Topics in Computer Science

Lecture 10

Melvin Ayala

Lecture 10: Sampling and Hypothesis Testing

Sections

1. Sampling

1.1. Definitions

Sampling:

- selecting the group that you will collect data from in your research.
- primary types of sampling methods:
  - probability sampling: involves random selection, allowing you to make strong statistical inferences about the whole group.
  - non-probability sampling: non-random selection based on convenience or other criteria
- allows you to test a hypothesis about the characteristics of a population.
- is the selection of a subset (a statistical sample) of individuals from within a statistical population to estimate characteristics of the whole population.
- statisticians attempt to collect samples that are representative of the population in question.

Population vs. Sample

- **population**: the entire group that you want to draw conclusions about.
- **sample**: the specific group of individuals that you will collect data from.
- population can be defined in terms of geographical location, age, income, or many other characteristics.

Sample size:

- the number of individuals you should include in your sample
- depends on various factors, including the size and variability of the population and your research design.
- there are different sample size calculators and formulas depending on what you want to achieve with statistical analysis.

1.2. Probability Sampling Methods

1. Simple random sampling

every member of the population has an equal chance of being selected.
should include the whole population.

2. Systematic sampling

similar to simple random sampling, but it is usually slightly easier to conduct.
every member of the population is listed with a number
individuals are chosen at regular intervals.
make sure that there is no hidden pattern in the list that might skew the sample.

3. Stratified sampling

involves dividing the population into subpopulations that may differ in important ways.
divides the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).

4. Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample.
Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

2. Hypothesis Testing

2.1. What is a Hypothesis?

Hypothesis:

- is an educated guess about something in the world around you. It should be testable, either by experiment or observation.
- for example:
  - a new medicine you think might work.
  - a way of teaching you think might be better.
  - a possible location of new species.
  - a fairer way to administer standardized tests.

2.2. What is a Hypothesis Statement?

Hypothesis Statement:

- the hypothesis and the level of assurance that we want to give to it.
- example:
    - hypothesis: the difference of two population means is due to chance
    - we must specify how much certain we want to be, e.g., 95%.
    - that would mean that we only have a probability of 5% of making a mistake by accepting or rejecting that hypothesis.

2.3. What is Hypothesis Testing?

Hypothesis Testing:

- a way for you to test the results of a survey or experiment to see if you have meaningful results.
- testing whether your results are valid
- if your results may have happened by chance, the experiment won't be repeatable and so has little use.
- Hypothesis testing is done with statistical tests

Examples #1:

- a researcher thinks that if knee surgery patients go to physical therapy twice a week (instead of 3 times), their recovery period will be longer.
- average recovery times for knee surgery patients is 8.2 weeks.
- hypothesis statement:
  - the researcher believes the average recovery time is more than 8.2 weeks.
  - this can be written in mathematical terms as: $H1: \mu > 8.2$

Null hypothesis:

That's what will happen if the researcher is wrong: $H0\ \mu \leq 8.2$

2.4. The Null Hypothesis

Statistical tests:

- provide mechanism for making quantitative decisions about a process
- intent: to determine whether there is enough evidence to "reject" a conjecture about the process
- conjecture is called the null hypothesis

Null Hypothesis:

- often denoted H_0
- the claim that no relationship exists between two sets of data or variables being analyzed.
- any experimentally observed difference is due to chance alone
- an underlying causative relationship does not exist, hence the term "null".
- In addition to the null hypothesis, an alternative hypothesis is also developed, which claims that a relationship does exist between two variables.

How to Accept or Reject the Null Hypothesis:

- take a random sample from the population, and then:
  - if the sample data are consistent with the null hypothesis, then do not reject the null hypothesis
  - if the sample data are inconsistent with the null hypothesis, then reject the null hypothesis and conclude that the alternative hypothesis is true.

The Null Hypothesis: Examples

Example 1:

A gambler may be interested in whether a game of chance is fair:

- if fair, then the expected earnings per play come to zero for both players. → the null hypothesis (H_0)
- if not, then the expected earnings are positive for only one player → the alternative hypothesis (H_a)

if the average earnings are sufficiently far from zero, then the gambler will reject the null hypothesis H_0 and conclude the alternative hypothesis H_a is true (but by how much?).

