Lec 37

Example 104.

Frequently, patients must wait a long time for elective surgery. Suppose the wait time for patients needing a knee replacement at two hospitals was investigated. Independent random samples of patients were obtained and the wait time for each (in weeks) was recorded. The resulting summary statistics are given in the following table.

Hospital	Sample Size	Sample Mean	Sample Variance
Hospital 1	15	17.4	34.81
Hospital 2	17	12.1	46.24

$n_1 = 15$ $\quad S_1^2 = 34.81$

$n_2 = 17$ $\quad S_2^2 = 46.24$

$\bar{X}_1 = 17.4$

$\bar{X}_2 = 12.1$

(a) Do the samples give enough evidence to conclude that the mean wait time at hospital 1 is longer than that of hospital 2? Assume that the unknown population variances are equal and populations are normal and test using $\alpha = 0.05$. [12 points]

(b) Find bounds on the p-value associated with the test. [2 points]

(c) How much is the mean wait time for hospital 1 longer than the mean for hospital 2? Calculate a 90% confidence interval to estimate the true value of $\mu_1 - \mu_2$. [4 points]

(d) Besides the population variances being equal, what else must be true (or what else must we assume) about the populations for parts (a) and (b) to be valid? [2 points]

$H_0: \mu_1 = \mu_2$ $\quad \mu_1 - \mu_2 = 0$ $\qquad H_a: \mu_1 > \mu_2$

$$t = \frac{17.4 - 12.1 - 0}{\sqrt{S_p^2 \cdot \left(\frac{1}{15} + \frac{1}{17}\right)}} = 2.339$$

$$S_p^2 = \frac{14 \times 34.81 + 16 \times 46.24}{15 + 17 - 2} = 40.906$$

$S_p = 6.4$

$t_{0.05,\ \underbrace{15+17-2}_{30}} = 1.697$

p-value

reject if $2.339 > 1.697$

2.339

we do reject

P-value b/w 0.01 & 0.025

90% CI :

$$\bar{X}_1 - \bar{X}_2 \pm t_{0.05,30}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$17.4 - 12.1 \pm 1.697\sqrt{40.906 \times \left(\frac{1}{15}+\frac{1}{17}\right)}$$

9.3 Analysis of Paired Data

We are interested in the difference between two observations of each subject. Suppose the data consists of n independently selected pairs $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$, with $E(X_i) = \mu_1$ and $E(Y_i) = \mu_2$. Let $D_1 = X_1 - Y_1$, $D_2 = X_2 - Y_2$ $\cdots$ $D_n = X_n - Y_n$, so the D_i's are the difference within pairs.

The Paired t test

Because different pairs are independent, the D_i's are independent of one another. Let $D = X - Y$, where X and Y are the first and second observations, respectively, within an arbitrary pair. Then the expected difference is

$$\mu_D = E[X - Y] = E[X] - E[Y] = \mu_1 - \mu_2$$

Null Hypothesis: $H_0: \mu_D = \Delta$

$\bar{D} = \bar{X} - \bar{Y}$

Test statistic value:

large n: $$Z = \frac{\bar{D} - \Delta}{S_D/\sqrt{n}}$$

S_D^2, sample variance for D_i's

small n: $$T = \frac{\bar{D} - \Delta}{S_D/\sqrt{n}}$$

Significance level:

Alternative Hypothesis: reject if

$H_a: \mu_D > \Delta$ $\quad Z > z_\alpha$ or $T > t_{\alpha, n-1}$

$H_a: \mu_D < \Delta$ $\quad Z < -z_\alpha$ or $T < -t_{\alpha, n-1}$

$H_a: \mu_D \neq \Delta$ $\quad |Z| > z_{\alpha/2}$ or $|T| > t_{\alpha/2, n-1}$

Assumptions: The D_i's constitute a random sample from a normal "difference" population (if n small).

Example 105.

Seven patients who claim to be suffering from job related stress were selected at random. After an initial resting pulse rate (in beats per minute) was obtained, each person participated in a relaxation therapy program. A final resting pulse rate was taken at the end of the program. The data is given in the following table.

