statsElements Documentation

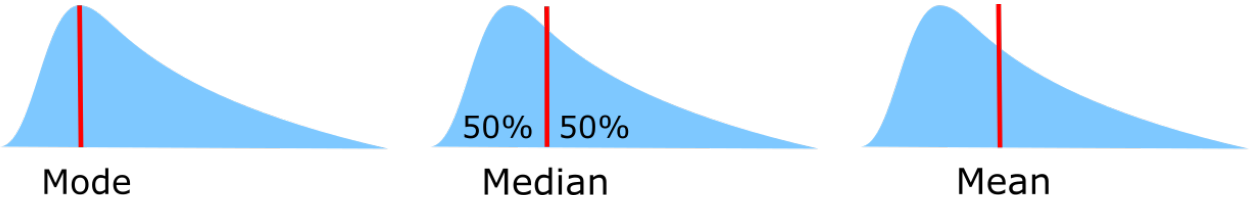
remy wahnoun

Nov 09, 2018

Contents:

CHAPTER 1

Measuring Central Tendency



Mode Most common value

Median Central Value (less sensitive to outliers)

Mean Sum observations / number of observations

CHAPTER 2

Measuring Variability

Range Largest observation – smallest observation

Quantiles Split data like into equally numbered groups. Median into two, quartiles into 4

Interquartile Range Range between top and bottom quartile. Shows where the middle 50% of the data lies. Not influenced by outliers

Standard Deviation Average deviation from the mean. Measures homogeneity of individual values.

$$std = \sqrt{\frac{\sum (x_i - x_{mean})^2}{n-1}}$$

CHAPTER 3

Distribution

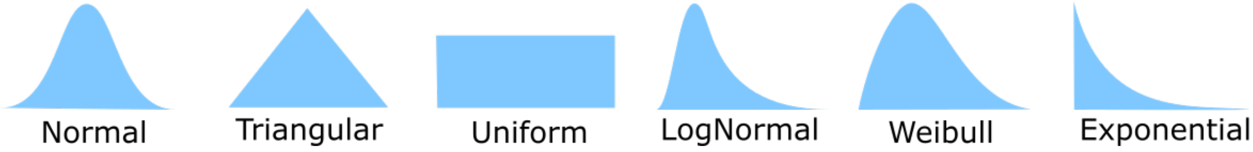


Fig. 1: Some distributions

Normal (=Gaussian) distribution Most common, unimodal, symmetrical. Other distributions tend to normalize when we increase sample size. Entirely defined by two parameters: means and std.

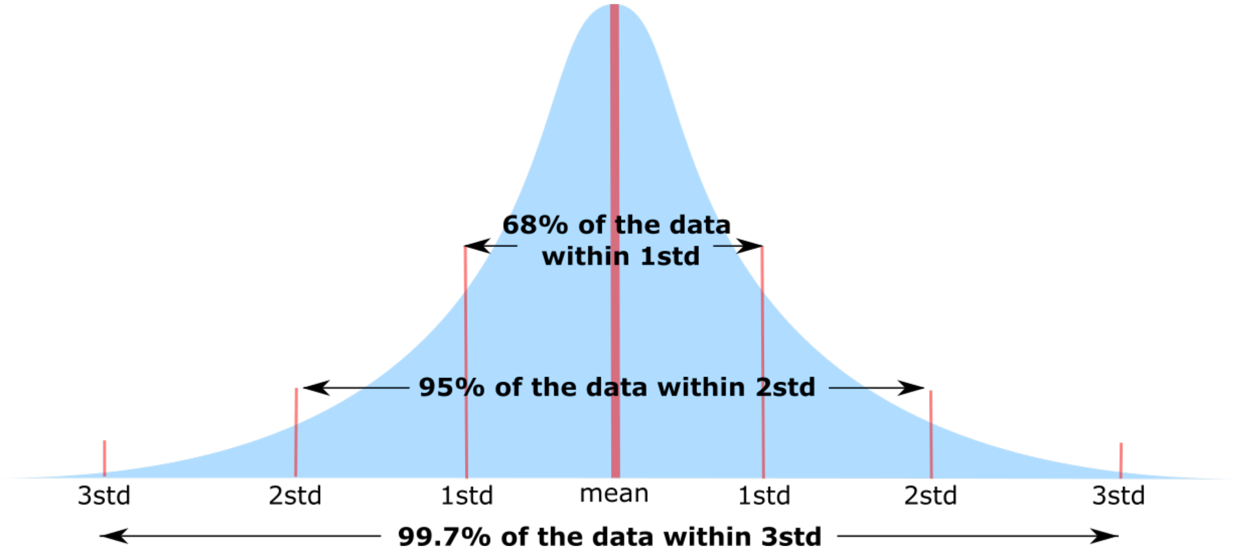


Fig. 2: Normal Distribution

See the *Statistical Tests* section for info on testing or visualizing a distribution vs another.

