Mathematical statistics

May 8th, 2019

Lecture 30: The two-sample t-test

Att.

- Final exam:

  Wednesday, 5/29/2019, Wednesday, 10:30am –12:30pm
  Ewing Hall Room 101
- Course evaluation
- Last homework due this Friday
- Department colloquium

Overview

Week 1	Probability reviews
Week 2	Chapter 6: Statistics and Sampling Distributions
Week 4	Chapter 7: Point Estimation
Week 7	Chapter 8: Confidence Intervals
Week 10	Chapter 9, 10: Test of Hypothesis

Inferences based on two samples

10.1 Difference between two population means

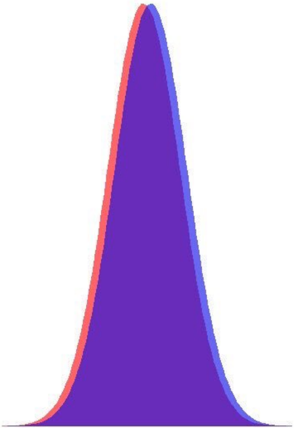
- z-test
- confidence intervals

10.2 The two-sample t test and confidence interval

10.3 Analysis of paired data

Two-sample inference

Difference between two population means



- Testing:

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

- Works well, even if $|\mu_1 - \mu_2| << \sigma_1, \sigma_2$

Settings

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.*

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

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

Assume further that the distributions of X and Y are normal and σ_1, σ_2 are known:

Problem

(a) *What is the distribution of*

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

(b) *Compute*

$$P\left[-1.96 \le \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \le 1.96\right]$$

(c) *Construct a 95% CI for $\mu_1 - \mu_2$ (in terms of $\bar{x}$, $\bar{y}$, m, n, σ_1, σ_2).*

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)

Sample solution

1. The parameter of interest is $\mu_1 - \mu_2$, the difference between true mean GPA for the <10 (conceptual) population and true mean GPA for the ≥ 10 population.

2. The null hypothesis is H_0: $\mu_1 - \mu_2 = 0$.

3. The alternative hypothesis is H_a: $\mu_1 - \mu_2 \neq 0$; if H_a is true then μ_1 and μ_2 are different. Although it would seem unlikely that $\mu_1 - \mu_2 > 0$ (those with low study hours have higher mean GPA) we will allow it as a possibility and do a two-tailed test.

4. With $\Delta_0 = 0$, the test statistic value is

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

5. The inequality in H_a implies that the test is two-tailed. For $\alpha = .05$, $\alpha/2 = .025$ and $z_{\alpha/2} = z_{.025} = 1.96$. H_0 will be rejected if $z \geq 1.96$ or $z \leq -1.96$.

Solution

6. Substituting $m = 10$, $\bar{x} = 2.97$, $\sigma_1^2 = .36$, $n = 11$, $\bar{y} = 3.06$, and $\sigma_2^2 = .36$ into the formula for z yields

$$z = \frac{2.97 - 3.06}{\sqrt{\frac{.36}{10} + \frac{.36}{11}}} = \frac{-.09}{.262} = -.34$$

   That is, the value of $\bar{x} - \bar{y}$ is only one-third of a standard deviation below what would be expected when H_0 is true.

7. Because the value of z is not even close to the rejection region, there is no reason to reject the null hypothesis. This test shows no evidence of any relationship between study hours and GPA. ■

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.

Large-sample CIs

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

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

Remember Chapter 8?

- 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

- 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).

2-sample t test: degree of freedom

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

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

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

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

where

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

(round ν 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.

2-sample t procedures

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

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

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

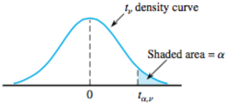
A paper reported the following data on tensile strength (psi) of liner specimens both when a certain fusion process was used and when this process was not used:

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. With confidence level $\alpha = 0.05$, carry out a test of hypotheses to see whether the data supports this conclusion (and provide the P-value of the test)

t-table

Table A.5 Critical Values for *t* Distributions



ν \ α	.10	.05	.025	.01	.005	.001	.0005
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745

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

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

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

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

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

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

so the test will be based on 15 df.

Example

The following data summarizes the proportional stress limits for specimens constructed using two different types of wood:

Type of wood	Sample size	Sample mean	Sample sd
Red oak	14	8.48	0.79
Douglas fir	10	6.65	1.28

Assuming that both samples were selected from normal distributions, carry out a test of hypotheses with significance level $\alpha = 0.05$ to decide whether the true average proportional stress limit for red oak joints exceeds that for Douglas fir joints by more than 1 MPa. Provide the P-value of the test.

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. $X \sim \chi_1^2$ distribution.*
- *If $Z_1, Z_2, \ldots, Z_n$ are independent and each has the standard normal distribution, then*

$$Z_1^2 + Z_2^2 + \ldots + Z_n^2 \sim \chi_n^2$$

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

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

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

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

- Variance of the RHS

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