Lecture 7 practice

Stats 7 Summer Session II 2022

CLT

If we are interested in the mean of some numerical variable for a population

- When the population standard deviation σ is known:
  - SE = σ / √n
  - Test statistic has a Z-distribution
- When the population standard deviation σ is unknown (most common):
  - SE = s / √n
  - Test statistic has a t-distribution

Identify the critical value

An independent random sample is selected from an approximately normal population with unknown standard deviation. Find the degrees of freedom and the critical t-value (t^*) for the given sample size and confidence level.

(a) n = 21, CL = 90%
df = n - 1 = 21 - 1 = 20

```
> qt((1 - 0.9)/2, df = 20)
[1] -1.724718
```

$t^* = 1.72$

(b) n = 21, CL = 98%
df = n - 1 = 21 - 1 = 20

```
> qt((1 - 0.98)/2, df = 20)
[1] -2.527977
```

$t^* = 2.53$

(b) n = 11, CL = 98%
df = n - 1 = 11 - 1 = 10

```
> qt((1 - 0.98)/2, df = 10)
[1] -2.763769
```

$t^* = 2.76$

Find the p-value

An independent random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given sample size, test statistic, and H_A. Also determine if H_0 would be rejected at $\alpha = 0.05$.

(a) $H_A: \mu < \mu_0$, $n = 20$, $T = -2.04$

```
> pt(-2.04, df = 19)
[1] 0.02774972
```

df = n - 1 = 20 - 1 = 19

P-value = 0.0277 < α = 0.05 so reject H_0, found strong evidence to support H_A.

(b) $H_A: \mu > \mu_0$, $n = 20$, $T = -2.04$

```
> 1 - pt(-2.04, df = 19)
[1] 0.9722503
```

df = n - 1 = 20 - 1 = 19

P-value = 0.9723 > α = 0.05 so fail to reject H_0, did not find strong evidence to support H_A.

Find the p-value

An independent random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given sample size, test statistic, and H_A. Also determine if H_0 would be rejected at $\alpha = 0.05$.

(c) $H_A: \mu \neq \mu_0$, $n = 20$, $T = -2.04$

```
> 2 * pt(-2.04, df = 19)
[1] 0.05549944
```

df = n - 1 = 20 - 1 = 19

P-value = 0.0555 > α = 0.05 so fail to reject H_0,

did not find strong evidence to support H_A.

(d) $H_A: \mu \neq \mu_0$, $n = 20$, $T = 2.04$

```
> 2 * (1 - pt(2.04, df = 19))
[1] 0.05549944
```

df = n - 1 = 20 - 1 = 19

P-value = 0.0555 > α = 0.05 so fail to reject H_0,

did not find strong evidence to support H_A.

Normality check

RULES OF THUMB: HOW TO PERFORM THE NORMALITY CHECK

There is no perfect way to check the normality condition, so instead we use two rules of thumb:

$n < 30$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

$n \geq 30$: If the sample size n is at least 30 and there are no *particularly extreme* outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

Sleep habits of New Yorkers

New York is known as “the city that never sleeps”. A random sample of 25 New Yorkers were asked how much sleep they get per night. Statistical summaries of these data are shown below. The point estimate suggests New Yorkers sleep less than 8 hours a night on average. We want to know is the result statistically significant?

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

(a) Write the hypotheses in symbols and in words.

H_0: $\mu = 8$ (New Yorkers sleep 8 hrs per night on average.)

H_A: $\mu \neq 8$ (New Yorkers sleep less or more than 8 hrs per night on average.)

Sleep habits of New Yorkers

New York is known as “the city that never sleeps”. A random sample of 25 New Yorkers were asked how much sleep they get per night. Statistical summaries of these data are shown below. The point estimate suggests New Yorkers sleep less than 8 hours a night on average. We want to know is the result statistically significant?

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

(b) What are the conditions to check for a one sample t-distribution test?

- Independence (usually met if the sample is random)
- No concerning outliers (usually use rule of thumb to flag unusual values: 7.73 ± 0.77 x 2.5)

Sleep habits of New Yorkers

H_0: μ = 8 (New Yorkers sleep 8 hrs per night on average.)
H_A: μ ≠ 8 (New Yorkers sleep less or more than 8 hrs per night on average.)

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

(c) We are assuming the conditions were met. Calculate the test statistic, T, and the associated degrees of freedom.

df = n - 1 = 25 - 1 = 24

$$T = \frac{\text{Point estimate} - \text{null value } \mu_0}{\text{SE}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7.73 - 8}{\frac{0.77}{\sqrt{25}}} = -1.75$$

Sleep habits of New Yorkers

H_0: $\mu = 8$ (New Yorkers sleep 8 hrs per night on average.)