Law Of Large Numbers As a sample size grows, its mean will get closer and closer to the average of the whole population.

Central Limit Theorem In probability theory, the central limit theorem (CLT) establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed. The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.

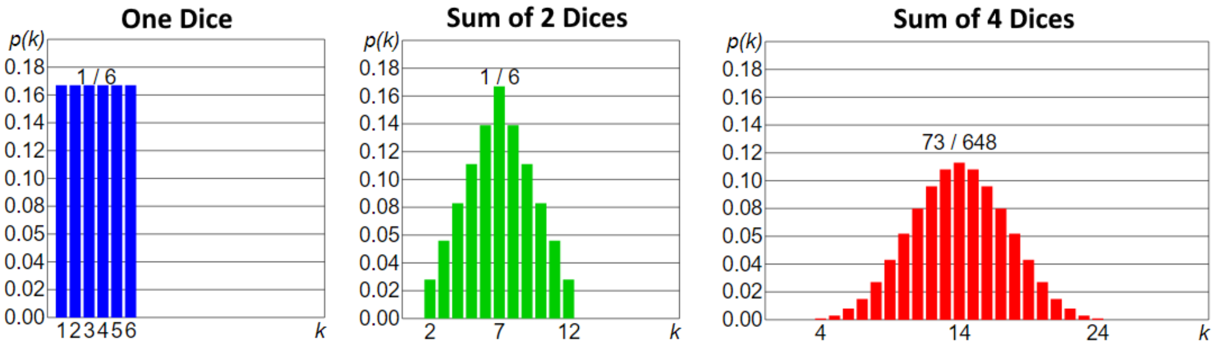


Fig. 3: Central Limit Theorem (wiki)

Standard Error Standard deviation of the sampling distribution of a statistic, most commonly of the mean. It can be seen as how far the sample mean is likely to be from the population mean

$$SE = \frac{std}{\sqrt{n}}$$

95% Confidence Interval For a gaussian distribution, the range in which 95% of the true population mean is likely to lie is defined by:

$$CI = [mean - 1.96 * \frac{std}{\sqrt{n}}, mean + 1.96 * \frac{std}{\sqrt{n}}]$$

1.96 can be replaced by other values for different percentages: 99%:2.576, 98%:2.326, 95%:1.96, 90%:1.645

May not be good for small sample size (<30) and very non normal distributions. In that case we can use the t-distribution to replace the 1.96

CHAPTER 4

Correlation

See *Statistical Tests* to choose the appropriate method. TODO ADD MEAT TO CORRELATION

Per Cohen (1992, Power primer): 0.0 < abs(corr) < 0.3: Weak 0.3 < abs(corr) <= 0.5: Moderate 0.5 < abs(corr) <= 0.9: Strong 0.9 < abs(corr) <= 1.0: Very strong

We can use a scatter plot to visualize correlation.

CHAPTER 5

Statistical Inference

Planning carefully the way we will analyze data is very important to obtain results that we can trust.

Note: In general, hypothesis testing requires the following steps:

- Clearly define the problem and the hypotheses, and the type of data that will be analyzed.
- Selecting an appropriate test and checking its assumptions.
- If planning a study or QTP (a-priori), estimate the minimum number of samples required to obtain significance. If data is already available (post-hoc), we can estimate the power of the test on the provided data.
- Run the test and conclude.

5.1 1. Define the problem

The first step is to clearly define what we want to test and how we want to test it. This step is crucial as everything else will depend on it.

Population This is the total set of observations that can be made. For example if we want to know the average weight of humans, this is the average of the weight of every human on earth.

Sample This is the set of collected data. In this example, this is the weights of a small group of randomly selected people.

We want to infer information on the population based on a selected sample.

We then describe the data that will be used. There are four basic data types:

	Scale		Categorical	
	Continuous	**Discrete**	**Ordinal**	**Nominal**
Data	takes any value	integers	obvious order	unordered
Example	height	number of children	low,medium,high	red,green,blue

5.2 2. State the Hypotheses

Then we state the Null (Ho) and the Alternative Hypothesis (Ha).

