Two Parameter Estimations

Presented By:

Dr. Vinay Kulkarni

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Recap: Single Parameter Estimation

- The Mean, when variance is known
  - Test statistic $Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
- The Mean, when variance is unknown
  - Large sample size $Z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$ (normal dist)
  - Small sample size $T = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$ (t-distribution)
- The Variance
  - Test statistic $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ (chi-squared)
- The Proportion
  - Test statistic $Z = \frac{\hat{p} - p_0}{\sqrt{[p_0(1-p_0)/n]}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$

$\hat{p} = x/n$

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Confidence Interval: Two Parameters

- Two means
- Two proportions
- Two Variances / Standard Deviations

Example

- Change in the manufacturing process for certain parts is being considered
- Samples are taken
  - Using the old and the new procedures
- Goal:
  - To identify the difference induced by the new procedure
- Sampling results:
  - Old procedure: 75 out of 1500 were found defective
  - New procedure: 80 out of 2000 were found defective
- Find:
  - 90% Confidence Interval on the true difference in the fraction of defectives between the old and new procedures

Confidence Interval: Difference between proportions

If

- x1, x2 are number of successes
- n1, n2 are number of trials
- P1 = x1/n1 and P2 = x2/n2 are the proportions
- p1 and p2 are the “true” value of probabilities

Then

- Confidence interval for (p1-p2) is given by

$$\text{Lower Limit: } (\hat{P}_1 - \hat{P}_2) - Z_{\alpha/2}\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$$

$$\text{Upper Limit: } (\hat{P}_1 - \hat{P}_2) + Z_{\alpha/2}\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}\,.$$

Example: Solution

- p1 = proportion of defectives in old procedure
- p2 = proportion of defectives in new procedures
- Find :
  - 90% confidence for (p1 – p2)
- x1 = 75; n1 = 1500; $\hat{p}_1$ = 75/1500 = 0.05
- x2 = 80; n2 = 2000; $\hat{p}_2$ = 80/2000 = 0.04
- For 90% CI, $z_{0.05}$ = 1.645

- p1 = proportion of defectives in old procedure
- p2 = proportion of defectives in new procedures
- Find :
  - 90% confidence for (p1 – p2)
- x1 = 75; n1 = 1500; $\hat{p}_1$ = 75/1500 = 0.05
- x2 = 80; n2 = 2000; $\hat{p}_2$ = 80/2000 = 0.04
- For 90% CI, $z_{0.05}$ = 1.645

- Limits = (0.01) ± 1.645 * $\sqrt{\frac{0.05*0.95}{1500} + \frac{0.04*0.96}{2000}}$
- Lower Limit = 0.01 – 0.0117323; Upper Limit = 0.01 + 0.0117323
- Therefore: -0.0017 < p1 – p2 < 0.0217

- Limits = (0.01) ± 1.645 * $\sqrt{\frac{0.05*0.95}{1500} + \frac{0.04*0.96}{2000}}$
- Lower Limit = 0.01 – 0.0117323; Upper Limit = 0.01 + 0.0117323
- Therefore: -0.0017 < p1 – p2 < 0.0217

Confidence Interval: Difference between Means

- Example:
  - There are two varieties of crops
  - The mean yield of first variety is measured
  - The mean yield of second variety is measured
  - We are interested in finding the confidence interval of the difference in their yields

Confidence Interval: Difference between means

- Following cases may arise
  - Variance of population is known
    - Use it
  - Variance of population is unknown, but sample size is greater than 30
    - Use sample variance of the sample
  - Variance of population is unknown, and sample size is less than 30
    - Use formula for calculating degrees of freedom
    - Use t-distribution

Confidence Intervals: Difference between means

- Population variances known

$$(\bar{X}-\bar{Y}) - Z_{\alpha/2}\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2} < \mu_1-\mu_2 < (\bar{X}-\bar{Y}) + Z_{\alpha/2}\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}$$

- Population variances not known; sample > 30

$$(\bar{X}-\bar{Y}) - Z\sqrt{S_1^2/n_1+S_2^2/n_2} < \mu_1-\mu_2 < (\bar{X}-\bar{Y}) + Z\sqrt{S_1^2/n_1+S_2^2/n_2}$$

