Statistical tests

Intro to Statistical Inference Part 1

Chuvakin Sergey

«School of Advanced Studies»

December 14, 2020

Outline

- Why do we need it?
- Statistical inference (what about sample?)
- Hypothesis
- Type of errors
- Box-plot explanation
- Compare two means
- T-test
- Degrees of freedom
- Dependent and Independent samples
- Variance check
- Normallity check

Why do we need it?

Suppose you have a question (aka research question).
There are tons of way to answer it.

⚠ But - how to do it scientifically?

Statistical inference

Statistical Inference - is a way to answer question using a data. Statistical inference - is a core of Data Driven Approuch in a business.

⚠ Helps to establish the fact of *significance* of **change** of some variable or **difference** between some variables (colud also be a relation between varaibles). Main goal is to expand inference from Sample to Population

Example: How people waste their money on insurance?

Hypothesis

Hypothesis - formal way to state a scientific question. Could and should be tested!

⚠ Research Question $\neq$ Hypothesis

Typically, a statistical hypothesis is the statement about (a) the relationship between two variables or (b) the characteristics of a distribution of a variable.

Hypothesis

All hypothesis contain two parts - **Alternative** Hypothesis and **Null** Hypothesis

- Substantive hypothesis (a.k.a. alternative; H1) is the research hypothesis, that is (typically), the statement that there is some relation between the phenomena under investigation
- Null hypothesis (H0) is the statement that there is no relation between the phenomena under investigation. Simply speaking, H0 states that H1 is false.

Hypothesis

Example

Research Question: How people waste their money on insurance?
H1: Men tend to waste more money on insurance
H0: There is no differences between men' s and women behavior
H1: People in southwest region tend to spend more momey on insurance
H0: There is no differences between people in different regions

Hypothesis

Important!

⚠ **Neither H1 nor H0 can be true or false**. Hypothesis can only be rejected.

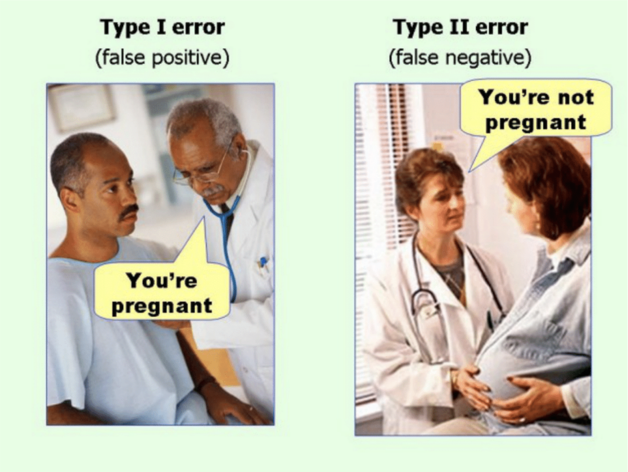
1. **True** means that a hypothesis can not be reasonably rejected given the observed data
2. **False** means that a hypothesis can be reasonably rejected given the observed data

NB: If H0 is rejected by the data, one can accept H1. However, if H1 is rejected by the data, it does not mean that one can accept H0

Type of errors

	H1 is 'true'	H1 is 'false'
Reject H0	Correct Inference	Type I error (False Positive)
Reject H1	Type II error (False Negative)	Correct Inference

Type of errors

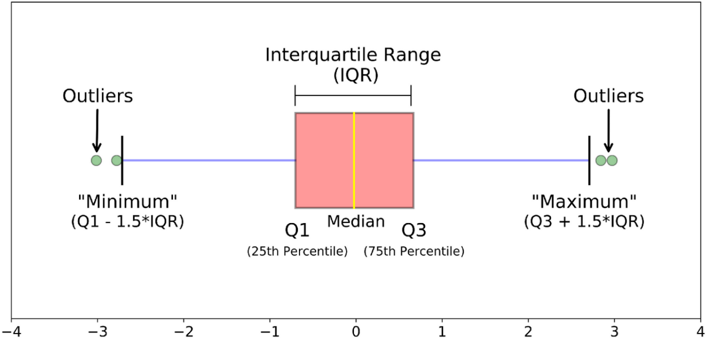


Compare two means

Box-plot explanation

Compare two means

Box-plot explanation



Different parts of a boxplot

Compare two means

What is an outlier?

Outlier - is an extra observation(s) that extremely differs from other in one variable.

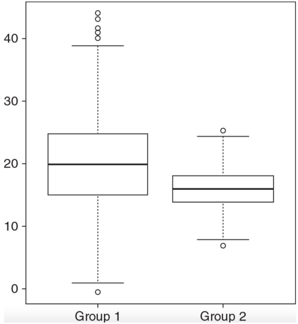
Rule of $1.5 \times IQR$

Follow here tutorial

Compare two means

T-test

Back to our first H1 - Men tend to waste more money on insurance?



