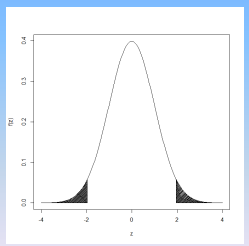
STATS 101
Introductory Statistics

Lecture 7

Confidence Interval Estimation

2-sample problems



In the previous lecture, we discussed the confidence interval estimation of one normal mean, and one binomial proportion.

In this lecture, comparison of two populations via confidence interval estimation of the difference (in two means, or two proportions) will be covered.

COMPARISON OF TWO POPULATIONS

Look at a few examples:

1. A realtor wants to compare the average selling price of homes in Henderson and Boulder City.
2. A motel chain wants to compare LCD TVs of two brands in terms of proportion of defective TVs.

These problems can be answered by calculating a confidence interval

(Problem 1) for the difference in 2 population means $\mu_1 - \mu_2$, and

(Problem 2) for the difference in 2 population proportions $p_1 - p_2$.

We will use the following formula from the previous chapter for the CI Estimate of a parameter:
estimate ± (reliability coefficient) × sd(estimate)
where the reliability coefficient comes from the probability distribution of the estimate.

Confidence Interval Estimation of Difference in 2 Population Means

The formula for CI estimate of $\mu_1 - \mu_2$ depends on how the two samples are collected:

Case 1: the two samples are independent of each other

Case 2: the two samples are PAIRED

Example of a paired sample : to determine if a weight loss program is effective, a random sample of 25 subjects is weighed before and after the weight loss program. The resulting sample is (x_j, y_j).., 25

x_j =before weight, y_j = after weight of subject j

Case 1: C I for $\mu_1 - \mu_2$ for independent samples

Sample 1:	Sample 2:
n_1 observations	n_2 observations
sample mean = $\bar{x}_1$, sd s_1	sample mean = $\bar{x}_2$, sd s_2

Estimate of $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$

$$\text{var}(\bar{x}_1) = \frac{\sigma_1^2}{n_1}, \text{var}(\bar{x}_2) = \frac{\sigma_2^2}{n_2}$$

Since the two samples are independent,
$\bar{x}_1$ and $\bar{x}_2$ are independent, and
then the variance of difference equals sum of variances:

$$\text{var}(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}, \text{ sd}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

CI ESTIMATION OF DIFFERENCE OF TWO NORMAL MEANS $\mu_1 - \mu_2$ (2 sample problem)

Case 1A: Two independent samples: two variances are equal

When $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (two variances are equal) $\text{var}(\bar{x}_1 - \bar{x}_2) = \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})$

Use the estimate of σ from the two samples:

$s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}}$, and the reliability coefficient comes from the t-distribution with degrees of freedom $n_1 + n_2 - 2$. Hence the 95% CI for $\mu_1 - \mu_2$ is

$\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2,.05} s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Confidence Interval Estimation of Difference in 2 Population Means

Example 1: CI Estimation of the difference in mean selling prices of homes in a Henderson and Boulder City, $\mu_{Henderson} - \mu_{BC}$

Suppose the realtor collects 2 representative independent random samples –

n_1 = 15 homes from Henderson, and

n_2 = 23 homes from Boulder City

Assumptions

[CI Estimation of $\mu_1 - \mu_2$, continued]

1. Assume that the random variable of interest X (selling price) is approximately normally distributed

   with mean μ_1 and sd σ_1 (Henderson sample)

   with mean μ_2 and sd σ_2 (Boulder City sample)

2. The two random samples are independent of one another.

Henderson	BC
131.515	168.195
195.058	168.193
170.623	145.11
174.3	161.917
165.894	131.031
196.372	190.279
187.97	139.783
156.613	181.898
138.788	170.188
186.699	121.059
180.709	199.435
164.636	202.593
186.104	145.565
203.83	171.162
232.011	158.708
	164.246
	155.334
	135.297
	156.256
	190.521
	170.762
	156.718
	170.015