H_A: $\mu \neq 8$ (New Yorkers sleep less or more than 8 hrs per night on average.)

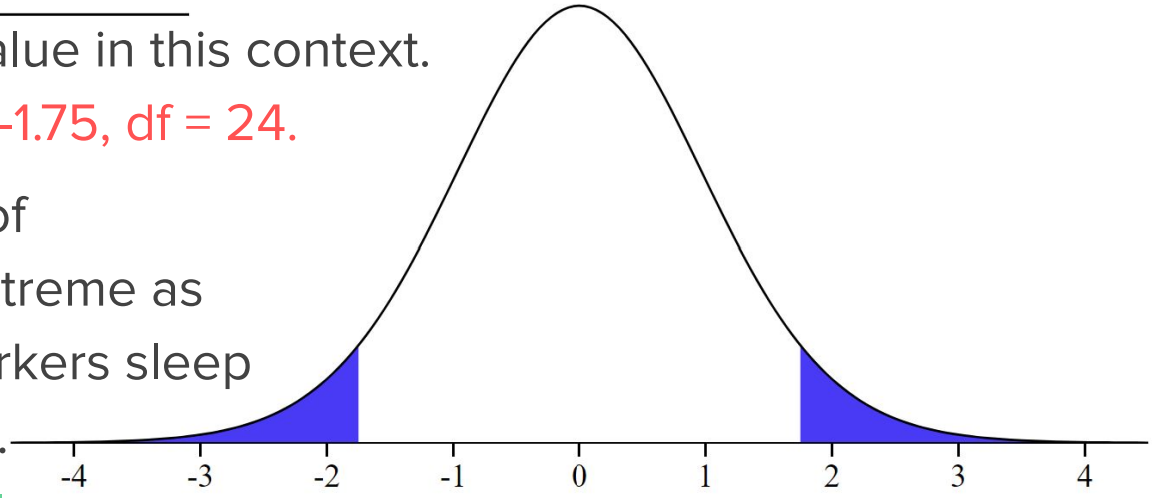
n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

```
> 2 * pt(-1.75, df = 24)
[1] 0.09289509
```

(d) Find and interpret the p-value in this context.

This is a two tail test with T = -1.75, df = 24.

There is a 9.29% probability of observing data as or more extreme as what we observed, if New Yorkers sleep 8 hours per night on average.



Air quality

Air quality measurements were collected in a random sample of 25 country capitals in 2013, and then again in the same cities in 2014. We would like to use these data to compare average air quality between the two years. Should we use a paired or non-paired test? Explain your reasoning.

Paired, data are recorded in the same cities at two different time points. The temperature in a city at one point is not independent of the temperature in the same city at another time point.

Paired or not?

In each of the following scenarios, determine if the data are paired.

(a) Compare pre- (beginning of semester) and post-test (end of semester) scores of students.

Since it’s the same students at the beginning and the end of the semester, there is a pairing between the datasets, for a given student their beginning and end of semester grades are dependent.

Paired or not?

In each of the following scenarios, determine if the data are paired.

(b) Assess gender-related salary gap by comparing salaries of randomly sampled men and women.

Since the subjects were sampled randomly, each observation in the men's group does not have a special correspondence with exactly one observation in the other (women's) group.

Paired or not?

In each of the following scenarios, determine if the data are paired.

(c) Compare artery thicknesses at the beginning of a study and after 2 years of taking Vitamin E for the same group of patients.

Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester artery thickness are dependent.

Paired or not?

In each of the following scenarios, determine if the data are paired.

(d) Assess effectiveness of a diet regimen by comparing the before and after weights of subjects.

Since it’s the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester weights are dependent.

True or false: paired

Determine if the following statements are true or false. If false, explain.

(a) In a paired analysis we first take the difference of each pair of observations, and then we do inference on these differences.

True, we use the difference of each pair as our one sample.

True or false: paired

Determine if the following statements are true or false. If false, explain.

(b) Two data sets of different sizes cannot be analyzed as paired data.

True, since we do inference on the difference for each pair of observations we need full pairs meaning equal sample sizes.

True or false: paired

Determine if the following statements are true or false. If false, explain.

(c) Consider two sets of data that are paired with each other. Each observation in one data set has a natural correspondence with exactly one observation from the other data set.

True

True or false: paired

Determine if the following statements are true or false. If false, explain.

(d) Consider two sets of data that are paired with each other. Each observation in one data set is subtracted from the average of the other data set's observations

False, each observation in one data set is subtracted from the corresponding one in the other data set.