Subject	1	2	3	4	5	6	7
Initial Pulse Rate	67	71	67	84	71	68	73
Final Pulse Rate	61	72	70	76	74	59	61
Difference (Initial – Final)	6	–1	–3	8	–3	9	12

a) $\bar{D} = 4$, $S_D^2 = \dfrac{(6-4)^2+(-1-4)^2+(-3-4)^2+\cdots+(12-4)^2}{7-1} = 38.67$

$H_0: \mu_1 = \mu_2$ $H_a: \mu_1 > \mu_2$ $S_D = 6.22$

$n = 7$ small

$\mu_1 - \mu_2$

$$T = \frac{\bar{D} - 0}{S_D / \sqrt{7}} = \frac{4}{6.22/\sqrt{7}} = 1.701$$

$\alpha = 0.05$

$t_{0.05,6} = 1.943$

since $1.701 < 1.943$

we fail to reject H_0 in favor of H_a

9.4 Inferences Concerning a Difference Between Population Proportions

Population 1:

$n_1 = \#$ of observations in sample 1

$X_1 = \#$ of subjects in sample 1 that have a certain characteristic we are interested

$\hat{p}_1 = \dfrac{X_1}{n_1} =$ sample 1 proportion

$p_1 =$ population 1 proportion (unknown)

Population 2:

$n_2 = \#$ of observations in sample 2

$X_2 = \#$ of subjects in sample 2 that have the same characteristic we are interested

$\hat{p}_2 = \dfrac{X_2}{n_2} =$ sample 2 proportion

$p_2 =$ population 2 proportion (unknown)

The natural estimator for $p_1 - p_2$, the difference in population proportions, is the corresponding difference in sample proportions $\hat{p}_1 - \hat{p}_2$. Since we know $X_1 \sim Bin(n_1, p_1)$, $X_2 \sim Bin(n_2, p_2)$ with X_1 and X_2 are ______ variables. Then

$H_0: p_1 = p_2$ $\quad H_a: p_1 > p_2$

or

$p_1 < p_2$

or

$p_1 \neq p_2$

suppose $p_1 = p_2 = p$, then

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

A Large-Sample Test Procedure

Suppose we want to test $p_1 = p_2$ or equivalently $p_1 - p_2 = 0$. When H_0 is true, let p denote the common value of p_1 and p_2, i.e. $p_1 = p_2 = p$. Then we have the standardized variable

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

has approximately a standard normal distribution when H_0 is true.

Note: $Var(\hat{p}_1 - \hat{p}_2) = Var\left(\frac{Bin(n_1,p)}{n_1} - \frac{Bin(n_2,p)}{n_2}\right) = \frac{Var(Bin(n_1,p))}{n_1^2} + \frac{Var(Bin(n_2,p))}{n_2^2}$

$= \frac{n_1 p(1-p)}{n_1^2} + \frac{n_2 p(1-p)}{n_2^2} = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

estimate of the p, by pooling the data into one data set

Null Hypothesis: $H_0: p_1 = p_2$
Test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Significant level:
Alternative Hypothesis: reject if

$H_A: p_1 > p_2$ $\quad z \geq z_\alpha$

$H_A: p_1 < p_2$ $\quad z \leq -z_\alpha$

$H_A: p_1 \neq p_2$ $\quad |z| \geq z_{\alpha/2}$

Assumptions:

- Two independent samples
- $x_1 = \hat{p}_1 n_1 \geq 10$ $\quad x_2 = \hat{p}_2 n_2 \geq 10$ $\quad n_1 - x_1 = (1-\hat{p}_1) n_1 \geq 10$

  $n_2 - x_2 = (1-\hat{p}_2) n_2 \geq 10$

Provided the above assumptions satisfied, a CI for $p_1 - p_2$ with a confidence level of $100(1-\alpha)\%$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Example 106.

A survey of 400 elementary school teachers (Group 1) and 300 high school teachers (Group 2) was conducted. Of the elementary teachers, 224 said they were very satisfied with their jobs; whereas, 141 of the high school teachers were very satisfied with their work.

(a) Do the samples give enough evidence to conclude that a larger proportion of elementary school teachers are more satisfied with their jobs? Test using α = 0.05.

(b) (i) What conditions must be verified for the test in part (a) to be valid?

(ii) Check the conditions and indicate whether or not they are satisfied.

prop. of highschool teachers

$p_1 > p_2$?

proportion of elementary school teachers

$H_0: p_1 = p_2$ $\quad H_a: p_1 > p_2$

a)

$n_1 = 400$ $\quad n_2 = 300$

$\hat{p}_1 = \frac{224}{400}$ $\quad \hat{p}_2 = \frac{141}{300}$

$\hat{p} = \frac{224+141}{400+300} = \frac{365}{700}$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = 2.36$$

$z_{0.05} = 1.6449$

since $2.36 > 1.6449$
we reject H_0 in favor of H_a

b) $X_1 = 224$ $\quad X_2 = 141$

$n_1 - X_1 = 400 - 224 = 176$

$n_2 - X_2 = 300 - 141 = 159$

all ≥ 10 ✓

samples are independent ✓