Significance tests

Part 1: t-Tests

By: Noureddin Sadawi, PhD

University of London

t-Tests

- ‘There are numerous types of significance tests, depending on whether the data comprises count data or measured data, how many samples there are, and what’s being measured.
- A very common one is the t-test, named after Student’s t-distribution, originally developed by W.S. Gosset to **approximate the distribution of a single sample mean**.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Key terms for t-Tests

- ‘**Test statistic:** A metric for the difference or effect of interest.
- **t-statistic:** A standardized version of common test statistics such as means.
- **t-distribution:** A reference distribution (in this case derived from the null hypothesis), to which the observed t-statistic can be compared.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

t-test

- A t-test is used to determine **whether the difference between the means of two groups is a significant**.
- The two groups may be related in some features.
- The t-test is a **hypothesis test** in statistics (one of many).
- Calculating a t-test requires three key values from the data:
  a. The difference between the group mean values (called the mean difference).
  b. the standard deviation of each group.
  c. the number of data values in each group (they do not have to be the same).
- There are many types of t-test.

Significance tests

- ‘All significance tests require that you specify a test statistic to measure the effect you are interested in and help you determine whether that observed effect lies within the range of normal chance variation.
- In a resampling test (see the discussion of permutation in “Permutation Test”), the scale of the data does not matter.
- You create the reference (null hypothesis) distribution from the data itself and use the test statistic as is.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

How t-test was developed

- ‘In the 1920s and 1930s, when statistical hypothesis testing was being developed, it was not feasible to randomly shuffle data thousands of times to do a resampling test.
- Statisticians found that a good approximation to the permutation (shuffled) distribution was the t-test, based on Gosset’s t-distribution.
- It is used for the very common two-sample comparison—A/B test—in which the data is numeric.
- But in order for the t-distribution to be used without regard to scale, a standardized form of the test statistic must be used.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Example: one sample t-test

- Let us say there is a volleyball team (assume the total number of players in the team is 20) and we would like to determine if the players in this team are shorter than players in other teams.
- We take the heights of the players in this team and compare their height to the average player height in the league.
- Let us assume that the average player height in the league is 201cm.
- H_0: $Mean_{team} = Mean_{league}$
- H_a: $Mean_{team} \neq Mean_{league}$

	PlayerHeight
0	188
1	198
2	190
3	199
4	180
5	195
6	180
7	185
8	186
9	202
10	181
11	180
12	202
13	188
14	199
15	192
16	197
17	186

Example: one sample t-test

- H_0: $Mean_{team} = Mean_{league}$
- H_a: $Mean_{team} \neq Mean_{league}$
- Population mean (μ) = 201cm.
- Sample mean = 190.44cm.
- STD of the team = 7.58 and n = 20.
- p-value for single sided test = 0.0000120.
- p-value < 0.05 so we reject the null hypothesis (i.e. we have sufficient evidence to conclude the alternative).
- There is sufficient evidence to suggest there is a difference in the height of players in this team and the height of other players in the league.

$$t = \frac{Z}{s} = \frac{\overline{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

t statistic

	PlayerHeight
0	188
1	198
2	190
3	199
4	180
5	195
6	180
7	185
8	186
9	202
10	181
11	180
12	202
13	188
14	199
15	192
16	197
17	186

Example: two sample t-test

- Obese men who suffer from sleep apnoea (two groups).
- One group (13 individuals) is given a specific diet for 9 weeks, the other group (14 individuals) have a normal diet for the same period of time.
- Number of sleep apnoea events are counted for each individual in each group (see on the right).

No Diet	Diet
7	3
6	7
7	3
6	5
8	4
6	6
9	8
5	6
8	7
5	7
9	7
8	7
4	8
8	

Example: two sample t-test

- H_0: $Mean_{Diet} = Mean_{No\ Diet}$
- H_a: $Mean_{Diet} \neq Mean_{No\ Diet}$
- Single sided t-test results in p-value = 0.0953.
- P-value > 0.05 so we fail to reject the null hypothesis (i.e. we do not have sufficient evidence to conclude the alternative).
- There is not sufficient evidence to suggest there is a difference in the number of sleep apnoea events in the two groups.

No Diet	Diet
7	3
6	7
7	3
6	5
8	4
6	6
9	8
5	6
8	7
5	7
9	7
8	7
4	8
8	