Null Hypothesis (Ho) An hypothesis associated with a contradiction to a theory we want to prove

Alternative Hypothesis (Ha) An hypothesis associated with a theory we want to prove

5.3 3. Select the appropriate statistical test

We then need to choose a statistical test.

5.3.1 a. Tails

One-sided test We want to test if a parameter is inferior to a reference value or we want to test if it is superior to a reference value.

Two-sided test We want to test if a parameter doesn't equal a reference value

5.3.2 b. Parametric vs Nonparametric test

Nonparametric tests do not assume that data follow a normal distribution. We can use parametric tests on non normal data, but the statistical power of the results will be reduced So we use a parametric test when: - Data are normally distributed - Sample size is large enough to satisfy the central limit theorem

Note:

Central Limit Theorem Given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution.

Parametric Tests

- Perform well with skewed and non normal distributions if the sample size is large enough
- Perform well when the spread of each group is different (not always the case on nonparametric)
- Stronger statistical power
- For ordinal variables with seven or more categories with normal distributions it is normally advised to use parametric tests.

Nonparametric Tests

- Chosen when median represents the population better than the mean (e.g. many outliers)
- Chosen when we have a small sample size
- Chosen when we have ordinal data, ranked data, or outliers that we can't remove

5.3.3 c. Choose the right test

This will be described in the *Statistical Tests* section

5.4 4. Check the tests assumptions

Most tests make specific assumptions on the data. Some are more sensitive to deviation from their assumptions than others. See the *Statistical Tests* section for info on specific tests. This section also shows how to visualize our data distribution against for example a normal distribution using a QQ-plot, or to test its fit. Examples of assumptions:

> Independence of observations from each other Independence of observational error from confounding effect Normality of observations

5.5 5. Power Analysis and Statistical Power

5.5.1 a. State the desired α and β

	Ho is True	Ha is True
Accept Ho	good	Type II Error
Reject Ho	Type I Error	good

α, Probability of Type I error. This is the error when the test rejects Ho while it is actually true.

β, Probability of Type II error This is the probability of not rejecting Ho when Ho is actually false.

Power, 1-β Probability of correctly rejecting a False hypothesis.

Warning:

- When a test outcome is not significant, it doesnt mean that Ho is True, the test is inconclusive.
- If several concurrent tests are performed, consider a Bonferroni correction (i.e. divide the significance level by the number of concurrent tests)

5.5.2 b. Establish the Effect Size

The effect size is a measure of the strength of the effect of an independent variable on a dependant variable. It helps assess whether a statistically significant result is meaningful. We can use for example the *G*Power: Statistical Power Analyses* software to calculate effect size. For reference, <0.3 is often seen as a small effect, 0.5 seen as medium and >.8 as large (Cohen).

5.5.3 c. Create Sampling Plan, determine sample size

When the data is not yet available, for example when we are preparing a clinical study, we want to estimate how many samples (or subjects) we need to obtain significant results. This is the hardest part as it often requires prior knowledge on the results. This can come from a preliminary study, or from the literature.

5.6 5. Run the test

Now we need to run the chosen test, estimate the test statistic, determine the p-value and conclude

Test Statistic Value calculated from a sample often to summarize the sample

P-value

- Smallest level of significance that would lead to a rejection of Ho with the given data.
- Probability of wrongly rejecting Ho. Small p-value indicates strong evidence against Ho

If p-Value is < than alpha-risk, reject Ho and accept Ha

If p-Value is > than alpha-risk, fail to reject the Null, Ho

CHAPTER 6

Statistical Tests

6.1 Test for Normality

To get an indication on the shape of a distribution, or to compare to a given distribution, we can plot a histogram or a QQ-plot (quantile-quantile). When plotting a QQ-plot against a normal distribution, if all the samples fall close to the reference line, we can assume normality.

<ADD STATISTICAL TESTS FOR NORMALITY FROM ONENOTE>

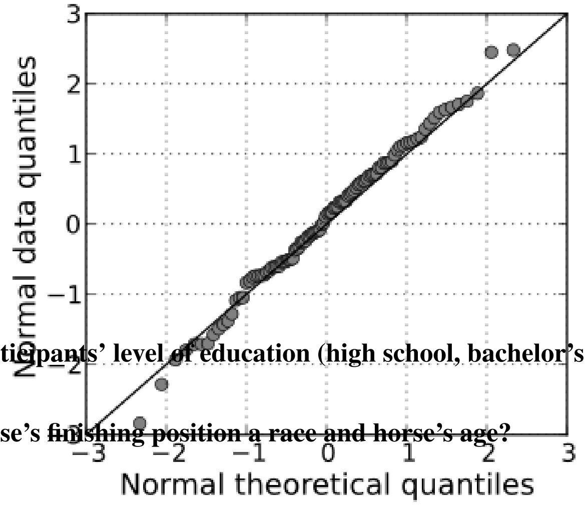


Fig. 1: src:wikipedia/Q–Q_plot
QQ plot of random normal data against a normal distribution