Example 2:

Given the test scores of two random samples, one of men and one of women, does one group differ from the other?

A possible null hypothesis is that the mean male score is the same as the mean female score:

H_0: $\mu_1 = \mu_2$

where:

H_0: the null hypothesis

μ_1: the mean of population 1

μ_2: the mean of population 2

2.5. Categories of Statistical Tests

Two broad categories:

- parameter/interval estimation hypothesis
- hypothesis tests

Parameter/interval estimation hypothesis

- is an estimate of the true parameter value is made using the sample data.
- called a point estimate or a sample estimate.
- confidence levels need to be provided
- confidence level:
  - you accept the null hypothesis if it is between the confidence level
- most commonly used estimation parameter: location is the mean
- example:
  - null hypothesis: the mean is $\mu = 10 \pm 0.3$ with a confidence level of 95%

Hypothesis tests

- attempts to refute a specific claim about a population parameter
- for example, the hypothesis might be one of the following:
  - the population mean is equal to 10
  - the population standard deviation is equal to 5
  - the means from two populations are equal
  - the standard deviations from 5 populations are equal

What does it mean to:

- accept a hypothesis? we do not have evidence to believe otherwise (we are x% sure)
- reject a hypothesis? we do not have evidence to believe it (we are not x% sure)

2.6. Common Format for a Hypothesis

Common Format for a Hypothesis:

H_0:	A statement of the null hypothesis, e.g., two population means are equal.
H_a:	A statement of the alternative hypothesis, e.g., two population means are not equal.
Test Statistic:	The test statistic is based on the specific hypothesis test.
Significance Level:	The significance level, α, defines the sensitivity of the test.

- a value of $\alpha = 0.05$ means that we inadvertently reject the null hypothesis 5% of the time when it is in fact true.
- also called the type I error.
- the choice of α is somewhat arbitrary, although in practice values of 0.1, 0.05, and 0.01 are commonly used.

3. Types of Statistical Tests

Most widely used tests:

- t-test
- Z-test
- chi-square test
- ANOVA test (analysis of variance)
- binomial test
- one sample median test
- Tests for mean range
- Tests for variance range

In General:

- most test statistics have the form:

$$t = \frac{Z}{s} = \frac{\bar{x} - \mu}{\hat{\delta}/\sqrt{n}}$$

where:

$\bar{x}$	sample means from a sample $x_1, x_2, \ldots, x_n$
s	standard error of the mean
$\hat{\delta}$	estimate of the standard deviation
μ	population mean

4. The t-Test

4.1. Background

t-test:

- any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis.
- is the most popular statistical test
- used to compare the means of two groups.
- often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.
- most commonly applied when:
    - the test statistic ollows a normal distribution
    - the value of a scaling term is known
- when the scaling term is estimated based on the data, the test statistic—under certain conditions—follows a Student's t distribution.
- t-test's most common application: to test whether the means of two populations are different.
- William Sealy Gosset first published it in 1908 using the pseudonym "Student" because his employer preferred staff to use pen names when publishing scientific papers.

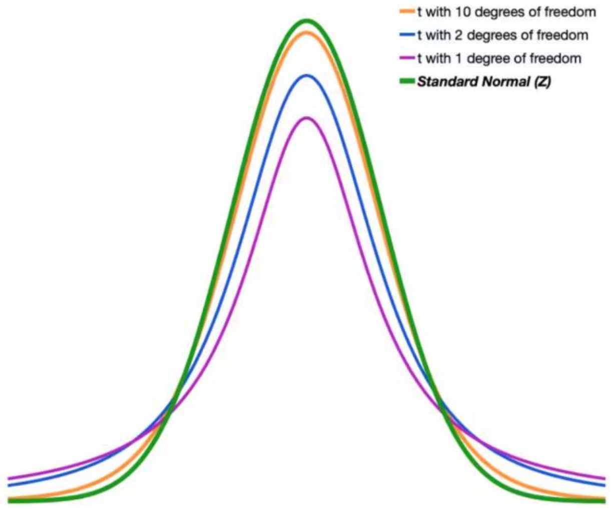
William Sealy Gosset, developed the "t-statistic" and published it under the pseudonym of "Student“.