Global warming

Let's consider a limited set of climate data, examining temperature differences in 1948 vs 2018. We randomly 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 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.

(a) Is there a relationship between the observations collected in 1948 and 2018? Or are the observations in the two groups independent? Explain.

For each observation in one data set, there is exactly one specially corresponding observation in the other data set for the same geographic location. The data are paired.

Global warming

Let's consider a limited set of climate data, examining temperature differences in 1948 vs 2018. We randomly 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 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.

(b) Write hypotheses for this research in symbols and in words.

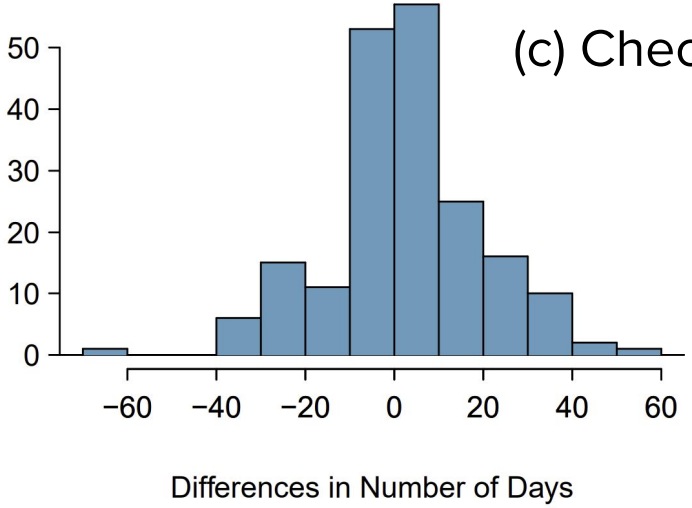
We will examine the differences (2018 value - 1948 value) for each location.

H_0 : $\mu_{difference} = 0$ (There is no difference in average number of days exceeding 90°F in 1948 and 2018 for NOAA stations.)

H_A: $\mu_{difference} > 0$ (2018 had a higher average number of days exceeding 90°F than 1948, for NOAA stations.

Global warming

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. $n = 197, \bar{x} = 2.9, s = 17.2$



(c) Check the conditions required to complete this test.

Locations were randomly sampled, so independence is reasonable. The sample size is at least 30, so we're just looking for particularly extreme outliers: none are present (the observation off left in the histogram would be considered a clear outlier, but not a particularly extreme one). The conditions are satisfied.

Global warming

(d) Calculate the test statistic and find the p-value.

$$n = 197, \bar{x} = 2.9, s = 17.2$$

$$\text{Test statistic} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2.9 - 0}{17.2/\sqrt{197}} = 2.37$$

$$df = n - 1 = 197 - 1 = 196$$

The alternative hypothesis is greater than so right tail test

```
> 1 - pt(2.37, df = 196)
[1] 0.009379488
```

Global warming

(e) Use $\alpha = 0.05$ to evaluate the test, and interpret your conclusion in context.

The p-value was 0.0094, which is less than the significance level of 0.05.

This means we reject the null hypothesis in support of the alternative.

We found strong evidence (p-value = 0.0094 < $\alpha = 0.05$) that 2018 had a higher average number of days exceeding 90℉ than 1948, for NOAA stations.

Global warming

(f) What type of error might we have made? Explain in context what the error means

Type 1 Error, since we may have incorrectly rejected H_0. This error would mean that NOAA stations did not actually observe an increase, but the sample we took just so happened to make it appear that this was the case.

Global warming

(g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the number of days exceeding 90°F in 2018 and 1948 to include 0? Explain your reasoning.

No, since we rejected H_0, which had a null value of 0.

Global warming

(h) Calculate the 95% confidence interval for the average difference between the number of days exceeding 90°F in 2018 and 1948.

$$n = 197, \bar{x} = 2.9, s = 17.2$$

```
> qt((1 - 0.95) / 2, df = 196)
[1] -1.972141
```

$$\text{point estimate} \pm t^* \times SE$$

$$\bar{x} \pm t* \times \frac{s}{\sqrt{n}}$$

$$2.9 \pm 1.97 * \frac{17.2}{\sqrt{197}}$$

$$2.9 \pm 2.41$$

$$(0.49, 5.31)$$

We are 95% confident that the average difference between the number of days exceeding 90°F in 2018 and 1948 is between 0.49 and 5.31.

Chicken diet and weight

Chicken farming is a multi-billion dollar industry, and any methods that increase the growth rate of young chicks can reduce consumer costs while increasing company profits, possibly by millions of dollars. An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. We are specifically interested in learning if the average weights of chickens that were fed linseed and horsebean are different.

