Human Data Science

Lab 1

Luca Casini

luca.casini7@unibo.it

Introduction

This lesson be a recap of statistics and probability with a practical approach.

References:

1. **Practical Statistics for Data Scientists, 2nd Edition.** Peter Bruce, Andrew Bruce, Peter Gedeck, 2020. O'Reilly Media
2. **All of Statistics.** Larry Wasserman, 2004. Springer

[1] is the basis for this lesson, [2] contains all the mathematics and theory in case you want to dive deeper.

We will discuss tools and libraries for data science in **Python**

Software

For data science and machine learning the standard is using either **Python or R**

We will focus on Python but for your project feel free to use whichever you prefer

In either case I suggest you download **Anaconda**, a software distribution that makes managing the development environment easier

We will mention a number of libraries, the most important are:

- **Pandas** for data management
- **Numpy**, **scipy** and **statsmodels** for the procedures we will use
- **Matplotlib** and **seaborn** for visualization

Code examples may be shown during the lesson depending on timing. Every resource will be uploaded on the course website

1. Exploratory Data Analysis

Data

Numeric Data that are expressed on a numeric scale.

- **Continuous** Data that can take on any value in an interval. (Synonyms: interval, float,numeric)
- **Discrete** Data that can take on only integer values, such as counts. (Synonyms: integer,count)

Categorical Data that can take on only a specific set of values representing a set of possible categories. (Synonyms: enums, enumerated, factors, nominal)

- **Binary** A special case of categorical data with just two categories of values, e.g., 0/1,true/false. (Synonyms: dichotomous, logical, indicator, boolean)
- **Ordinal** Categorical data that has an explicit ordering. (Synonym: ordered factor)

Estimates of Location

Mean The sum of all values divided by the number of values. Synonym average

Weighted mean The sum of all values times a weight divided by the sum of the weights. Synonym weighted average

Median The value such that one-half of the data lies above and below.Synonym 50th percentile

Percentile The value such that P percent of the data lies below. Synonym quantile

Weighted median The value such that one-half of the sum of the weights lies above and below the sorted data.

Trimmed mean The average of all values after dropping a fixed number of extreme values. Synonym truncated mean

Robust Not sensitive to extreme values. Synonym resistant

Outlier A data value that is very different from most of the data.Synonym extreme value

Estimates of Variability

Deviations The difference between the observed values and the estimate of location. Synonyms errors, residuals

Variance The sum of squared deviations from the mean divided by n – 1 where n is the number of data values. Synonym mean-squared-error

Standard deviation The square root of the variance.

Mean absolute deviation The mean of the absolute values of the deviations from the mean. Synonyms l1-norm, Manhattan norm

Median absolute deviation from the median The median of the absolute values of the deviations from the median.

Estimates of Variability (2)

Range The difference between the largest and the smallest value in a data set.

Order statistics Metrics based on the data values sorted from smallest to biggest. Synonym ranks

Percentile The value such that P percent of the values take on this value or less and (100–P) percent take on this value or more. Synonym quantile

Interquartile range The difference between the 75th percentile and the 25th percentile. Synonym IQR

Exploring the Data Distribution

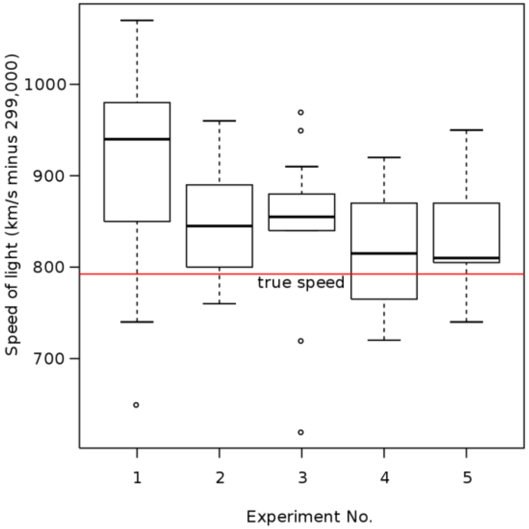
Boxplot A plot introduced by Tukey as a quick way to visualize the distribution of data.

