STA13: Elementary Statistics

Lecture 19
Book sections 7.1-7.3

Dmitriy Izyumin

March 05 2018

Independent Samples

Sometimes we have two independent random samples from different populations, and want to answer the following.

- Is there a difference between the means of the populations?
- In other words, are observations from the two populations different on average?
- Are values from population 1 on average larger than values from population 2?

Independent Samples

Examples:

- Salaries of graduates of two different majors.
- Blood pressure of patients subjected to treatment 1, and patients subjected to treatment 2.

Notation

- Population 1 has mean μ_1 and s.d σ_1.
- Sample 1 is taken from population 1, and has size n_1, sample mean $\bar{x}_1$, and sample s.d s_1.
- Population 2 has mean μ_2 and s.d σ_2.
- Sample 2 is taken from population 2, and has size n_2, sample mean $\bar{x}_2$, and sample s.d s_2.

Parameter

$$\mu_1 - \mu_2$$

- μ_1 is the mean of the first population.
- μ_2 is the mean of the second population.
- We want to make inferences about $\mu_1 - \mu_2$, the difference in means between the two populations.

Statistic

$$\bar{x}_1 - \bar{x}_2$$

- $\bar{x}_1$ is the sample mean of the first sample.
- $\bar{x}_2$ is the sample mean of the second sample.
- $\bar{x}_1 - \bar{x}_2$ is a statistic and has a sampling distribution.
- We use the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to make inferences about $\mu_1 = \mu_2$.

Sampling Distribution

Properties of the Sampling Distribution of $(\bar{x}_1 - \bar{x}_2)$

1. The mean of the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ is $(\mu_1 - \mu_2)$.
2. If the two samples are independent, the standard deviation of the sampling distribution is

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

   where σ_1^2 and σ_2^2 are the variances of the two populations being sampled and n_1 and n_2 are the respective sample sizes. We also refer to $\sigma_{(\bar{x}_1 - \bar{x}_2)}$ as the **standard error of the statistic** $(\bar{x}_1 - \bar{x}_2)$.
3. By the Central Limit Theorem, the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ is approximately normal *for large samples*.

Large Sample - Conditions

Conditions Required for Valid Large-Sample Inferences about $(\mu_1 - \mu_2)$

1. The two samples are randomly selected in an independent manner from the two target populations.
2. The sample sizes, n_1 and n_2, are both large (i.e., $n_1 \geq 30$ and $n_2 \geq 30$). (By the Central Limit Theorem, this condition guarantees that the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ will be approximately normal, regardless of the shapes of the underlying probability distributions of the populations. Also, s_1^2 and s_2^2 will provide good approximations to σ_1^2 and σ_2^2 when both samples are large.)

Large Sample - CI

Large, Independent Samples Confidence Interval for $(\mu_1 - \mu_2)$: Normal (z) Statistic

$$\sigma_1^2 \text{ and } \sigma_2^2 \text{ known: } (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\,\sigma_{(\bar{x}_1-\bar{x}_2)} = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\sigma_1^2 \text{ and } \sigma_2^2 \text{ unknown: } (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\,\sigma_{(\bar{x}_1-\bar{x}_2)} \approx (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Large Sample - HT

Large, Independent Samples Test of Hypothesis for $(\mu_1 - \mu_2)$: Normal (z) Statistic

One-Tailed Test

H_0: $(\mu_1 - \mu_2) = D_0$

H_a: $(\mu_1 - \mu_2) < D_0$

[or H_a: $(\mu_1 - \mu_2) > D_0$]

Two-Tailed Test

H_0: $(\mu_1 - \mu_2) = D_0$

H_a: $(\mu_1 - \mu_2) \neq D_0$

where D_0 = Hypothesized difference between the means (this difference is often hypothesized to be equal to 0)

Test statistic:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sigma_{(\bar{x}_1 - \bar{x}_2)}} \quad \text{where} \quad \sigma_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \text{ if both } \sigma_1^2 \text{ and } \sigma_2^2 \text{ are known}$$

$$\approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \text{ if } \sigma_1^2 \text{ and } \sigma_2^2 \text{ are unknown}$$

Rejection region: $z < -z_\alpha$

[or $z > z_\alpha$ when H_a: $(\mu_1 - \mu_2) > D_0$]

Rejection region: $|z| > z_{\alpha/2}$

Small Sample - Conditions

Conditions Required for Valid Small-Sample Inferences about $(\mu_1 - \mu_2)$

1. The two samples are randomly selected in an independent manner from the two target populations.
2. Both sampled populations have distributions that are approximately normal.
3. The population variances are equal (i.e., $\sigma_1^2 = \sigma_2^2$).

