T-tests statistics

Jolla Kullgren
Department of Chemistry - Ångström

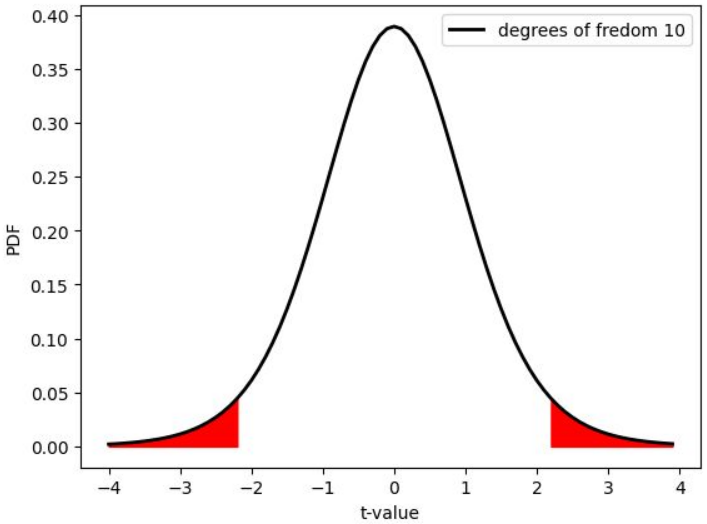
Central Limit Theorem (CLT)

- For independent and identically distributed random variables, mean values tend to be normally distributed.
- This holds even if the random variable itself is not randomly distributed.

Student's t-test

- After William Sealy Gosset
- "t-statistic" is abbreviated from "hypothesis test statistic"
- The t-test can be used to test whether the means of two populations are different or to test whether the mean of a single distribution is larger, smaller or equal to a claimed value.

One-sample t-test



The t-distribution for 10 degrees of freedom. If our test value are in the red areas we can reject the null hypothesis. Use only left/right when testing for mean smaller/larger than claimed value.

- With the one-sample t-test we test a claim on the mean value. We need to compute the parameter:

$$t_{n-1}^{test} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

where is the mean value, is the claim, s is the variance in the data and n the number of samples (degrees of freedom=n-1).

- This value is then used to compute a probability, P, for the null hypothesis (i.e. the risk of our claim to be wrong).

$$\bar{x} > \mu_0 \Rightarrow P(t_{n-1} > t_{n-1}^{test})$$
$$\bar{x} < \mu_0 \Rightarrow P(t_{n-1} < t_{n-1}^{test})$$
$$\bar{x} = \mu_0 \Rightarrow 2P(t_{n-1} > |t_{n-1}^{test}|)$$

UPPSALA UNIVERSITET

Example (One-sample t-test):

Assume that we have the following data:

$$\bar{x} = 44.9,\ s = 8.9,\ n = 15$$

How likely is it that the true mean is larger than 40?

$$t_{14}^{test} = \frac{44.9 - 40}{8.9/\sqrt{15}} = 2.13$$

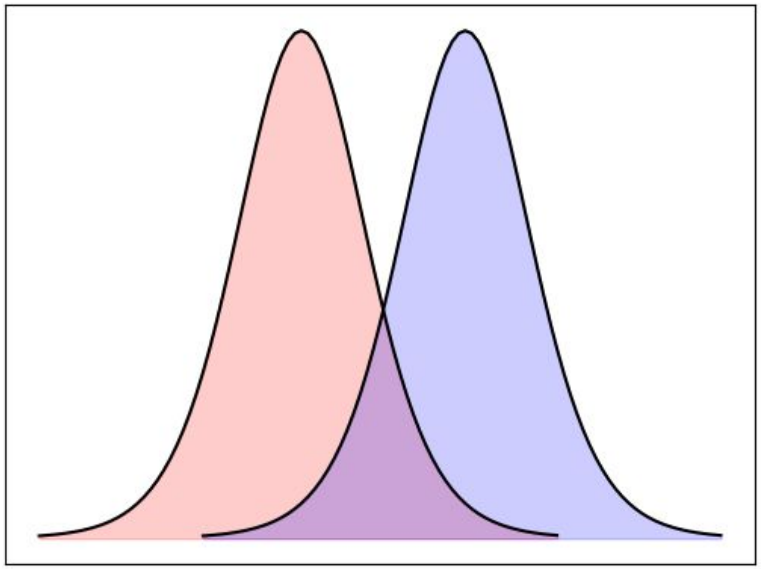
$$P(t_{14} > 2.13) = 0.026 < 0.05$$

Reject null hypothesis! Claim is significant.

Example (One-sample t-test) in python:

```
import numpy as np
from scipy import stats

data=[0.52508266, 0.09735529, 0.18190318, 0.55766562, 0.37788684, 0.44707692]


t_test, p_value = stats.ttest_1samp(data , popmean=0.5 , alternative='less')

print("t-test: ",t_test," -> P-value: ",p_value)



==========================================================================
=

t-test:  -1.775161048201416  -> P-value:  0.06802008896186788
```

Independent two-sample t-test



With the ***independent two-sample t-test*** we test if two data-sets could share the same mean-value.