The t-Test (Cont.)

Example:

You want to know whether the mean petal length of iris flowers differs according to their species:

- you find two different species of irises growing in a garden and measure 25 petals of each species.
- you can test the difference between these two groups using a t-test and null and alterative hypotheses.
- the null hypothesis (H_0) is that the true difference between these group means is zero.
- the alternate hypothesis (Ha) is that the true difference is different from zero.
- assumptions: your data:
  - are independent
  - are (approximately) normally distributed
  - have a similar amount of variance within each group being compared (a.k.a. homogeneity of variance)



The t-distribution depends on the degrees of freedom (df = n - 1)

4.2. Types of t-Tests

By the samples involved:

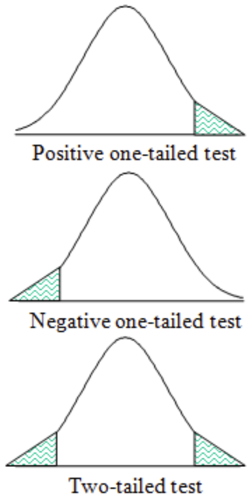
One-sample, two-sample, or paired t-test?

- **one-sample t-test:**
  If there is one group being compared against a standard value (e.g., comparing the acidity of a liquid to a neutral pH of 7), perform a one-sample t-test.
- **two-sample t-test:**
  If the groups come from two different populations (e.g., two different species, or people from two separate cities), perform a two-sample t-test (a.k.a. independent t-test). This is a between-subjects design.
- **paired t-test:**
  If the groups come from a single population (e.g., measuring before and after an experimental treatment), perform a paired t-test. This is a within-subjects design.

By the sign of the difference:

One-tailed or two-tailed t-test?

- **two-tailed t-test:**
  If you only care whether the two populations are different from one another, perform a two-tailed t-test.
- **one-tailed t-test:**
  If you want to know whether one population mean is greater than or less than the other, perform a one-tailed t-test.



In our test of whether petal length differs by species:

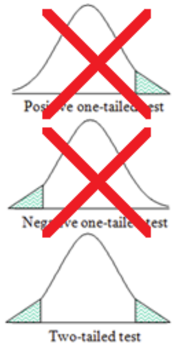
- the observations come from two separate populations (separate species) → **two-sample t-test**.
- we don't care about the direction of the difference → **two-tailed t-test.**

One-sample, two-sample, or paired t-test?

- **one-sample t-test:**
  If there is one group being compared against a standard value (e.g., comparing the acidity of a liquid to a neutral pH of 7), perform a one-sample t-test.
- **two-sample t-test:**
  If the groups come from two different populations (e.g., two different species, or people from two separate cities), perform a two-sample t-test (a.k.a. independent t-test). This is a between-subjects design.
- **paired t-test:**
  If the groups come from a single population (e.g., measuring before and after an experimental treatment), perform a paired t-test. This is a within-subjects design.

One-tailed or two-tailed t-test?

- **two-tailed t-test:**
  If you only care whether the two populations are different from one another, perform a two-tailed t-test.
- **one-tailed t-test:**
  If you want to know whether one population mean is greater than or less than the other, perform a one-tailed t-test.



How does it work?

Requires three fundamental data values:

- difference between the mean values from each data set
- standard deviation of each group
- the number of data values of each group

Produces a value as its output:

- t-value, or t-score:
  - a ratio of the difference between the mean of the two sample sets and the variation that exists within the sample sets.