- Population variances not known; sample < 30

$$(\bar{X}-\bar{Y}) - t\sqrt{S_1^2/n_1+S_2^2/n_2} < \mu_1-\mu_2 < (\bar{X}-\bar{Y}) + t\sqrt{S_1^2/n_1+S_2^2/n_2}$$

where the d.o.f is given by

$$\frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{(n_1-1)}+\frac{\left(\frac{S_2^2}{n_2}\right)^2}{(n_2-1)}}$$

Confidence Intervals: Difference between Means

- When it is known that the two population variances are the same, the samples' variance can be pooled. In this case …

$$S^2_{pooled} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}.$$

In this case we have a new random variable given by

$$T = \frac{(X - Y) - (\mu_1 - \mu_2)}{S_{pooled}\sqrt{1/n_1 + 1/n_2}},$$

This test Statistic follows t-distribution with d.o.f ($n_1 + n_2 - 2$)

CI of Difference of Means of "paired samples"

- All earlier cases assumed "independent samples"
- If the samples are dependent, or related
  - Eg. Same individual measured under different conditions
- Then … two different samples reduced to
  - One sample of paired data
- In this case:

$$d_i = X_i - Y_i,\ i = 1, 2, \ldots, n.$$

$$\bar{d} = (1/n)\sum_{i=1}^{n} d_i \text{, and}$$

$$s_d^2 = \{1/(n-1)\}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

$$T = \frac{\bar{d} - (\mu_1 - \mu_2)}{S_d / \sqrt{n}} \qquad \textbf{d.o.f = n - 1}$$

$$\bar{d} - t_{\alpha/2} S_d / \sqrt{n} < \mu_d < \bar{d} + t_{\alpha/2} S_d / \sqrt{n}$$

Example: Difference between means

Data was collected to compare the wear of two different materials. The summary came up to be

Sample	Mean	Size	Sample Standard deviation
I	85	32	5
II	81	30	4

Calculate the 95% C. I on the difference between the two means.

Example: Differences of means, large sample

- $\bar{x}$ = 85; n_1 = 32; s_1 = 5 … population sd unknown
- $\bar{y}$ = 81; n_2 = 30; s_2 = 4 … population sd unknown
- Required: 95% CI on the true difference between the means
- Since population sd is not known, and sample size is > 30

$$(\bar{X}-\bar{Y}) - Z\sqrt{S_1^2/n_1 + S_2^2/n_2} < \mu_1 - \mu_2 < (\bar{X}-\bar{Y}) + Z\sqrt{S_1^2/n_1 + S_2^2/n_2}$$

- $z_{0.025}$ = -1.96; $z_{0.975}$ = 1.96
- $\sqrt{\frac{25}{32} + \frac{16}{30}}$ = 1.146
- (CI for difference of means) = (85-81) ± 1.96 * 1.146 = 4 ± 2.246
- (CI for difference of means) = [1.754, 6.246]

Difference between means: Pooled Variances

- Conditions under which "Pooled Variance" applicable
  - Population variances are unknown
  - Sample sizes are 'small'
  - Populations from which samples are drawn
    - Normal distribution
  - The two population variances are equal
- In such cases:
  - Pooled variance : $$S^2_{pooled} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}$$
  - Statistic : $$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_{pooled}\sqrt{1/n_1 + 1/n_2}}$$

    This has t-distribution with d.o.f = $(n_1 + n_2 - 2)$
  - C.I : $$(\bar{X} - \bar{Y}) - t_{\alpha/2} . S_{pooled}\sqrt{1/n_1 + 1/n_2} < \mu_1 - \mu_2 < (\bar{X} - \bar{Y}) + t_{\alpha/2} . S_{pooled}\sqrt{1/n_1 + 1/n_2}$$

Example: Difference between means

Data was collected to compare the wear of two different materials. The summary came up to be

Sample	Mean	Size	Sample Standard deviation
I	85	12	5
II	81	10	4

Calculate the 95% C. I on the difference between the two means, by

a) Pooling for the common variance of the two populations,

b) Not pooling.

