Simple Statistical Fallacies

Lesson 10

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Overview of Statistics

Statistics is a branch of Math, geared towards more hands-on aspects of the field

Many Stats methods are applicable and useful in data science

Two main kinds of Stats:

- Descriptive
- Inferential

In data science we make use of both

- Stats that are focused on the mathematical properties of distributions and on theorems are not encountered in data science

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Statistical Fallacies

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

STATISTICAL FALLACIES

Fallacy: a conclusion that is not valid for the population it is related to

Different kinds out there. Most important ones:

- Type I and type II Errors
- Correlation - Causation relationship
- Overestimating the importance of the average (mean) value
- Confusing statistical significance with importance

Fallacies are not always easy to spot as they are methodological errors

- Fallacies can easily break a data science project, or delay its process significantly

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Type I and type II errors

Related to how we deal with the Null Hypothesis (what we are trying to disprove)

Similar to a binary classification problem

- Type I error: if the Null Hypothesis is fine (i.e. it cannot be rejected with the data at hand) but we reject it anyway
  - Whether something is a type I error depends on the significance threshold we use (α)
- Type II error: if the Null Hypothesis is false (i.e. it can be rejected with the data at hand) but we don't reject it
  - Whether something is a type II error depends on the power of the test (1 - β)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Correlation and Causation

Two variables can be highly correlated but the phenomena they represent may not have a cause-effect relationship to each other

- Two phenomena may have a cause-effect relationship but the corresponding variables may not have a strong correlation

Proving that something is the cause of something else = a very challenging problem, involving many experiments

- The relationship between 2 variables may not be linear (and not be reflected in the Stats correlation metrics)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Overestimating the importance of the average (mean) value

The mean is one of many central tendency metrics

- Mean only captures part of the whole picture. Oftentimes, we use other metrics too (e.g. median, mode, weighted mean, etc.)

Metrics of central tendency by themselves don't give a clear view of a variable => need for measures of spread (e.g. variance, standard deviation, range, etc.)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Overestimating the importance of other statistical metrics

Variance by itself is also very misleading

- Every other measure of spread (e.g. range) can be unreliable, when used by itself

All conventional statistical metrics are influenced by outliers

- Even if you used metrics from all metric families, to describe a variable, you still wouldn't see the whole picture

The best way to understand a variable is through a histogram or a box plot

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Confusing statistical significance with importance

- Statistical significance in a test = confidence in rejecting the Null Hypothesis
- Importance of a test result = how useful it is as a finding or as a feature to be used in a model
  - Something can be important without being statistically significant
  - Something can be statistically significant without being particularly important
- Need to exercise judgment in order to assess importance in a test result

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Some considerations about statistical fallacies

- There are several other statistical fallacies
- They have to do with how we use Stats and how we interpret the results of a statistical process
- Fallacies are serious methodological errors and we need to be careful about them in every data science project
- Having a solid understanding of the mindset behind statistical analysis is a good deterrent of various statistical fallacies

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

The importance of stating your assumptions

Assumptions are often essential in data science projects

- Most statistical methods involve a number of assumptions => it's easy to forget them

Assumption often influence the result noticeably

- Other people may not be aware of the underlining assumptions of a data analytics process

Adhering to the assumptions is bound to lead to better interpretation of the results and avoidance of statistical fallacies

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Normal Distribution

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

NORMAL DISTRIBUTION & P-VALUE

Understanding distributions, esp. the Normal distribution, is very useful

Distributions are everywhere so being able to treat data according to the distribution it follows, can:

- save us time
- enable better understanding of the data

P-value = probability value of an event to occur within a given range in a distribution

- P-values are useful in many data science processes

Distributions often encountered

Two types of distributions in general:

- Continuous variables
- Discreet variables

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Distributions for continuous variables:

- Normal (aka Gaussian)
- Poisson
- Zipfian (power law)
- Weibull
- Uniform (less common)

Distributions for discreet variables:

- Binomial
- Chi-squared

Normal distribution summary

Variables characterized by it are anywhere in the Real number spectrum (i.e. -Inf to Inf)