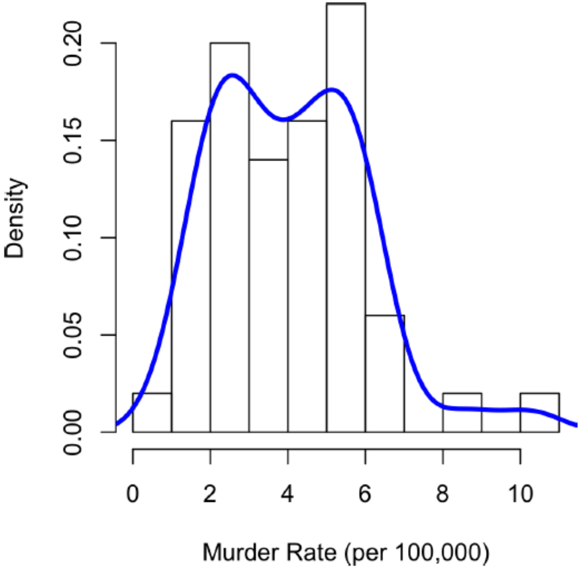
Frequency table A tally of the count of numeric data values that fall into a set of intervals (bins).

Histogram A plot of the frequency table with the bins on the x-axis and the count (or pro-portion) on the y-axis. While visually similar, bar charts should not be confused with histograms.

Density plot A smoothed version of the histogram, often based on a kernel density estimate.

Boxplot, Histogram and Density





Exploring Binary and Categorical Data

Mode The most commonly occurring category or value in a data set.

Expected value When the categories can be associated with a numeric value, this gives an average value based on a category’s probability of occurrence.

Bar charts The frequency or proportion for each category plotted as bars.

Pie charts The frequency or proportion for each category plotted as wedges in a pie.

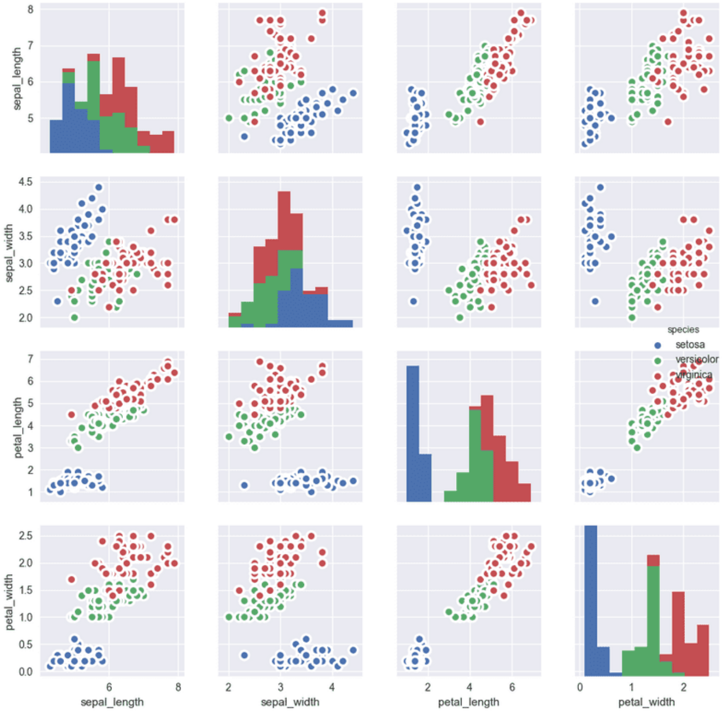
Correlation

Correlation coefficient A metric that measures the extent to which numeric variables are associated with one another (ranges from –1 to +1).

Correlation matrix A table where the variables are shown on both rows and columns, and the cellvalues are the correlations between the variables.

Scatterplot A plot in which the x-axis is the value of one variable, and the y-axis the value of another.