6.2 Examples of Test Selection

Is there a statistically significant relationship between participants' level of education (high school, bachelor's, or graduate degr

Spearman

Is there a statistically significant relationship between horse's finishing position a race and horse's age?

Spearman

CHAPTER 7

Examples of Statistical Inference

7.1 Influence of Teacher Reputation on Rating

From onlinestatbook

Questions:

- Does an instructor's prior reputation affect student rating?
- Does the size of this effect depend on student characteristics?

Experimental Design:

- subjects viewed video of a lecture of a teacher after reading an evaluation of the instructor, then they rated the instructor.
- subjects where randomly assigned two conditions: Charismatic (reading a good evaluation), Punitive (reading a bad review of the instructor)

Descriptive Statistics:

- Boxplot showing the student rating vs the two conditions. Ratings seem higher for the charismatic teacher.
- N, Mean, Median, Skewness, Kurtosis etc are calculated for both conditions

Inferential Statistics:

- Independent samples t-test used to test the differences. The result is significant, supporting the conclusion that instructor reputation affects ratings.
- **Assumptions:**
  - Each score is sampled independently and randomly: Ok, as the students are randomly assigned to a condition.
  - Normal distribution of the scores within each condition: Violated to a moderate degree because of the skewness.

  They assessed that this was not important using the data analysis lab. - Equality of variance between the two populations: Ok

7.2 Mediterranean Diet and Health

From onlinestatbook

Question:

- Is a mediterranean diet healthier than a diet with high-saturated fat?

Experimental Design:

- 605 survivors of a heart attack assigned to either the AHA diet or the Mediterranean diet
- Over a 4 year period, patients following the Mediterranean diet were seen initially, then after two months, then once a year to check observance.
- The other group was assumed to follow the diet.
- Information was collected on number of deaths from cardiovascular causes, non fatal heart-related episodes and tumors.

Descriptive Statistics:

- Histogram, frequencies tables. 20% of the AHA diet patients had at least one illness, compared to 10% on the Mediterranean.

Inferential Statistics:

- A Chi-Square test can be used to check if there is a relationship between diet and outcome..
- Conclusion that outcome is related to diet and that Mediterranean diet is superior to the AHA diet.

7.3 Who is buying iMacs

From onlinestatbook

Question:

- Are the buyers of the latest Mac new buyers, or did they previously have a Mac product

Experimental design:

- They asked 500 of the new Mac purchasers if they owned or had owned a Mac

Results:

- 83 new computer owners, 60 who had a Windows computer, 357 who had owned a Mac
- Proportion of first time computer owners = 0.167.
- The 95% confidence interval on the proportion is calculated (0.13<CI<.20) therefore, it is likely that between 13% and 20% of new Mac buyers are first time computer owners.

Assumptions:

- No reason seen that would violate the assumptions of normality or independence.

Epilog:

- After one year, Apple reports that 1/3 of new buyers are first time computer buyers. This is outside the CI range. Is this a sampling error ? Other factors?