Solution: Pooled sd

- $\bar{x}$ = 85; n_1 = 12; s_1 = 5 … population sd unknown
- $\bar{y}$ = 81; n_2 = 10; s_2 = 4 … population sd unknown
- Population sd of both are assumed to be same … hence pooling
- Required: 95% CI on the true difference between the means
- Since population sd is not known, and <u>sample size is < 30</u>

$$(\bar{X} - \bar{Y}) - t_{\alpha/2} . S_{pooled} \sqrt{1/n_1 + 1/n_2} < \mu_1 - \mu_2 < (\bar{X} - \bar{Y}) + t_{\alpha/2} . S_{pooled} \sqrt{1/n_1 + 1/n_2}$$

- $S^2_{pooled} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)} = \frac{(12-1)*25+(10-1)*16}{12+10-2}$ = 20.95; $S_{pooled} = \sqrt{20.95} = 4.577$
- $t_{0.025,20}$ = -2.086; $t_{0.975,20}$ = 2.086
- $\sqrt{\frac{1}{n1} + \frac{1}{n2}} = \sqrt{\frac{1}{12} + \frac{1}{10}} = 0.428$
- (CI for difference of means) = (85-81) ± 2.086 * $\sqrt{20.95}$ * 0.428 = 4 ± 4.086
- (CI for difference of means) = [0.086, 8.086]

Solution: Non-pooled sd

- $\bar{x}$ = 85; n_1 = 12; s_1 = 5 … population sd unknown
- $\bar{y}$ = 81; n_2 = 10; s_2 = 4 … population sd unknown
- No pooling of standard deviation
- Required: 95% CI on the true difference between the means
- Since population sd is not known, sample size is < 30, no pooling

- d.o.f = $\dfrac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{(n_1-1)}+\frac{\left(\frac{S_2^2}{n_2}\right)^2}{(n_2-1)}} = \dfrac{\left(\frac{25}{12}+\frac{16}{10}\right)^2}{\frac{\left(\frac{25}{12}\right)^2}{(12-1)}+\frac{\left(\frac{16}{10}\right)^2}{(10-1)}} = \dfrac{13.567}{0.3945+0.284} = 19.99 \sim 20$
- $(\bar{X}-\bar{Y})-t\sqrt{S_1^2/n_1+S_2^2/n_2} < \mu_1-\mu_2 < (\bar{X}-\bar{Y})+t\sqrt{S_1^2/n_1+S_2^2/n_2}$.
- $t_{0.025,20}$ = -2.086; $t_{0.975,20}$ = 2.086
- (CI for difference of means) = (85-81) ± 2.086 * $\sqrt{\frac{25}{12}+\frac{16}{10}}$ = 4 ± 4.003
- (CI for difference of means) = [-0.00345, 8.00345]

Confidence Interval: Ratio of Variances

- In the case of variances : we are interested in the ratio, and not the difference
- Test statistic = $\left(\frac{s1^2}{\sigma 1^2}\right)/\left(\frac{s2^2}{\sigma 2^2}\right) = F = \frac{U/r_1}{V/r_2}$
- Where $U = \frac{(n_1-1)S_1^2}{\sigma_1^2}$, and $V = \frac{(n_2-1)S_2^2}{\sigma_2^2}$ with $r_1 = (n_1-1)$ and $r_2 = (n_2-1)$ degrees of freedom
- The statistic F follows F-distribution
  - It is characterized by α, r_1 and r_2
  - It can be seen that F is a ratio of two Chi-square terms
- And the confidence interval of the ratio of variances is given by

$$\frac{S_2^2}{S_1^2}\frac{1}{F(1-\alpha/2, r_2, r_1)} \le \frac{\sigma_2^2}{\sigma_1^2} \le \frac{S_2^2}{S_1^2}F(1-\alpha/2, r_1, r_2)$$

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Example: Ratio of variances

A standard and a new method of teaching statistics are being compared with respect to their variability as measured by the final examination scores produced by the two methods. One class is taught by each method, and both classes take the same final examination. The observed sample variances are $S_1^2 = 100$, and $S_2^2 = 144$, where the first data coming from the standard method and the second is from the new method. If the classes contained 121 and 61 students respectively, find a 95% C.I. for $\frac{\sigma_2^2}{\sigma_1^2}$.

- $s_1{}^2 = 100$; $s_2{}^2 = 144$; n1 = 121; r1 = 120; n2 = 61; r2 = 60
- $\frac{sigma2^2}{sigma1^2} = F_{1,2} * \frac{s_2{}^2}{s_1{}^2}$ (statistic to be evaluated at the two CI boundaries)
- $F_{0.025,120,60}$ = qf(0.025, 120, 60) = 0.6536 ; $F_{0.975,120,60}$ = qf(0.975,120,60) = 1.581
- $\frac{144}{100} * 0.6536 \leq Ratio \leq \frac{144}{100} * 1.581$
- $0.94 \leq Ratio \leq 2.276$