DSA 8010 - Inference on two means

Two sample t procedures

Paired data

Paired data

Paired data

Open Intro Statistics, 4th edition, Diez et al: Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Examples of paired data:

- Measure employees' scores on a competence assessment before and after implementing a new training program.
- For a random sample of 100 congressional districts, measure the voter turnout in 2016 and again in 2020.
- Select 45 cities at random. In each city, record the nightly price of Hyatt hotel rooms and Hilton hotel rooms.

Paired data in rectangular format

- Often, paired data should be organized so that each row corresponds to one pair and each column corresponds to one of the numeric measurements on that pair.
- The two numeric measurements can be thought of as two different variables, measured on the same pair.

	A	B	C
1	School	2016-2017 Graduation rate	2017-2018 Graduation Rate
2	A	70.9	70.9
3	B	90.3	86.3
4	C	93.4	82.4
5	D	69.9	74.3
6	E	96.2	97.1
7	F	76	80.7
8	G	80.9	83.6
9	H	85.4	85.2
10	I	57.2	58.9
11	J	53.6	53.9
12	K	80.9	83
13	L	95.9	95.1
14			

Describing paired data

When performing a descriptive analysis of paired data, keep in mind the relationship between the measurements.

- Univariate summaries of each numeric vector.

  This can be helpful, but gives limited information.

- Make scatterplots or calculate correlation among the paired measurements.
- Summarize the difference between the two measurements.

  Do the differences tend to be small or large in magnitude? Do they tend to be positive or negative?

Inference on paired data

Data. n paired observations.

Notation. y_{1i}, y_{2i} are the two paired measurements from the ith pair.

Let $d_i = y_{1i} - y_{2i}$, $i = 1, \ldots, n$. The d_i are observed differences between the two measurements.

$\bar{d}$ = sample mean of the d_i.
s_d = sample standard deviation of the d_i.

Statistical model. $d_1, \ldots, d_n$ are i.i.d. and approximately $N(\mu_d, \sigma_d^2)$.

The inferential question is on average, does measurement 1 differ from measurement 2?

Paired t confidence interval

A $(1 - \quad) \times 100\%$ CI for μ_d is

$$\bar{d} \pm t^*_{n-1,\alpha/2} s_d / \sqrt{n},$$

where $t^*_{n-1,\alpha/2}$ is the $1 - \quad /2$th percentile of t distribution with $n - 1$ degrees of freedom.

If the confidence interval does not contain zero, there is a significant difference between the means of the two measurements.

Paired t test

Hypotheses. $H_0 : \mu_d = D_0$

$H_A : \mu_d \neq D_0 \quad (>, <)$

Test statistic.

$$t_0 = \frac{\bar{d} - D_0}{s_d/\sqrt{n}}$$

p-value. For a two-sided alternative: $2 * P(T > |t_0|)$, where T has a t distribution with degrees of freedom$=n-1$.

`2*pt(abs(t0),df=n-1)`.

Decision. Reject H_0 if the p-value is $<$.

Example: temperatures

From *Open Intro Statistics, 4th edition, Diez et al*: Let's consider a limited set of climate data, examining temperature differences in 1948 vs 2018. We sampled 197 locations from the National Oceanic and Atmospheric Administration's (NOAA) historical data, where the data was available for both years of interest. We want to know: were there more days with temperatures exceeding 90°F in 2018 or in 1948? The difference in number of days exceeding 90°F (number of days in 2018 - number of days in 1948) was calculated for each of the 197 locations.

Example: temperatures

From *Open Intro Statistics, 4th edition, Diez et al*: Let's consider a limited set of climate data, examining temperature differences in 1948 vs 2018. We sampled 197 locations from the National Oceanic and Atmospheric Administration's (NOAA) historical data, where the data was available for both years of interest. We want to know: were there more days with temperatures exceeding 90°F in 2018 or in 1948? The difference in number of days exceeding 90°F (number of days in 2018 - number of days in 1948) was calculated for each of the 197 locations.

The average of these differences was 2.9 days with a standard deviation of 17.2 days. We are interested in determining whether these data provide strong evidence that there were more days in 2018 that exceeded 90°F from NOAA's weather stations.

Use a hypothesis test $= 0.01$ to investigate if there is evidence of more $>$ 90 days in 2018.

Independent samples

Two independent samples

- (Crabs): Can certain morphological features (e.g. size) distinguish between crab species? A sample of n_1 blue crabs and n_2 orange crabs is taken, and the rear width is measured on each crab.
- (Airbnbs): Collect data from n randomly-chosen Airbnb listings in New Orleans. I want to know whether the average price is significantly higher for the private room listings than the entire home listings.
- (Arthritis treatment): Does a new arthritis compound work better than the current drug? Randomly assign n_1 patients to the new drug and n_2 to the current therapy and measure grip strength in both groups.