CHAPTER 8

Examples of Study Design

http://onlinestatbook.com/case_studies_rvls/index.html

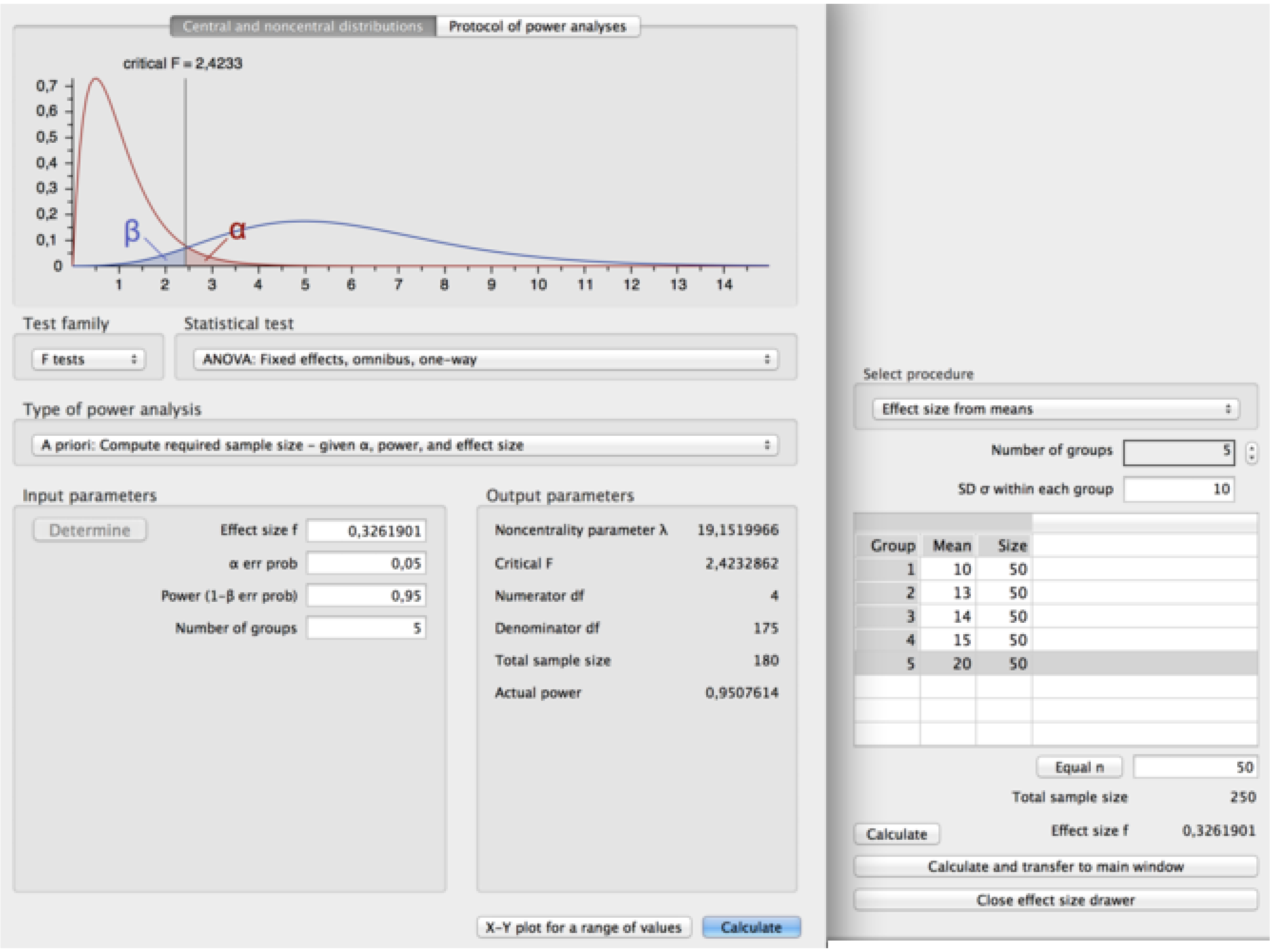
CHAPTER 9

Useful Tools

9.1 G*Power: Statistical Power Analyses

G*Power is a tool to compute statistical power analyses for many different t tests, F tests, $\chi 2$ tests, z tests and some exact tests. G*Power can also be used to compute effect sizes and to display graphically the results of power analyses.

http://www.gpower.hhu.de/


Central and noncentral distributions
Protocol of power analyses
critical F = 2,4233
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
β
α
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Determine
X-Y plot for a range of values
Calculate
Calculate
Calculate and transfer to main window
Close effect size drawer
Equal n

Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input parameters

Effect size f	0,3261901
α err prob	0,05
Power (1-β err prob)	0,95
Number of groups	5

Output parameters

Noncentrality parameter λ	19,1519966
Critical F	2,4232862
Numerator df	4
Denominator df	175
Total sample size	180
Actual power	0,9507614

Select procedure: Effect size from means

Number of groups: 5

SD σ within each group: 10

Group	Mean	Size
1	10	50
2	13	50
3	14	50
4	15	50
5	20	50

Equal n: 50

Total sample size: 250

Effect size f: 0,3261901

CHAPTER 10

References

Basics:

- https://en.wikipedia.org/wiki/Statistical_hypothesis_testing

Detailed:

- http://onlinestatbook.com

Examples of Experimental Design:

- http://onlinestatbook.com/case_studies_rvls/index.html

CHAPTER 11

Indices and tables

- genindex
- modindex
- search

e mara