Lecture 9:
Inference for Two Means
STAT 630, Fall 2021

Difference of Two Means from Independent Samples

- In this lecture we discuss how to construct confidence intervals and perform hypothesis tests for the difference between two populations means $\mu_1 - \mu_2$, where the data come from two independent samples.
- Just as with a single sample, we need to check whether certain conditions are satisfied for the confidence interval or hypothesis test to be valid.
- An important question we address is whether the difference between the two population means is significantly different than 0.

Difference of Two Means from Independent Samples

$1 - \alpha$ confidence interval for $\mu_1 - \mu_2$:

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2;df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- When both sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$) we can use either a z or t critical value.
- The degrees of freedom can be computed as $df = \min(n_1 - 1, n_2 - 1)$
- The official formula for the degrees of freedom computed using software (`t.test()` function in R) is more complex and given by the Welch-Satterthwaite approximation.[1]

[1] https://en.wikipedia.org/wiki/Welch%27s_t-test

Difference of Two Means from Independent Samples

$H_0: \mu_1 - \mu_2 = 0 \Rightarrow \mu_1 = \mu_2$

$H_A: \mu_1 - \mu_2 \neq 0 \Rightarrow \mu_1 \neq \mu_2$

Hypothesis Test:
$H_0 : \mu_1 - \mu_2 = \delta$
$H_A : \mu_1 - \mu_2 > \delta$, or $\mu_1 - \mu_2 < \delta$, or $\mu_1 - \mu_2 \neq \delta$

Test Statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Most commonly, $\delta = 0$, which is a hypothesis test for whether there is a difference between the two population means. For example, when $\delta = 0$, a two-sided test is $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$
- When both sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$) we can use either a z or t test statistic.
- The degrees of freedom are the same as the confidence interval.

$df = \min(n_1 - 1, n_2 - 1)$

Conditions

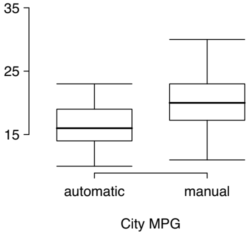
Conditions for a confidence interval or hypothesis test for the difference between to population means:

- The data in each group comes from a random sample, or randomized experiment. Additionally, the two groups are independent of each other (the cases in the first group are not related to the cases in the second group).
- The sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$), and there are no extreme outliers. This implies that the sampling distribution for $\bar{X}_1 - \bar{X}_2$ is approximately normal according to the central limit theorem.
- Otherwise, when the sample sizes are small ($n_1 < 30$ or $n_2 < 30$), the data in each group should follow an approximate normal distribution. Use graphical methods to check this (e.g., side-by-side box plots, histograms).

Example

Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel efficiency (city miles per gallon) from random samples of cars with manual and automatic transmissions. Use a hypothesis test to determine whether there is a statistically significant difference between the two means.

	City MPG	
	Automatic	Manual
Mean	16.12	19.85
SD	3.58	4.51
n	26	26



Example

(a) Write the null and alternative hypothesis for a two sided test.

$H_0: \mu_M = \mu_A \qquad H_A: \mu_M \neq \mu_A$

(b) Check the conditions for the test.

– Data come from 2 independent random samples ✓

– Data in each group are approx normal ✓
(should check since $n_M = n_A = 26 < 30$)

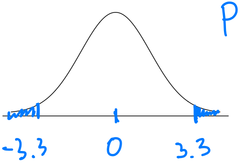
(c) Calculate the test statistic.

$$t = \frac{\bar{X}_M - \bar{X}_A}{\sqrt{\frac{S_M^2}{n_M} + \frac{S_A^2}{n_A}}} = \frac{19.85 - 16.12}{\sqrt{\frac{4.51^2}{26} + \frac{3.58^2}{26}}} = 3.3$$

Example

$df = \min(26-1, 26-1) = 25$

(d) Calculate the p-value and make a decision using $\alpha = 0.01$ significance level.



$p\text{-value} = 2 * pt(-3.3, df = 25)$

$= 0.003$

Since p-value < 0.01, we reject H_0

(e) What is the conclusion of the test in the context of the data?

The data provide strong evidence of a difference between the average city mpg of cars with automatic and manual transmission. So the difference between the two means is statistically significant.