The calculated t-value is then compared against a value obtained from a critical value table called the T-distribution table:

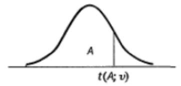
- smaller t-values → more similarity between the two sample sets.
  → accept H_0
- higher t-values → more differences between the two sample sets.
  → reject H_0

Obtain the critical t-values by looking up the percentiles of the t-Distribution using:

- the significance level
- the degrees of freedom

Percentiles of the t Distribution

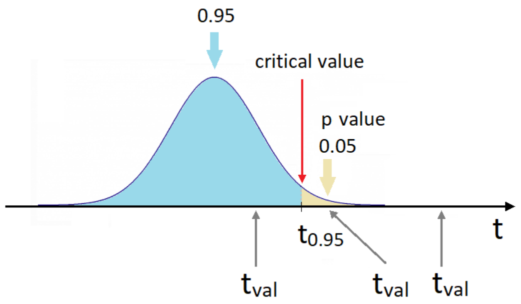
Entry is $t(A;\nu)$ where $P\{t(\nu) \leq t(A;\nu)\} = A$



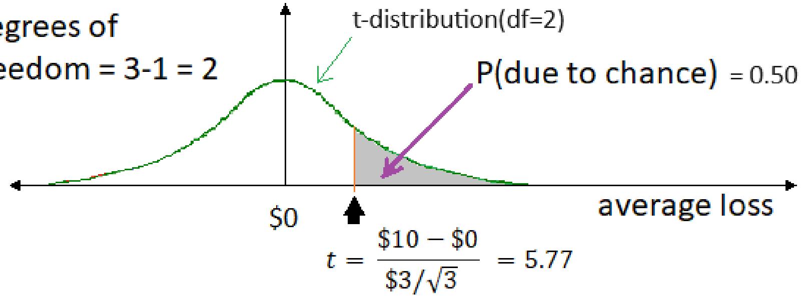
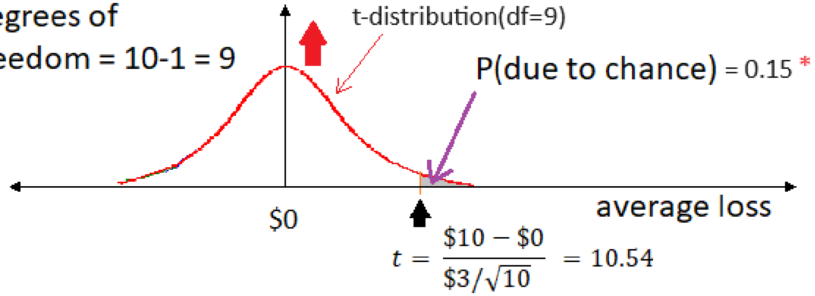
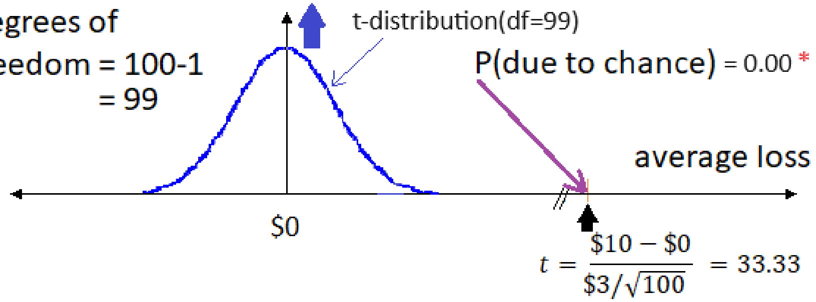
	A						
ν	.60	.70	.80	.85	.90	.95	.975
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069

Accepting or rejecting H_0:
depends on:

- the significance level required
- degrees of freedom
- t_{val} (i.e., $t_{val} < t_{0.95}$?)