Correlation



Exploring Two or More Variables

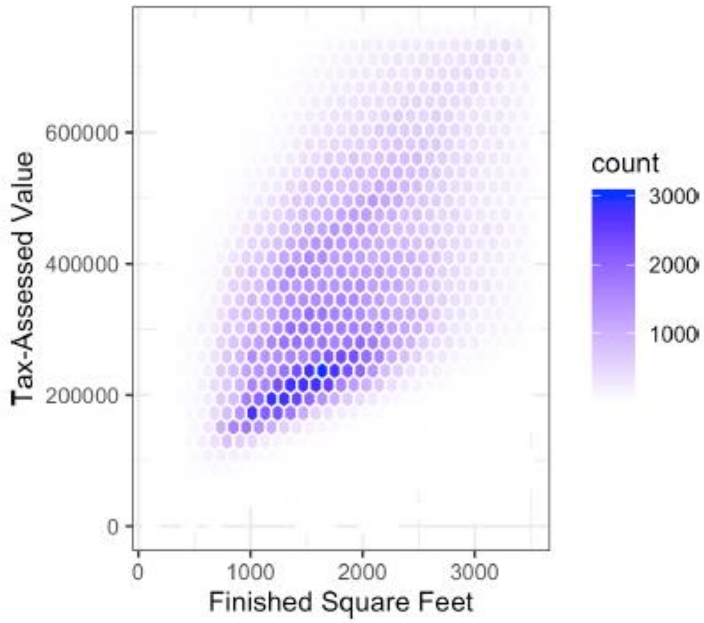
Contingency table A tally of counts between two or more categorical variables.

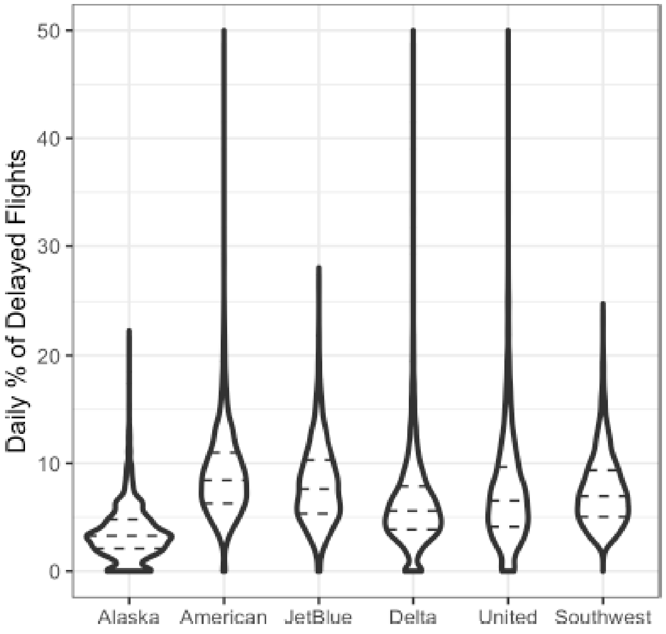
Hexagonal binning A plot of two numeric variables with the records binned into hexagons.

Contour plot A plot showing the density of two numeric variables like a topographical map.

Violin plot Similar to a boxplot but showing the density estimate.

Hexagonal binning & violin plot





2. Data and Sampling Distributions

Random Sampling

Sample A subset from a larger data set.

Population The larger data set or idea of a data set.N (n)The size of the population (sample).

Random sampling Drawing elements into a sample at random.

Stratified sampling Dividing the population into strata and randomly sampling from each strata.

Bias Systematic error.

Sample bias A sample that misrepresents the population.

Selection Bias

Selection bias Bias resulting from the way in which observations are selected.

Data snooping Extensive hunting through data in search of something interesting.

Vast search effect Bias or nonreproducibility resulting from repeated data modeling, or modeling data with large numbers of predictor variables.

Regression to the mean refers to a phenomenon involving successive measurements on a given variable: extreme observations tend to be followed by more central ones.

Sampling Distribution of a Statistic

The term sampling distribution of a statistic refers to the distribution of some sample statistic over many samples drawn from the same population. Much of classical statistics is concerned with making inferences from (small) samples to (very large) populations.

Sample statistic A metric calculated for a sample of data drawn from a larger population.