Calculate the means and variances of the two samples (shown in lecture 8)

Henderson		Boulder City	
n1	15	n2	23
xbar1	178.0748	xbar2	163.2289
VAR1	643.5705	VAR2	450.4879

We are now ready to calculate the 95% CI for $\mu_{Henderson} - \mu_{BC}$

The formula for CI depends on whether we have

Case a: the two population variances are equal, or
Case b: the two population variances are not equal

Example 1:[CI Estimation of $\mu_1 - \mu_2$, continued]

Case

$\bar{x}_1 - \bar{x}_2 \pm t_{df} s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

where

$df = n_1 + n_2 - 2$

t_{df} = value from t-table corresponding to 95% confidence (2-sided value)

$s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}}$

Example 1:[CI Estimation of $\mu_1 - \mu_2$, continued]

Case

We first calculate the estimate of the common sd σ from the two samples:

$$s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}}$$

$$= \sqrt{\frac{14 \times 643.5705 + 22 \times 450.4879}{(15+23-2)}}$$

$$= \sqrt{\frac{18920.7208}{36}} = \sqrt{525.5756} = 22.9254$$

95% CI for $\mu_1 - \mu_2$

Case 1a: Independent samples, Equal variances

The reliability coefficient comes from the t-table with degrees of freedom = 15+23-2 = 36, and can be obtained in excel by going to any cell and typing =tinv(.05,36); Note that excel returns value corresponding to 2-sided confdence of 95%.

Estimate of $\mu_1 - \mu_2$ is: $\overline{x}_1 - \overline{x}_2 = 14.8459$

Reliability coefficient = $t_{36,.05} = 2.0281$

$$sd(\overline{x}_1 - \overline{x}_2) = s_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 22.9254 \times \sqrt{.1101} = 7.6085$$

Hence 95% confidence interval for $\mu_1 - \mu_2$ is: estimate ± reliability coeff × sd

$$= 14.8459 \pm 2.0281 \times 7.607 = 14.8459 \pm 15.4308$$

$$= (-.5849, 30.2767)$$

In words: we are 95% confident that $\mu_1 - \mu_2$ lies in the range -0.5849 to + 30.2767; since 0 lies inside this CI, 0 is a likely value of $\mu_1 - \mu_2$ or it is likely that $\mu_1 = \mu_2$.

95% CI for $\mu_1 - \mu_2$

Case 1b: Independent samples, Unequal variances

This problem is known as the Behrens- Fisher Problem. In this case, several formulas for CI of $\mu_1 - \mu_2$ exist in the statistics literature; we will illustrate Cochran's* method and Satterthwaite's** method for data of Example 1.

$$n_1 = 15, \overline{x}_1 = 178.0748, s_1^2 = 643.5705$$

$$n_2 = 23, \overline{x}_2 = 163.2289, s_2^2 = 450.4879$$

*Cochran, W. G. 1964. Approximate significance levels of the Behrens-Fisher test. Biometrics, 20: 191.

** Armitage P, Berry G. *Statistical Methods in Medical Research* (3rd edition). Blackwell 1994.

Cochran's approximate CI for data of Example 1: Unequal Variances (Example 1, continued)

$t_1 = t_{n_1-1,.05} = t_{14,.05} = 2.1448$ (t-notation for 2-sided probability)

$t_2 = t_{n_2-1,.05} = t_{23,.05} = 2.0739$

$$w_1 = \frac{s_1^2}{n_1} = \frac{643.5705}{15} = 42.9047$$

$$w_2 = \frac{s_2^2}{n_1} = \frac{450.4879}{23} = 19.5864$$

$$t'_{1-\frac{\alpha}{2}} = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2} = \frac{42.9047 \times 2.1448 + 19.5864 \times 2.0739}{42.9047 + 19.5864} = 2.1226$$

$$\sqrt{w_1 + w_2} = \sqrt{42.9047 + 19.5864} = \sqrt{62.4911} = 7.9051$$

$$CI = \overline{x}_1 - \overline{x}_2 \pm t'_{1-\frac{\alpha}{2}}\sqrt{w_1 + w_2} = (178.0748 - 163.2289) \pm 2.1226 \times 7.9051$$

$$= 14.85 \pm 16.78 = (-1.93, +31.63)$$

Satterthwaite's approximate CI for data of Example 1: Unequal Variances (Example 1, continued)