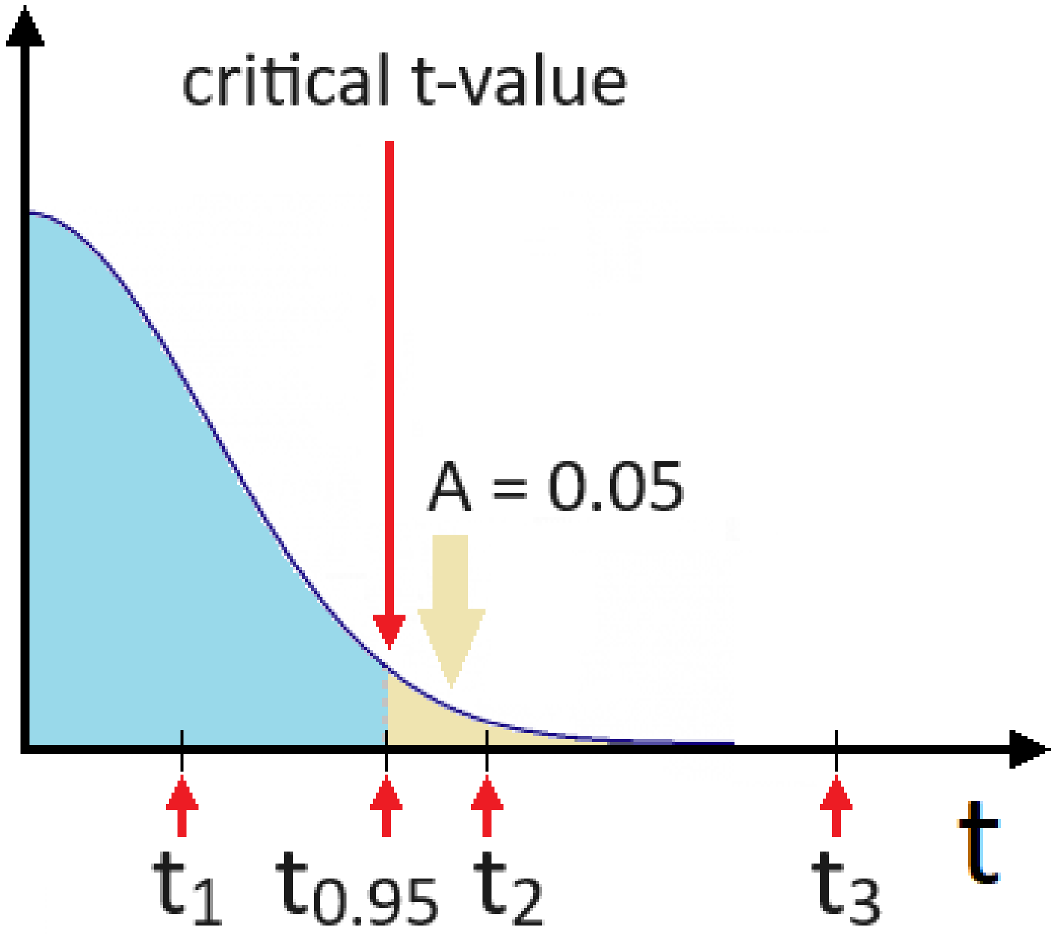
The probabilities depend on the degrees of freedom

times played in the casino	average loss	total loss	probability of that loss being due to chance	how is the probability computed?
3 times	$10	$30	low	degrees of freedom = 3-1 = 2 
10 times	$10	$100	medium	degrees of freedom = 10-1 = 9 
100 times	$10	$1,000	high	degrees of freedom = 100-1 = 99 

assume σ = $3

* in a one-tailed t-test

- the probability that a difference in means is **due to chance** is the integral under the curve from the t-value to the right.
- the higher the t-value, the smaller that probability.



A Numerical Example

Suppose we want to know whether or not the mean weight of a certain species of turtle is equal to 310 pounds.
To test this, will perform a one-sample t-test at significance level $\alpha = 0.05$ using the following steps:

Step 1: Gather the sample data:

Sample size $n = 40$

Sample mean weight $x = 300$

Sample standard deviation $s = 18.5$

Step 2: Define the Hypothesis:

We will perform the one sample t-test with the following hypotheses:

$H_0: \mu = 310$ (population mean is equal to 310 pounds)

$H_1: \mu \neq 310$ (population mean is not equal to 310 pounds)

Step 3: Calculate the test statistic t:

$t = (x - \mu) / (s/\sqrt{n}) = (300-310) / (18.5/\sqrt{40}) = -3.4187$

Step 4: Calculate the p-value of the test statistic t:

According to the T Score to P Value Calculator, the p-value associated with $t = -3.4817$ and degrees of freedom $= n-1 = 40-1 = 39$ is 0.00149.