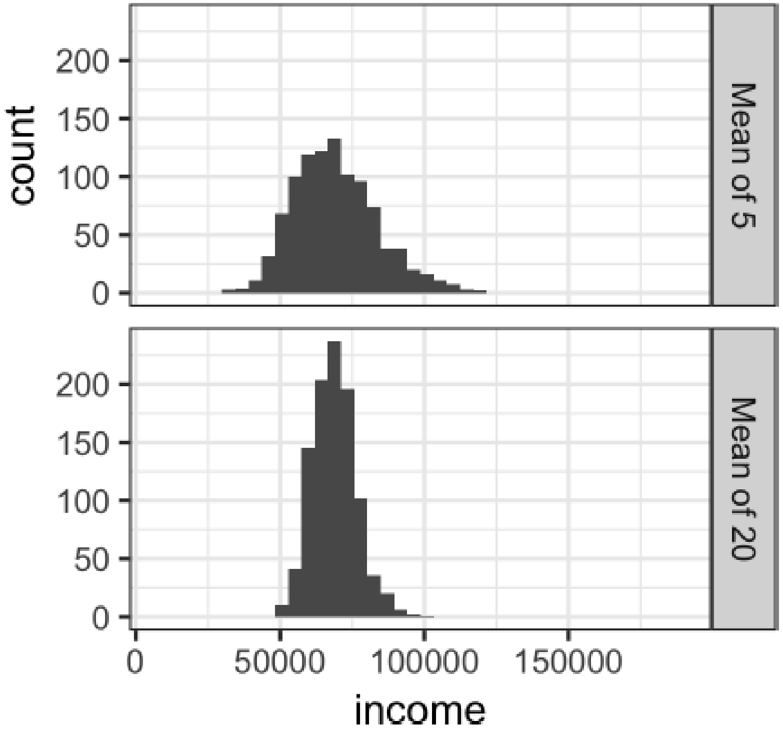
Data distribution The frequency distribution of individual values in a data set.

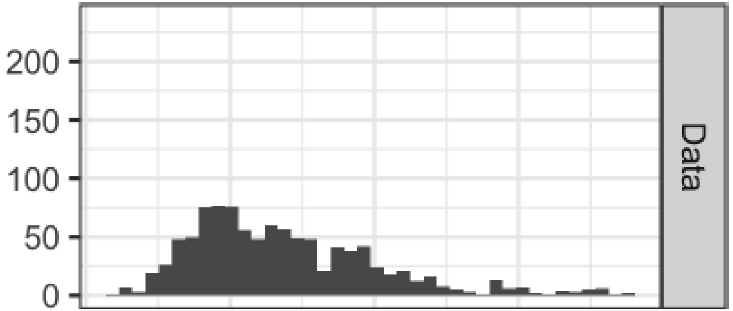
Sampling distribution The frequency distribution of a sample statistic over many samples or resamples.

Central limit theorem The tendency of the sampling distribution to take on a normal shape as sample size rises.

Standard error The variability (standard deviation) of a sample statistic over many samples
(not to be confused with standard deviation, which by itself, refers to variability of individual data values).

Sampling Distribution of a Statistic (2)





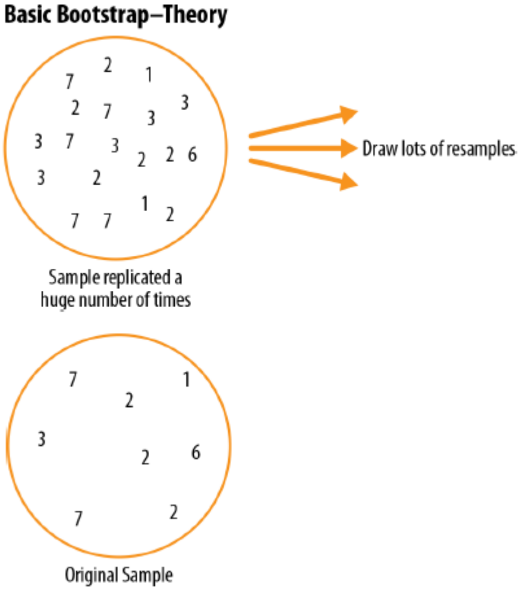
Histogram of annual incomes of 1,000 loan applicants (top), then 1,000 means of n=5 applicants (middle), and finally 1,000 means of n=20 applicants (bottom)

Bootstrap

The bootstrap (sampling with replacement from a data set) is a powerful tool for assessing the variability of a sample statistic.

The bootstrap can be applied in similar fashion in a wide variety of circumstances, without extensive study of mathematical approximations to sampling distributions.

It also allows us to estimate sampling distributions for statistics where no mathematical approximation has been developed.



Bootstrap (2)

1. Draw a sample value, record it, and then replace it.
2. Repeat n times.
3. Record the mean of the n resampled values.
4. Repeat steps 1–3 R times.
5. Use the R results to:
   a. Calculate their standard deviation
      (this estimates sample mean standard error).
   b. Produce a histogram or boxplot.
   c. Find a confidence interval.