Estimate of $\mu_1 - \mu_2 = \overline{x}_1 - \overline{x}_2 = 178.0748 - 163.2289 = 14.85$

$sd(estimate) = sd(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ is estimated by $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$$\frac{s_1^2}{n_1} = \frac{643.5705}{15} = 42.9047$$

$$\frac{s_2^2}{n_2} = \frac{450.4879}{23} = 19.5864$$

$$sd(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{42.9047 + 19.5864} = \sqrt{62.4911} = 7.9051$$

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{(62.4911)^2}{\frac{42.9047^2}{14} + \frac{19.5864^2}{22}} = \frac{3905.14}{131.49 + 17.44} = \frac{3905.14}{148.93} = 26.22$$

$$CI = \overline{x}_1 - \overline{x}_2 \pm t_{df,.05}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (178.0748 - 163.2289) \pm 2.06 \times 7.9051$$

$$= 14.85 \pm 16.25 = (-1.40, +31.10)$$

Hence 95% confidence interval for $\mu_1 - \mu_2$ is

$= (-.5849, 30.2767)$, two variances are equal

and if the two variances are not equal

$= (-1.93, +31.63)$, Cochran's t,

$= (-1.40, +31.09)$, Satterthwaite's t,

Note that in either case, since 0 is inside the 95% CI, we conclude with 95% confidence that the the mean selling prices in Henderson and Boulder City are equal.

Case 2: 95% CI for $\mu_1 - \mu_2$ from PAIRED SAMPLES

Example 2: The given table shows daily sales for 20 randomly selected days for two fast food restaurants at a given intersection in Las Vegas. Calculate a 95% CI for $\mu_1 - \mu_2$.

Day	X1	X2	Day	X1	X2
1	799.13	755.54	11	870.08	898.54
2	820.62	891.40	12	733.89	798.08
3	887.98	754.38	13	837.57	823.24
4	881.90	834.13	14	777.60	829.04
5	754.32	676.80	15	849.40	704.44
6	849.14	844.90	16	828.18	793.91
7	838.64	742.50	17	848.47	803.91
8	887.50	827.49	18	857.43	772.43
9	866.69	859.20	19	678.55	594.12
10	838.33	803.33	20	881.15	855.99

Case 2: 95% CI for $\mu_1 - \mu_2$ from PAIRED SAMPLES

In Example 2, the sales X1 and X2 are CORRELATED (both values are expected to be low if the intersection does not get too much traffic, and going to be high if the intersection is a busy one). Assuming both X1 and X2 to be normally distributed with means μ_1, μ_2 and standard deviations σ_1, σ_2 respectively, then the difference D = X1-X2 is also normal with Mean of D = $\mu_1 - \mu_2$, and sd of D = σ_D is unknown. Hence the paired t-test is simply a 1-sample t-test on the difference D.

19

Day	X1	X2	D	Day	X1	X2	D
1	799.13	755.54	43.59	11	870.08	898.54	-28.45
2	820.62	891.40	-70.78	12	733.89	798.08	-64.18
3	887.98	754.38	133.60	13	837.57	823.24	14.33
4	881.90	834.13	47.77	14	777.60	829.04	-51.45
5	754.32	676.80	77.52	15	849.40	704.44	144.96
6	849.14	844.90	4.24	16	828.18	793.91	34.27
7	838.64	742.50	96.14	17	848.47	803.91	44.56
8	887.50	827.49	60.01	18	857.43	772.43	85.00
9	866.69	859.20	7.49	19	678.55	594.12	84.43
10	838.33	803.33	35.00	20	881.15	855.99	25.16