(a) What type of inference could be appropriate here?

We are interested in comparing two samples of weights with unpaired data so a two-sample hypothesis test for a difference in mean could be appropriate.

Chicken diet and weight

...An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. We are specifically interested in learning if the average weights of chickens that were fed horsebean and linseed and are different.

(b) What are the hypotheses?

H_0: $\mu_{horsebean} - \mu_{linseed} = 0$.

H_A: $\mu_{horsebean} - \mu_{linseed} \neq 0$.

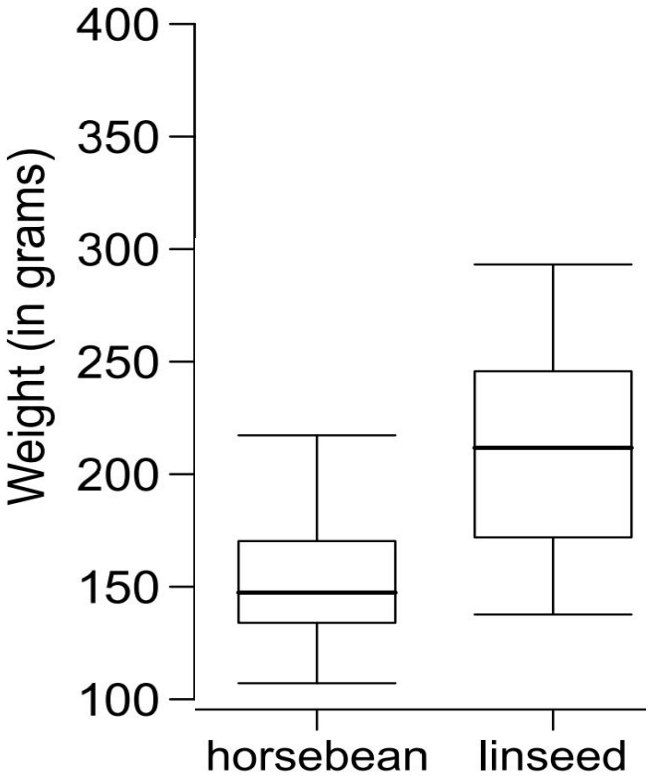
Chicken diet and weight

$$\text{Horsebean} : n = 10, \bar{x} = 160.20, s = 38.63$$
$$\text{Linseed} : n = 12, \bar{x} = 218.75, s = 52.24$$

To the left are box plots showing the distribution of weights by feed type.

(c) Check the conditions for this hypothesis test.

- We need independence within and between groups.
  - The newly hatched chicks were randomly allocated to feed group so there should independence within and between groups
- We need no clear outliers, the boxplot shows none



Chicken diet and weight

(d) Compute the test statistic and p-value.

$$\text{Horsebean} : n = 10, \bar{x} = 160.20, s = 38.63$$

$$\text{Linseed} : n = 12, \bar{x} = 218.75, s = 52.24$$

$$\text{T} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(160.20 - 218.75) - 0}{\sqrt{\frac{38.63^2}{10} + \frac{52.24^2}{12}}} = -3.02$$

$$df = min(n_1 - 1, n_2 - 1) = min(10 - 1, 12 - 1) = min(9, 11) = 9$$

Two tailed test so p-value is:

```
> 2 * pt(-3.02, df = 9)
[1] 0.01447932
```

Chicken diet and weight

(e) Do these data provide strong evidence that the average weights of chickens that were fed horsebean and linseed are different? Use a 5% significance level.

The p-value is 0.0144 which is less than the significance level by a lot.

Reject the null hypothesis, we found evidence to support the alternative.

The data provide strong evidence (p-value = 0.0144 < α = 0.05) that there is a significant difference between the average weights of chickens that were fed horsebean and linseed.

Chicken diet and weight

(f) Say we wanted to estimate the difference in average weights of chickens that were fed horsebean and linseed. Compute and interpret a 90% confidence interval for this difference.

```
> qt((1 - 0.9)/2, df = 9)
[1] -1.833113
```

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(160.20 - 218.75) \pm 1.83 \times \sqrt{\frac{38.63^2}{10} + \frac{52.24^2}{12}}$$

$$(-94.07, -23.03)$$

We are 90% confident that the difference in average weights of chickens that were fed horsebean is between 94.1 and 23.0 grams less than the average for linseed fed chicks.

Credits

Examples adapted from OpenIntro Statistics (4th edition) by David Diez, Mine Cetinkaya-Rundel, and Christopher D Barr https://www.openintro.org/book/os/ protected under the Creative Commons License