Confidence Intervals

Confidence level The percentage of confidence intervals, constructed in the same way from the same population, that are expected to contain the statistic of interest.

Interval endpoints The top and bottom of the confidence interval.

CI can be easily computed using bootstrap without assumptions of the distribution

Historically, beacuse of CLT, they were computed using t-statistics with n-1 deg of freedom

$$\bar{x} \pm t_{n-1}(0.05)\frac{s}{\sqrt{n}}$$

Where n is the sample size, and s the sample std.dev

This will be discussed when we look at hypothesis testing

Normal Distribution

The bell-shaped normal distribution is iconic in traditional statistics. The fact that distributions of sample statistics are often normally shaped has made it a powerful tool in the development of mathematical formulas that approximate those distributions.

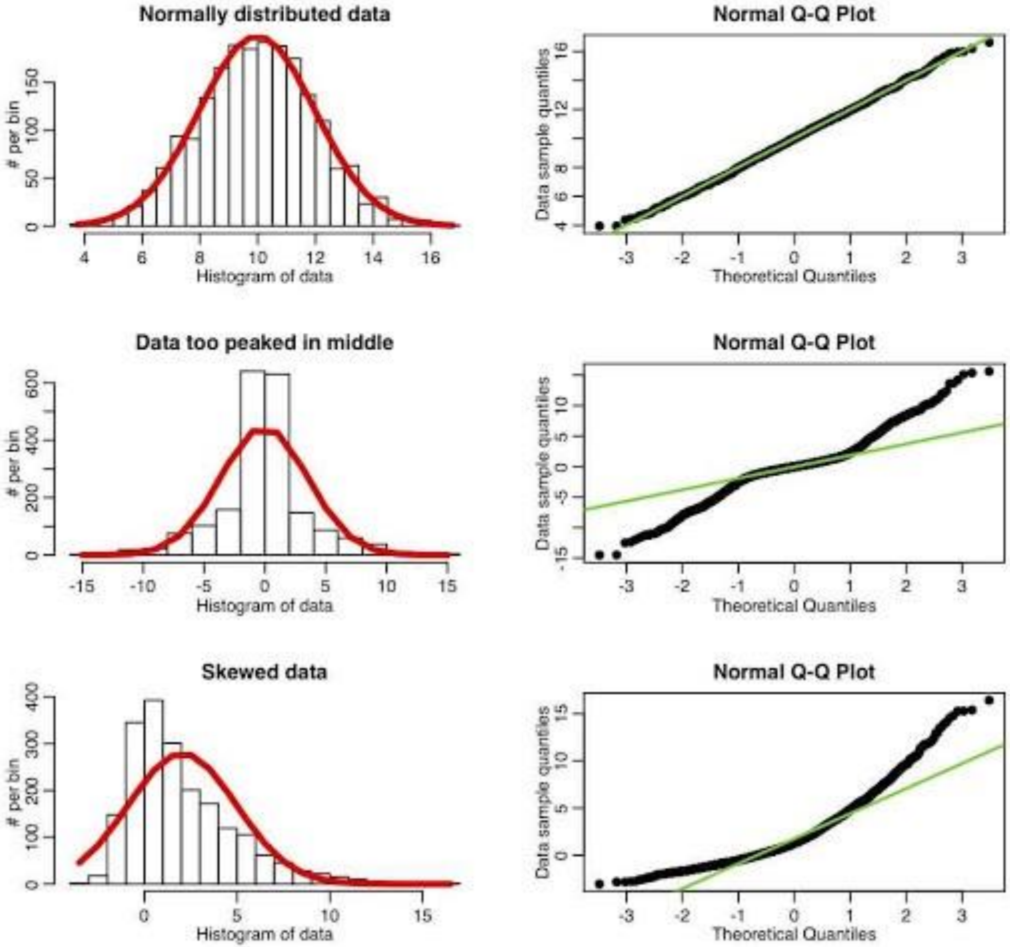
Error The difference between a data point and a predicted or average value.

Standardize Subtract the mean and divide by the standard deviation.

z-score The result of standardizing an individual data point.

Standard normal A normal distribution with mean = 0 and standard deviation = 1.