20

D-bar	36.16
sd(D)	59.82
reliability coefficient = tinv(.05,19)	2.093
rel coeff X sd(D)/sqrt(n)	27.99
L = Dbar - rel coeff X sd(D)/sqrt(n)	8.17
U = Dbar + rel coeff X sd(D)/sqrt(n)	64.15

95% CI for $\mu_1 - \mu_2$ is therefore (8.17, 64.15). Since the entire interval falls to the right of 0, we conclude that $\mu_1 > \mu_2$.

21

CI Estimation of One Binomial Proportion

- Recall that the number of successes in a series of experiments resulting in only 2 outcomes (SUCCESS or FAILUES) has a binomial distribution with n trials and p = P(Success) is the parameter to be estimated.
- Example: A restaurant owner has conducted an exit survey and found out that 88 out of 125 customers said YES to the question: WERE YOU HAPPY WITH FOOD AND SERVICE? The owner wants to estimate p,

22

CI Estimation of One Binomial Proportion (contd.)

the proportion of customers who are HAPPY WITH FOOD AND SERVICE. This proportion p is estimated by

$$\hat{p} = \frac{X}{n} = \frac{\#\text{ of successes}}{\#\text{ of trials}}$$

For the above example

$$\hat{p} = = \frac{88}{125}$$

The CI of p is calculated based on the Central Limit Theorem – for large n, the estimate $\hat{p}$ is approximately normal with mean p and variance p(1-p)/n (if np ≥5, n(1-p) ≥ 5).

23

$\hat{p}$ is normally distributed with mean p, and variance $\frac{p(1-p)}{n}$:

$$\hat{p} = \frac{X}{n} \sim N\left(p, sd = \sqrt{\frac{p(1-p)}{n}}\right)$$

Since the distribution of the estimate of p is normal, the reliability coefficient in the following formula comes from the Z-Table.

95% CI for p is

$$\hat{p} \pm \text{reliability coefficient} \times \text{sd}(\hat{p})$$

$$\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

One proportion

100(1-α)% CI for p is:

$$\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example 1:
In a sample of 125 customers, 88 stated they were HAPPY WITH FOOD AND SERVICE.

$$\hat{p} = \frac{x}{n} = \frac{88}{125} = 0.7040$$

$$sd(\hat{p}) = \sqrt{\frac{0.704 \times (1-0.704)}{125}} = \sqrt{0.00166707} = 0.0408$$

$$z_{.975} = 1.96$$

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.7040 \pm 1.96 \times 0.0408 = 0.7040 \pm 0.0800 = (0.624, 0.784)$$

Difference in 2 proportions

100(1-α)% CI for $p_1 - p_2$

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

Example 2: Of the 215 black subjects in a health study, 58 had diabetes mellitus; of the 1140 white subjects, 217 had diabetes mellitus. Construct A 90% CI for the difference between the two population proportions.

$$n_1\hat{p}_1 \geq 5, n_1(1-\hat{p}_1) \geq 5$$
$$\Rightarrow normal \quad approximation \quad valid$$
$$n_2\hat{p}_2 \geq 5, n_2(1-\hat{p}_2) \geq 5$$

$$n_1 = 215, x_1 = 58, \hat{p}_1 = \frac{58}{215} = 0.269767$$

$$n_2 = 1140, x_2 = 217, \hat{p}_2 = \frac{217}{1140} = 0.190351$$

$$sd(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{0.269767 \times (1-0.269767)}{215} + \frac{0.190351 \times (1-0.190351)}{1140}} = \sqrt{0.001051} = 0.0324$$

$$z_{0.95} = 1.645$$

$$(0.2698 - 0.1904) \pm 1.645 \times 0.0324 = 0.0794 \pm 0.0533 = (0.0261, 0.1327)$$

Since CI > 0, we can Conclude $p_1 > p_2$