- With the independent two-sample t-test we test if the mean value from two data-sets are significantly different.
- There are different formulations for the problem when sizes and expected variance of the data-set differ.
- The procedure is otherwise very similar to the one-sample case. In the case of equal data-size and variance we have:

$$t_{n-1}^{test} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{2}s/\sqrt{n}}$$

$$n = n_1 = n_2 \qquad s = \sqrt{\frac{s_1^2 + s_2^2}{2}}$$

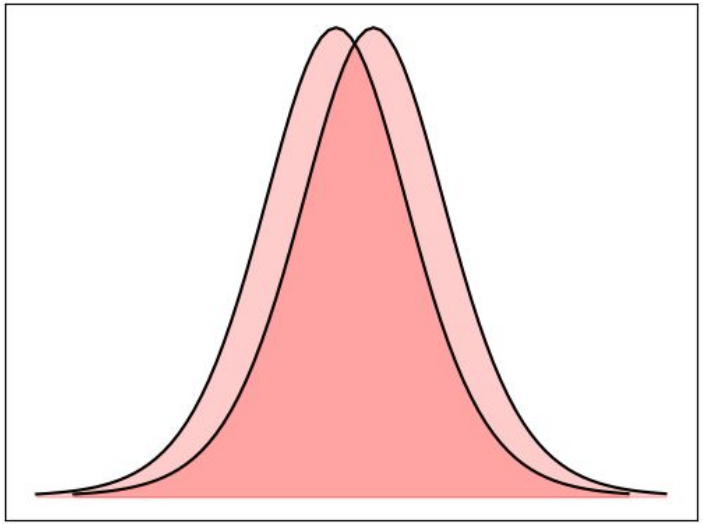
Example (independent two-sample t-test) in python:

```
import numpy as np
from scipy import stats

data_1=[0.61185506, 0.32452449, 1.18288771, 0.31038965, 0.26552593]
data_2=[0.03016453, 0.43172039, 0.92430701, 0.31756139]

t_test, p_value = stats.ttest_ind(data_1, data_2)

print("t-test: ",t_test," -> P-value: ",p_value)
```

```
==========================================================================
=

t-test:  0.4439095335692569  -> P-value:  0.6705058607443215
```

P-value suggest we cannot reject null hypothesis. The mean values are not significantly different.
Note that data-sets are small!

Dependent two-sample t-test



In the ***dependent two-sample t-test*** we check for consistency between two sets of measurement.

- In the dependent two-sample t-test we consider repeated measurements.

$$t_{n-1}^{test} = \frac{\bar{x_d}}{s_d/\sqrt{n}}$$

where the subscript d signifies that we now use the average and standard deviation of the differences between all pairs.

- Note #1: The P-value can be low when the values in x_2 is close to a scalar times x_1.
- Note #2: The P-value is zero if x_1 and x_2 differ by a single scalar.

UPPSALA UNIVERSITET

Example (dependent two-sample t-test) in python:

```
import numpy as np
from scipy import stats

data_1=[0.61185506, 0.32452449, 1.18288771, 0.31038965, 0.26552593]
data_2=[0.74016453, 0.43172039, 1.31430701, 0.41756139, 0.41846823]

t_test, p_value = stats.ttest_rel(data_1, data_2)

print("t-test: ",t_test," -> P-value: ",p_value)
```

```
==========================================================================
=

t-test:  -14.643136260522295  -> P-value:  0.00012654099656845676
```

P-value suggest we can reject the null hypothesis.

Example (dependent two-sample t-test) with scaled data in python:

```
import numpy as np
from scipy import stats

data_1=[0.61185506, 0.32452449, 1.18288771, 0.31038965, 0.26552593]
data_2=[2*x for x in data_1]
t_test, p_value = stats.ttest_rel(data_1, data_2)

print("t-test: ",t_test," -> P-value: ",p_value)
```

```
=============================================================================
=

t-test:  -3.13057624168559  -> P-value:  0.035165702564070 12
```

P-value becomes small for scaled data.

Example (dependent two-sample t-test) with shifted data in python:

```
import numpy as np
from scipy import stats

data_1=[0.61185506, 0.32452449, 1.18288771, 0.31038965, 0.26552593]
data_2=[x+1.0 for x in data_1]
t_test, p_value = stats.ttest_rel(data_1, data_2)

print("t-test: ",t_test," -> P-value: ",p_value)
```

```
================================================================================
=

t-test:  -1.64448207058845 42e+16  -> P-value:  8.204170532336355e-65
```

The P-value becomes zero for shifted data.

Summary

- The t-test can be used to test whether the means of two populations are different or to test whether the mean of a single distribution is larger, smaller or equal to a claimed value.