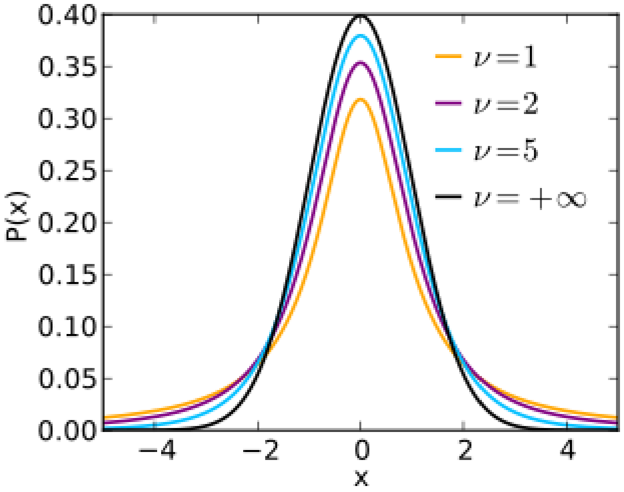
QQ-Plot A plot to visualize how close a sample distribution is to a specified distribution,e.g., the normal distribution.


Normally distributed data
# per bin
0
50
100
150
4
6
8
10
12
14
16
Histogram of data
Normal Q-Q Plot
Data sample quantiles
4
6
8
12
16
-3
-2
-1
0
1
2
3
Theoretical Quantiles
Data too peaked in middle
# per bin
0
200
400
600
-15
-10
-5
0
5
10
15
Histogram of data
Normal Q-Q Plot
Data sample quantiles
-15
-5
0
5
10
-3
-2
-1
0
1
2
3
Theoretical Quantiles
Skewed data
# per bin
0
100
200
300
400
0
5
10
15
Histogram of data
Normal Q-Q Plot
Data sample quantiles
0
5
10
15
-3
-2
-1
0
1
2
3
Theoretical Quantiles

Student’s t-Distribution

The t-distribution is a normally shaped distribution, except that it is a bit thicker and longer on the tails. It is used extensively in depicting distributions of sample statistics.

Is used in hypothesis testing and confidence intervals to get critical values from a table depending on degrees of freedom

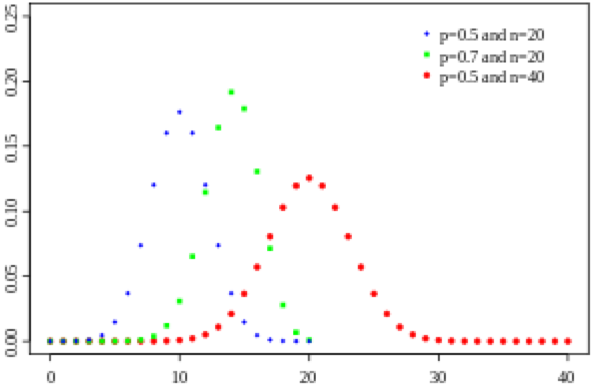


Binomial Distribution

Yes/no (binomial) outcomes lie at the heart of analytics since they are often the culmination of a decision or other process; buy/don’t buy, click/don’t click, survive/die, and so on.

Central to understanding the binomial distribution is the idea of a set of trials, each trial having two possible outcomes with definite probabilities.

E.g. Coin flipping

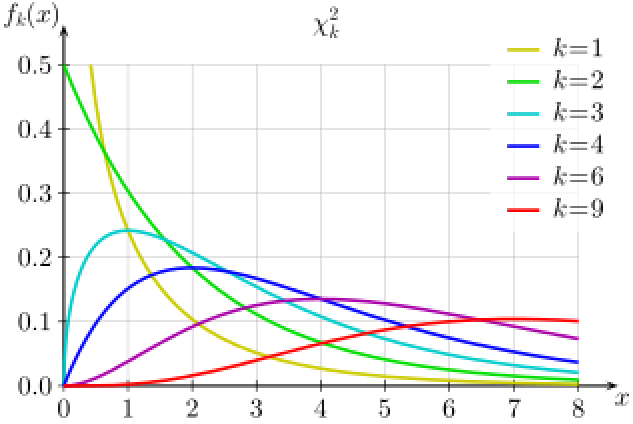


Chi-Square Distribution

The chi-square statistic is a measure of the extent to which a set of observed values “fits” a specified distribution (a “goodness-of-fit” test).

It is useful for determining whether multiple treatments (an “A/B/C… test”) differ from one another in their effects.