Example

Calculate and interpret a 99% confidence interval for the difference between the two means.

99% CI for $\mu_M - \mu_A$

$$\bar{X}_M - \bar{X}_A \pm t_{\alpha/2; df} \sqrt{\frac{s_M^2}{n_M} + \frac{s_A^2}{n_A}}$$

critical value

qt(0.995, df=25)

$= 2.79$

$$\Rightarrow 3.73 \pm 2.79(1.13) \Rightarrow \boxed{(0.58,\ 6.88)}$$

We are 99% confident that the mean city mileage for manual cars is between 0.58 and 6.88 mpg higher than automatic cars.

Note: since $0 \notin CI$, we reject $H_0: \mu_M = \mu_A$

Inference for Paired Data

- Two sets of observations are **paired** if each observation in one data set has a correspondence or connection with exactly one observation in the other data set.
- To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations.
- Paired data commonly arise in experiments that compare the difference in a variable before and after some treatment was applied to the same subjects (e.g., blood pressure before and after taking a drug).

Inference for Paired Data

Case	x_i	y_i	d_i
1	x_1	y_1	$d_1 = x_1 - y_1$
2	x_2	y_2	$d_2 = x_2 - y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	$d_n = x_n - y_n$

Compute the sample mean and standard deviation of the differences:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

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

Inference for Paired Data

$1 - \alpha$ confidence interval for μ_d:

$$\bar{d} \pm t_{\alpha/2;n-1} \frac{s_d}{\sqrt{n}}$$

Hypothesis test (paired t-test):
$H_0 : \mu_d = d_0$
$H_A : \mu_d > d_0$, or $\mu_d < d_0$, or $\mu_d \neq d_0$

Test Statistic:

$$t = \frac{\bar{d} - d_0}{\frac{s_d}{\sqrt{n}}};\ df = n - 1$$

Setting $d_0 = 0$ tests whether the mean difference is 0 (no difference).

Inference for Paired Data

Checking the Conditions:

- The paired t-test is just a one-sample t-test applied to the differences.
- Check that the sample size is large ($n \geq 30$), or for small samples, that the differences are approximately normal.
- The subjects, or pairs of observations, should be randomly sampled.

Example

A random sample of $n = 170$ British couples were interviewed. The ages of each husband and wife were recorded, and the first several entries of the data are shown below. The sample mean difference (husband's age – wife's age) is $\bar{d} = 2.24$ with standard deviation $s_d = 4.08$. Use a hypothesis test to determine whether there is a significant difference in age between British husbands and wives.

Case	Husband's Age	Wife's Age	Difference
1	49	43	6
2	25	28	-3
3	40	30	10
4	52	57	-5
5	58	52	6
⋮	⋮	⋮	⋮
169	40	39	1
170	59	56	3

Example

(a) Write the null and alternative hypothesis for a two sided test.

$H_0: \mu_d = 0$ $\quad$ $H_A: \mu_d \neq 0$

(b) Check the conditions for the test.

- Independence between pairs since couples were randomly sampled ✓
- $n = 170 > 30$, large # of pairs (couples) ✓

(c) Calculate the test statistic. $n = 170$, $\bar{d} = 2.24$, $s_d = 4.08$

$$z = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{2.24}{4.08/\sqrt{170}} = 7.16$$

(can use z-test since $n = 170$ is large)

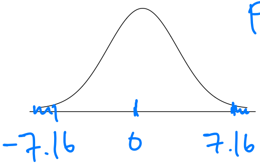
Example

(d) Calculate the p-value and make a decision using $\alpha = 0.05$ significance level.

p-value = 2 * pnorm(-7.16)

≈ 0

$\Rightarrow$ reject H_0



(e) What is the conclusion of the test in the context of the data?

The population mean difference in age, μ_d, is significantly different than 0.

Example

$qnorm(0.975) = 1.96$

Construct and interpret a 95% confidence interval for the population mean difference in age between British husbands and wives.

95% CI for μ_d:

$$\bar{d} \pm z_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}} \Rightarrow 2.24 \pm 1.96 \cdot \frac{4.08}{\sqrt{170}}$$

$$\Rightarrow \boxed{(1.63, 2.85)}$$

We are 95% confident that the husbands are between 1.63 and 2.85 years older than their wives, on average.