Data and notation

Neighbourhood	Room_Type	Price	Num
Navarre	Entire home/apt	300	
Leonidas	Entire home/apt	100	
St. Claude	Private room	115	
Bywater	Entire home/apt	50	
St. Roch	Entire home/apt	65	
Treme - Lafitte	Entire home/apt	325	
Bywater	Entire home/apt	200	
Bayou St. John	Entire home/apt	123	
St. Claude	Entire home/apt	130	
St. Roch	Entire home/apt	105	
Bywater	Private room	99	
Seventh Ward	Entire home/apt	91	
Marlyville - Fontainbleau	Private room	80	
St. Claude	Entire home/apt	102	
Marigny	Private room	75	
Tall Timbers - Brechtel	Entire home/apt	75	
Tall Timbers - Brechtel	Entire home/apt	55	

Data. Typically, one quantitative variable, measured on two groups of individuals.

Notation. The n_1 observations from group one are denoted $y_{11}, \ldots, y_{1n_1}$. The n_2 observations from group one are denoted $y_{21}, \ldots, y_{2n_2}$.

$\bar{y}_{1.}$ = sample mean in group 1; s_1 = sample standard deviation in group 1.
$\bar{y}_{2.}$ = sample mean in group 2; s_2 = sample standard deviation in group 2.

Two statistical models

There are two different statistical models that can be used in independent two-sample t procedures.

Equal variances	$y_{11}, \ldots, y_{1n_1}$ are i.i.d. and approximately $N(\mu_1, \sigma^2)$. $y_{21}, \ldots, y_{2n_2}$ are i.i.d. and approximately $N(\mu_2, \sigma^2)$.

Unequal variances	$y_{11}, \ldots, y_{1n_1}$ are i.i.d. and approximately $N(\mu_1, \sigma_1^2)$. $y_{21}, \ldots, y_{2n_2}$ are i.i.d. and approximately $N(\mu_2, \sigma_2^2)$.

Two independent samples

In two-sample problems, a common research question is: Is there a significant difference between the means of two populations? The parameter of interest is the difference between the two means ($\mu_1 - \mu_2$ or, equivalently, $\mu_2 - \mu_1$).

- If the two groups have different means ($\mu_1 - \mu_2$ is far from zero), then this indicates a type of association between the numeric variable and the grouping variable.
- If the two groups have equal or nearly equal means, there is weak or no association between the numeric variable and the grouping variable.

Descriptive analyses of two independent samples

- Use side-by-side histograms and boxplots to visually compare the distributions of the numeric variable.
- Compare summary statistics across the two groups.
- Look at the difference between $\bar{y}_{1.}$ and $\bar{y}_{2.}$.

CI for difference between means

t interval for difference between means

The $(1 - \quad) \times 100\%$ t confidence interval for $\mu_1 - \mu_2$ is

Equal variances	$\bar{y}_{1.} - \bar{y}_{2.} \pm t^*_{df,\alpha/2} s_p \sqrt{1/n_1 + 1/n_2}$
Unequal variances	$\bar{y}_{1.} - \bar{y}_{2.} \pm t^*_{df,\alpha/2} \sqrt{s_1^2/n_1 + s_2^2/n_2}$

where

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

and the t^* values are found using the t distribution with df given below:

Equal variances	t distribution with df$= n_1 + n_2 - 2$.
Unequal variances	t distribution with df $= \frac{(n_1-1)(n_2-1)}{(1-C)^2(n_1-1)+C^2(n_2-1)}$, where $C = \frac{s_1^2/n_1}{s_1^2/n_1+s_2^2/n_2}$.

Example (Airbnbs)

In a random sample of 47 Airbnb listings from New Orelans, the average price per night and room type were recorded. Use the following summary statistic to find a 95% CI for the difference in mean price between private rooms and entire home/apartment.

Type	mean	standard dev.	n
Private room	74.9	21.8	9
Entire apt/home	212.7	193.0	38

Interpretation of two-sample t intervals

- (Sodium): Does the average level of sodium consumption differ across omnivore and vegan diets? Conduct an observational study to measure the sodium consumption of n_1 randomly selected omnivores and of n_2 randomly selected vegans. The 95% CI for $\mu_1 - \mu_2$ is found to be $(-102, 89)$.

  Interpretation: I am 95% confident that the mean sodium consumption is between 102 mg lower and 89 mg high for omnivores than for vegans.