Step 5: Draw a conclusion:

Since this p-value is less than our significance level $\alpha = 0.05$, we reject the null hypothesis. We have sufficient evidence to say that the mean weight of this species of turtle is not equal to 310 pounds.

4.3. Using Python to run a t-Test

a professor wants to know if two different studying methods lead to a different mean exam scores.

he recruits 10 students to use method A and 10 students to use method B.

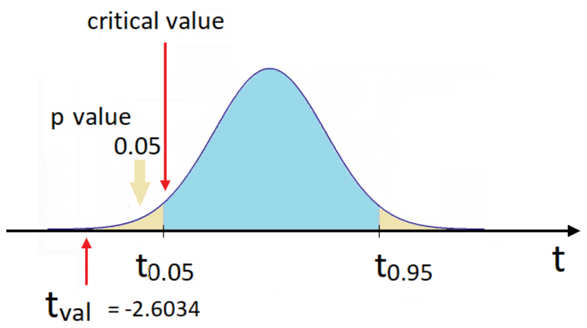
#	method	score
1	A	71
2	A	72
3	A	72
4	A	75
5	A	78
6	A	81
7	A	82
8	A	83
9	A	89
10	A	91
11	B	80
12	B	81
13	B	81
14	B	84
15	B	88
16	B	88
17	B	89
18	B	90
19	B	90
20	B	91

The code in Python could look like this:

```
import pandas as pd
from scipy.stats import ttest_ind
# create pandas dataframe
df = pd.DataFrame(
      {'method': ['A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B'],
       'score': [71, 72, 72, 75, 78, 81, 82, 83, 89, 91, 80, 81, 81, 84, 88, 88, 89, 90, 90, 91]})

# define samples by filtering
group1 = df[df['method'] == 'A']
group2 = df[df['method'] == 'B']

# perform independent two sample t-test
summary, results = ttest_ind(group1['score'], group2['score'])
print("t-value = ", summary)
print("p value = ", results)
print("end")
```



We obtain:

t-test statistic = -2.6034

p-value = 0.0179

Since the p value is less than 0.05, we reject the null hypothesis of the t-test and conclude that there is sufficient evidence to say that the two methods lead to different mean exam scores.

Code in Jupyter Notebooks

```
import pandas as pd
from scipy.stats import ttest_ind
# create pandas dataframe
df = pd.DataFrame(
        {'method': ['A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B'],
         'score': [71, 72, 72, 75, 78, 81, 82, 83, 89, 91, 80, 81, 81, 84, 88, 88, 89, 90, 90, 91]})

# define samples by filtering
group1 = df[df['method'] == 'A']
group2 = df[df['method'] == 'B']
# perform independent two sample t-test
summary, results = ttest_ind(group1['score'], group2['score'])
print("t-value = ", summary)
print("p value = ", results)
print("The probability that the difference in the scores of the two methods is just by chance is {:.2%}".format(results));
print("end")
```

```
t-value =  -2.6034304605397938
p value =  0.017969284594810425
The probability that the difference in the scores of the two methods is just by chance is 1.80%
end
```

Code in PyCharm

```python
import pandas as pd
from scipy.stats import ttest_ind
# create pandas dataframe
df = pd.DataFrame({'method': ['A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A',
                              'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B'],
                   'score': [71, 72, 72, 75, 78, 81, 82, 83, 89, 91, 80, 81, 81, 84, 88, 88, 89, 90, 90, 91]})

# define samples by filtering
group1 = df[df['method'] == 'A']
group2 = df[df['method'] == 'B']
# perform independent two sample t-test
summary, results = ttest_ind(group1['score'], group2['score'])
print("t-value = ", summary)
print("p value = ", results)
print("end")
```

Debug: ttest

Debugger Console

```
C:\Users\melvi\Anaconda3\envs\Concrete\python.exe "C:\Program Files\JetBrains\PyCharm Community Edition with Anaconda plu
pydev debugger: process 4660 is connecting

Connected to pydev debugger (build 193.7288.30)
t-value =  -2.6034304605397938
p value =  0.017969284594810425
end

Process finished with exit code 0
```

End of Lecture