Equal Variances

- If $\frac{\text{larger sample s.d.}}{\text{smaller sample s.d.}} < 2$, then we can assume the population variances are approximately equal.
- Otherwise we can't assume that population variances are equal.
- There is another version of the test for situations with unequal variances. It will not be covered in this class.

Small Sample - CI

Small, Independent Samples Confidence Interval for $(\mu_1 - \mu_2)$: Student's t-Statistic

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

and $t_{\alpha/2}$ is based on $(n_1 + n_2 - 2)$ degrees of freedom.

[*Note:* $s_p^2 = \dfrac{s_1^2 + s_2^2}{2}$ when $n_1 = n_2$]

- The degrees of freedom are not $n - 1$ anymore.
- s_p^2 is called the pooled sample estimate of the variance

Small Sample - HT

Small, Independent Samples Test of Hypothesis for $(\mu_1 - \mu_2)$: Student's t-Statistic

One-Tailed Test

H_0: $(\mu_1 - \mu_2) = D_0$
H_a: $(\mu_1 - \mu_2) < D_0$
[or H_a: $(\mu_1 - \mu_2) > D_0$]

Two-Tailed Test

H_0: $(\mu_1 - \mu_2) = D_0$
H_a: $(\mu_1 - \mu_2) \neq D_0$

$$\textit{Test statistic: } t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Rejection region: $t < -t_\alpha$
or $t > t_\alpha$ when
H_a: $(\mu_1 - \mu_2) > D_0$]

Rejection region: $|t| > t_{\alpha/2}$

where t_α and $t_{\alpha/2}$ are based on $(n_1 + n_2 - 2)$ degrees of freedom.

Paired Differences

Sometimes we have two samples of paired observations.

- Blood pressure measurements before and after treatment.
- Student scores on the midterm and the final.

Here, the two samples are NOT independent, as they are obtained based on the same subjects.

Procedure

- Start with two samples (size n each) of paired observations (ex. Before and After).
- Take differences (ex. After - Before).
- Now have one sample (size n) of differences.
- Compute
  - $\bar{x}_d$, the sample mean of the differences,
  - s_d, the sample s.d. of the differences.
- Proceed as before in the one-sample setting.

Paired - CI

Paired Difference Confidence Interval for $\mu_d = \mu_1 - \mu_2$

Large Sample, Normal (z) Statistic

$$\bar{x}_d \pm z_{\alpha/2}\frac{\sigma_d}{\sqrt{n_d}} \approx \bar{x}_d \pm z_{\alpha/2}\frac{s_d}{\sqrt{n_d}}$$

Small Sample, Student's t-Statistic

$$\bar{x}_d \pm t_{\alpha/2}\frac{s_d}{\sqrt{n_d}}$$

where $t_{\alpha/2}$ is based on $(n_d - 1)$ degrees of freedom

Paired - HT

Paired Difference Test of Hypothesis for $\mu_d = \mu_1 - \mu_2$

One-Tailed Test	**Two-Tailed Test**
$H_0: \mu_d = D_0$	$H_0: \mu_d = D_0$
$H_a: \mu_d < D_0$	$H_a: \mu_d \neq D_0$
[or $H_a: \mu_d > D_0$]	

Large Sample, Normal (*z*) Statistic

$$\text{Test statistic: } z = \frac{\bar{x}_d - D_0}{\sigma_d/\sqrt{n_d}} \approx \frac{\bar{x}_d - D_0}{s_d/\sqrt{n_d}}$$

Rejection region: $z < -z_\alpha$ [or $z > z_\alpha$ when $H_a: \mu_d > D_0$]	*Rejection region:* $\lvert z \rvert > z_{\alpha/2}$

Small Sample, Student's *t*-Statistic

$$\text{Test statistic: } t = \frac{\bar{x}_d - D_0}{s_d/\sqrt{n_d}}$$

Rejection region: $t < -t_\alpha$ [or $t > t_\alpha$ when $H_a: \mu_d > D_0$]	*Rejection region:* $\lvert t \rvert > t_{\alpha/2}$

where t_α and $t_{\alpha/2}$ are based on $(n_d - 1)$ degrees of freedom

Paired - Conditions

Conditions Required for Valid Large-Sample Inferences about μ_d

1. A random sample of differences is selected from the target population of differences.
2. The sample size n_d is large (i.e., $n_d \geq 30$). (By the Central Limit Theorem, this condition guarantees that the test statistic will be approximately normal, regardless of the shape of the underlying probability distribution of the population.)

Conditions Required for Valid Small-Sample Inferences about μ_d

1. A random sample of differences is selected from the target population of differences.
2. The population of differences has a distribution that is approximately normal.