with the previously stated upper-, lower-, and two-tailed rejection regions based on z critical values gives large-sample tests whose significance levels are approximately α. These tests are usually appropriate if both $m > 40$ and $n > 40$. A P-value is computed exactly as it was for our earlier z tests.

Proposition

Provided that m and n are both large, a CI for $\mu_1 - \mu_2$ with a confidence level of approximately $100(1-\alpha)\%$ is

$$\bar{x} - \bar{y} \ \pm \ z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

where $-$gives the lower limit and $+$ the upper limit of the interval. An upper or lower confidence bound can also be calculated by retaining the appropriate sign and replacing $z_{\alpha/2}$ by z_α.

Example

Example

Let μ_1 and μ_2 denote true average tread lives for two competing brands of size P205/65R15 radial tires.

(a) Test

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

at level 0.05 using the following data: $m = 45$, $\bar{x} = 42,500$, $s_1 = 2200$, $n = 45$, $\bar{y} = 40,400$, and $s_2 = 1900$.

(b) Construct a 95% CI for $\mu_1 - \mu_2$.

The two-sample t test and confidence interval

Overview

- Section 8.1
  - Normal distribution
  - σ is known
- Section 8.2
  - ~~Normal distribution~~
    $\rightarrow$ Using Central Limit Theorem $\rightarrow$ needs $n > 30$
  - ~~σ is known~~
    $\rightarrow$ needs $n > 40$
- Section 8.3
  - Normal distribution
  - ~~σ is known~~
  - n is small

$\rightarrow$ Introducing t-distribution

Principles

- For one-sample inferences:

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

- For two-sample inferences:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \sim t_\nu$$

  where ν is some appropriate degree of freedom (which depends on m and n).

Chi-squared distribution

Proposition

- *If Z has standard normal distribution $\mathcal{Z}(0,1)$ and $X = Z^2$, then X has Chi-squared distribution with 1 degree of freedom, i.e. $X \sim \chi^2_1$ distribution.*
- *If $Z_1, Z_2, \ldots, Z_n$ are independent and each has the standard normal distribution, then*

$$Z_1^2 + Z_2^2 + \ldots + Z_n^2 \sim \chi^2_n$$

t distributions

Definition

Let Z be a standard normal rv and let W be a χ^2_ν rv independent of Z. Then the t distribution with degrees of freedom ν is defined to be the distribution of the ratio

$$T = \frac{Z}{\sqrt{W/\nu}}$$

2 plus 2 is 4 minus 1 that's 3

Definition of t distributions:

$$\frac{Z}{\sqrt{W/\nu}} \sim t_\nu$$

Our statistic:

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{m}+\frac{S_2^2}{n}}} = \frac{\left[(\bar{X}-\bar{Y})-(\mu_1-\mu_2)\right]/\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}{\sqrt{\left(\frac{S_1^2}{m}+\frac{S_2^2}{n}\right)/\left(\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}\right)}}$$

What we need:

$$\left(\frac{S_1^2}{m}+\frac{S_2^2}{n}\right)/\left(\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}\right) = \frac{W}{\nu}$$

Quick maths

- What we need:

$$\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right) = \left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)\frac{W}{\nu}$$

- What we have
  - $E[W] = \nu,\ V[W] = 2\nu$
  - $E[S_1^2] = \sigma_1^2,\ V[S_1^2] = 2\sigma_1^4/(m-1)$
  - $E[S_2^2] = \sigma_2^2,\ V[S_2^2] = 2\sigma_2^4/(n-1)$
- Variance of the LHS

$$V\left[\frac{S_1^2}{m} + \frac{S_2^2}{n}\right] = \frac{2\sigma_1^4}{(m-1)m^2} + \frac{2\sigma_2^4}{(n-1)n^2}$$

- Variance of the RHS

$$V\left[\left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)\frac{W}{\nu}\right] = \left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)^2\frac{2\nu}{\nu^2}$$

2-sample t test: degree of freedom

THEOREM When the population distributions are both normal, the standardized variable

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \qquad (10.2)$$

has approximately a t distribution with df v estimated from the data by

$$v = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}} = \frac{\left[(se_1)^2 + (se_2)^2\right]^2}{\frac{(se_1)^4}{m-1} + \frac{(se_2)^4}{n-1}}$$

where

$$se_1 = \frac{s_1}{\sqrt{m}} \qquad se_2 = \frac{s_2}{\sqrt{n}}$$

(round v down to the nearest integer).

CIs for difference of the two population means

The **two-sample t confidence interval for $\mu_1 - \mu_2$** with confidence level $100(1-\alpha)\%$ is then

$$\bar{x} - \bar{y} \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

A one-sided confidence bound can be calculated as described earlier.

The **two-sample t test** for testing H_0: $\mu_1 - \mu_2 = \Delta_0$ is as follows:

Test statistic value: $t = \dfrac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\dfrac{s_1^2}{m} + \dfrac{s_2^2}{n}}}$

Alternative Hypothesis	**Rejection Region for Approximate Level α Test**
H_a: $\mu_1 - \mu_2 > \Delta_0$	$t \geq t_{\alpha,\nu}$ (upper-tailed test)
H_a: $\mu_1 - \mu_2 < \Delta_0$	$t \leq -t_{\alpha,\nu}$ (lower-tailed test)
H_a: $\mu_1 - \mu_2 \neq \Delta_0$	either $t \geq t_{\alpha/2,\nu}$ or $t \leq -t_{\alpha/2,\nu}$ (two-tailed test)

A P-value can be computed as described in Section 9.4 for the one-sample t test.

Example

A paper reported the following data on tensile strength (psi) of liner specimens both when a certain fusion process was used and when this process was not used:

No fusion	2748	2700	2655	2822	2511			
	3149	3257	3213	3220	2753			
	$m = 10$	$\bar{x} = 2902.8$	$s_1 = 277.3$					
Fused	3027	3356	3359	3297	3125	2910	2889	2902
	$n = 8$	$\bar{y} = 3108.1$	$s_2 = 205.9$					

The authors of the article stated that the fusion process increased the average tensile strength. Carry out a test of hypotheses to see whether the data supports this conclusion (and provide the P-value of the test)

Solution

1. Let μ_1 be the true average tensile strength of specimens when the no-fusion treatment is used and μ_2 denote the true average tensile strength when the fusion treatment is used.

2. H_0: $\mu_1 - \mu_2 = 0$ (no difference in the true average tensile strengths for the two treatments)

3. H_a: $\mu_1 - \mu_2 < 0$ (true average tensile strength for the no-fusion treatment is less than that for the fusion treatment, so that the investigators' conclusion is correct)

Solution

4. The null value is $\Delta_0 = 0$, so the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

5. We now compute both the test statistic value and the df for the test:

$$t = \frac{2902.8 - 3108.1}{\sqrt{\frac{277.3^2}{10} + \frac{205.9^2}{8}}} = \frac{-205.3}{113.97} = -1.8$$

Using $s_1^2/m = 7689.529$ and $s_2^2/n = 5299.351$,

$$\nu = \frac{(7689.529 + 5299.351)^2}{\frac{(7689.529)^2}{9} + \frac{(5299.351)^2}{7}} = \frac{168{,}711{,}004}{10{,}581{,}747} = 15.94$$

so the test will be based on 15 df.