Say it is real data - what can you say here?

Compare two means

T-test

Answer - almost nothing. The only way state the difference is to conduct statistical test.

T-test - statistical technic to answer the question, wheather two groups are different on some variable.

Anyway - you can notice that group different, but formal test suggest whether it *significant!*

Significance - statistical feature which states that with growing number of observation difference persists.

Compare two means

T-test

Formally:

- Null hypothesis (H0): $\mu_1 = \mu_2$
- Alternative hypothesis (H1): $\mu_1 \neq \mu_2$
- Test statistic is given by:

$$t = \frac{\hat{\mu_1} - \hat{\mu_2}}{\sqrt{\frac{\hat{\sigma_1^2}}{n_1} - \frac{\hat{\sigma_2^2}}{n_2}}}$$

  where μ_1 and μ_2 are sample (estimated) means, σ_1^2 and $\sigma_2^2$are sample (estimated) variances, and n1 and n2 are sample sizes for Groups 1 and 2.

- In t-test, there are two basic summaries of the data: test statistics and degrees of freedom

Compare two means

Afterwards

- Count Degrees of freedom
- Count T-test
- Choose Confidence level
- Look at this matrix
- Make inference

Compare two means

Thats all?

Not yet.

- How to find Degrees of freedom?
- Are groups independent?
- Are variances equal?
- Does target varaible normally distributed?

Compare two means

Degrees of freedom

Degrees of freedom (df) is a kind of measure of model complexity.

- Formally speaking, it is the number of values in the finalcalculation of a statistic that are free to vary.
- If you are a manager of a football team, you can freely determine positions of 9 out of 10 field players (if you choose sequentially).
- To put it simply,df is the difference between the number of observations (independent information peaces) you have and the number of parameters you use to estimates some test statistic of interest:

$$df = n_o - n_p$$

Compare two means

Degrees of freedom

The defailt df formula is given by the Welch–Satterthwaite equation:

$$df = \frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{\frac{s_1^2}{n_1}}{n_1 - 1} + \frac{\frac{s_2^2}{n_2}}{n_2 - 1}}$$

Basic intuition: $df = n_{row} - n_{col}$

Compare two means

Target varaible distribution

It should be continious, and at least peudo normally distributed.
You can check it in two ways:
Numerical:

- Shapiro-Wilk test. Null hypothesis: no large deviations fromthe normal distribution. If shapiro.test()results in large (i.e. insignificant) p-values, we cannot reasonably reject the null so we may safely assume that normality holds: This is theoretically incorrect (we actually test H1 of non-normality) but still the standard practice.
- Kolmogorov-Smirnov (K-S) normality test.

Compare two means

Target varaible distribution

Graphical:

- density plots/histograms: is the empirical density of Y close to the bell-shape curve? Not very useful with extremely small samples.
- QQ-plots (QQ forquantile-quantile): does individual observations are close enough to the 45-degree line?

Compare two means

Homogeneity of variances

Homogeneity of variance means that we assume that two populations under comparison (from which we sample comparison groups) may differ in their means but not variances.

We can check variances using F-test. In R - var.test(). Alternatives bartlett.test(), leveneTest() (car package), fligner.test()

Compare two means

Dependents Samples

- The standard t-test assumes that different individuals are randomly assigned to one of two conditions, so their Y scores are independent (i.e a score of ani-th individual is not influenced by a score of an i-th individual: no spillover effects)
- Dependent (paired) samples:
  - Same individuals sequentially exposed to two different conditions
  - There are many pairs consisting of two very similar (identical or matched) individuals. In each pair, one individual is(randomly) assigned to one condition and the other assigned to another condition (e.g., experiments with twins).

Compare two means

Dependents Samples

Test statistic is computed in a different way:

$$t = \frac{\hat{D} - \mu_0}{\frac{\sigma_D}{\sqrt{n_D}}}$$

where D is the sample average within-pair difference,μ_0 is some constant (typically 0, because the default H_0: D = 0; read about one-sample t-test for details),σ_Di s the estimatedstandard deviation of within pair differences, and n_D is thenumber of pairs

R implementation: set the paired argument of t.test() to True

Compare two means

what can be done

Robustness is a property of a statistical test meaning that the test can return correct results even if one or some of it's assumptions are not perfectly fitted

Homogeneity of variance: defaultt.test() settings correct for deviations from this assumption.

Compare two means

what can be done

Non-normality is more problematic:

- Non-normal data: use non-parametric tests, e.g. Wilcoxon test(non-parametricmeans that various distributional parameters,e.g. means and variances, are not used in the test statistic computation)
- Notice that the Wilcoxon signed-rank test (a.k.aMann–Whitney U test test) does not acutally compare means.
- Influential observations (with extreme values on Y): trimming(perfrom test keeping some proportion of extreme obs out),e.g. Yuen's test.