- Mean value (μ) = 0.0
- Standard Deviation (σ) = 1.0

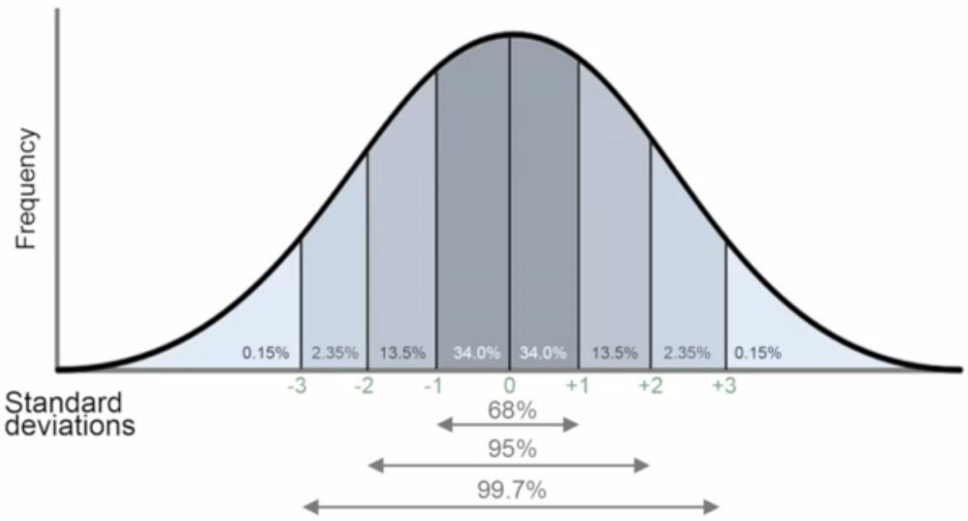
For this distribution only, Median = Mode = Mean

Most data points are closer to the mean with only very few of them being far from it, on either side

- Shaped like a bell (Bell Curve)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Normal distribution plot



Source: www.magoosh.com

P-Values

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

P-values

P-value = Probability of a variable being in a particular part of the distribution

- P-values are calculated using a reference table or a function in a programming language

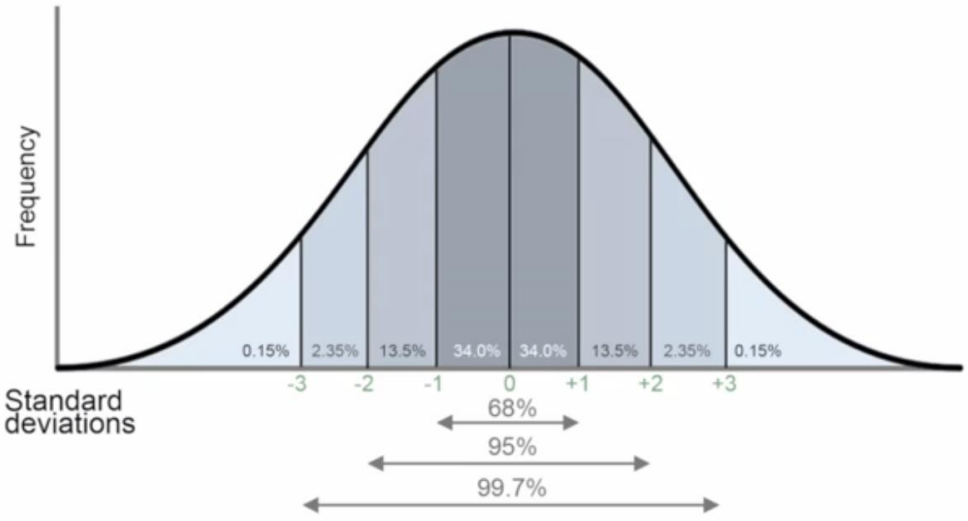
P-values are related to confidence scores in predictive models (e.g. a classifier)

- Usually they refer to sections like this: P(x > 5), or P(A) where A = values of x between -2 and 3, inclusive

All P-values are between 0 and 1, inclusive

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Normal distribution plot revisited



Source: www.magoosh.com

Central Limit Theorem

A fancy sounding name for an observation someone made and proved, about the distribution of the means of the samples of any distribution out there

- In a nutshell: the means of the samples follow a normal distribution, regardless of the shape of the original distribution

Very useful since this enables us to perform statistical analysis on various samples and also draw more reliable conclusions based on these samples

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Summary

- It's easy to be mislead by statistics
  - Correlation does not equal causation
  - Pay attention to significance and power with the Null Hypothesis
- Normal Distribution is always a bell shape
- P-value is the confidence of the classifier
- The means of a sample are always a Normal distribution.