Interpretation of two-sample t intervals

- (Sodium): Does the average level of sodium consumption differ across omnivore and vegan diets? Conduct an observational study to measure the sodium consumption of n_1 randomly selected omnivores and of n_2 randomly selected vegans. The 95% CI for $\mu_1 - \mu_2$ is found to be $(-102, 89)$.

  Interpretation: I am 95% confident that the mean sodium consumption is between 102 mg lower and 89 mg high for omnivores than for vegans.

- (Arthritis treatment): Does a new arthritis compound work better than the current drug? Randomly assign n_1 patients to the new drug and n_2 to the current therapy and measure the grip strength in both groups. The 95% CI for $\mu_1 - \mu_2$ is found to be $(2.1, 5.4)$.

  Interpretation: I am 95% confident that the mean grip strength is between 2.1 and 5.4 kg higher under the new drug than under the current therapy.

How to choose between the two models

- Best case scenario: use your prior knowledge about the sampled populations. Is there reason to believe the measurements from one group will be more variable?
- Compare the sample standard deviations s_1 and s_2. Use your judgment to discern whether they are close enough that it would be reasonable to assume that the population-level variability is equal across the groups.
- Perform a formal test of $H_0 : \sigma_1 = \sigma_2$. If you reject H_0, use the unequal variances model. This test is approximately accurate only if the data are very close to normal.
- When the two groups have equal sample sizes, the t procedures will give similar answers regardless of which model is chosen.

Test for difference between means

Two sample t test

Hypotheses. $H_0 : \mu_1 - \mu_2 = D_0$

$H_A : \mu_1 - \mu_2 \neq D_0 \quad (>, <)$

Test statistic.

Equal variances	$t_0 = \frac{\bar{y}_{1.} - \bar{y}_{2.} - D_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$.
Unequal variances	$t_0 = \frac{\bar{y}_{1.} - \bar{y}_{2.} - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

p-value and decision. For the two-sided alternative: find $P(T > |t_0|)$ where T has a t distribution with degrees of freedom given below. Reject H_0 if the p-value is less than .

Equal variances	df $= n_1 + n_2 - 2$.
Unequal variances	df$= \frac{(n_1-1)(n_2-1)}{(1-C)^2(n_1-1) + C^2(n_2-1)}$, where $C = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}$.

Example (Airbnb)

In a random sample of 47 Airbnb listings from New Orelans, the average price per night and room type were recorded. Is the average price for Entire homes significantly higher than that of private rooms? Use the following summary statistics.

Type	mean	standard dev.	n
Private room	74.9	21.8	9
Entire apt/home	212.7	193.0	38

Example (crabs)

A sample of n_1 blue crabs and n_2 orange crabs is taken and the crabs' rear widths are measured. Is there evidence of a significant difference in average rear width across the species? Use the data in `all_crabs.csv` on Canvas. Use $= 0.05$ for the test.

Independent or paired samples?

- The paired and independent t procedures both address the question of whether the means of two groups are equal.
- Determining whether data are paired or independent requires consideration of how the data were collected.
- Would it make sense to calculate individual differences? If so, the data may be paired.
- Would it make sense to arrange the numeric measurements in adjacent columns in a spreadsheet? If so, the data may be paired.

	A	B	C
1	School	2016-2017 Graduation rate	2017-2018 Graduation Rate
2	A	70.9	70.9
3	B	90.3	86.3
4	C	93.4	82.4
5	D	69.9	74.3
6	E	96.2	97.1
7	F	76	80.7
8	G	80.9	83.6
9	H	85.4	85.2
10	I	57.2	58.9
11	J	53.6	53.9
12	K	80.9	83
13	L	95.9	95.1
14			

Neighbourhood	Room_Type	Price	Num
Navarre	Entire home/apt	300	
Leonidas	Entire home/apt	100	
St. Claude	Private room	115	
Bywater	Entire home/apt	50	
St. Roch	Entire home/apt	65	
Treme - Lafitte	Entire home/apt	325	
Bywater	Entire home/apt	200	
Bayou St. John	Entire home/apt	123	
St. Claude	Entire home/apt	130	
St. Roch	Entire home/apt	105	
Bywater	Private room	99	
Seventh Ward	Entire home/apt	91	
Marlyville - Fontainbleau	Private room	80	
St. Claude	Entire home/apt	102	
Marigny	Private room	75	
Tall Timbers - Brechtel	Entire home/apt	75	
Tall Timbers - Brechtel	Entire home/apt	55	

Model checking

- In the two independent samples t procedures, approximate normality is assumed within both groups of measurements.
- To check normality graphically, you should make one normal quantile plot for each group.
- Always keep in mind the i.i.d. assumption – this requires good data collection.
- Outliers can have a strong influence on inferential results.