The chi-square distribution is the distribution of this statistic under repeated resampled draws from the null model



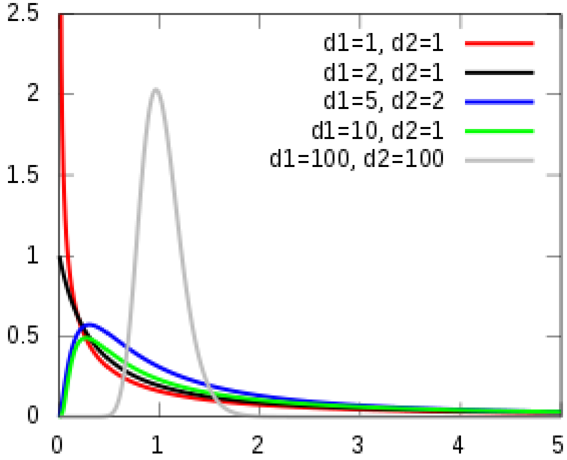
F-Distribution

A common procedure in scientific experimentation is to test multiple treatments across groups.

This is similar to test referred to in the chi-square distribution, except we are dealing with measured continuous values rather than counts.

In this case we are interested in the extent to which differences among group means are greater than we might expect under normal random variation.

The F-statistic measures this and is the ratio of the variability among the group means to the variability within each group (also called residual variability).



Poisson and Related Distributions

Many processes produce events randomly at a given overall rate—visitors arriving at a website, or cars arriving at a toll plaza (events spread over time); imperfections in a square meter of fabric, or typos per 100 lines of code (events spread over space).

Lambda The rate (per unit of time or space) at which events occur.

Poisson distribution The frequency distribution of the number of events in sampled units of time or space.

Exponential distribution The frequency distribution of the time or distance from one event to the next event.

Weibull distribution A generalized version of the exponential distribution in which the event rate is allowed to shift over time.

Statistical Experiments and Significance Testing

A/B Testing

imagine you want to test two version of the same thing to discover which one is better. For example: medical therapies, web ads,soil fertilizers, etc.

One of the two groups could be standard or without intervention, in this case is called a control group

We keep track of statistic for both groups and if there is a difference it can depend either on chance or on the effect we look for.

To answer this question we use statistical hypothesis testing

Hypothesis Tests

Hypothesis tests, also called significance tests, are ubiquitous in the traditional statistical analysis of published research. Their purpose is to help you learn whether random chance might be responsible for an observed effect.

Null hypothesis The hypothesis that chance is to blame.

Alternative hypothesis Counterpoint to the null (what you hope to prove).

One-way test Hypothesis test that counts chance results only in one direction.

Two-way test Hypothesis test that counts chance results in two directions.

Resampling

There are two main types of resampling procedures: the bootstrap and permutation tests. The bootstrap is used to assess the reliability of an estimate; Permutation tests are used to test hypotheses, typically involving two or more groups

One virtue of resampling, in contrast to formula approaches, is that it comes much closer to a one-size-fits-all approach to inference.

- Data can be numeric or binary.
- Sample sizes can be the same or different.
- Assumptions about normally distributed data are not needed.

Permutation Test

1. Combine the results from the different groups into a single data set.
2. Shuffle the combined data and then randomly draw (without replacement) a resample of the same size as group A (clearly it will contain some data from the other groups).
3. From the remaining data, randomly draw (without replacement) a resample of the same size as group B.
4. Do the same for groups C, D, and so on. You have now collected one set of resamples that mirror the sizes of the original samples.
5. Whatever statistic or estimate was calculated for the original samples (e.g., difference in group proportions), calculate it now for the resamples, and record; this constitutes one permutation iteration.
6. Repeat the previous steps R times to yield a permutation distribution of the test statistic.

Statistical Significance and p-Values

Statistical significance is how statisticians measure whether an experiment (or even a study of existing data) yields a result more extreme than what chance might produce. If the result is beyond the realm of chance variation, it is said to be statistically significant.

p-value Given a chance model that embodies the null hypothesis, the p-value is the probability of obtaining results as unusual or extreme as the observed results.

Alpha The probability threshold of “unusualness” that chance results must surpass for actual outcomes to be deemed statistically significant.

Type 1 error Mistakenly concluding an effect is real (when it is due to chance).

Type 2 error Mistakenly concluding an effect is due to chance (when it is real).

t-Tests

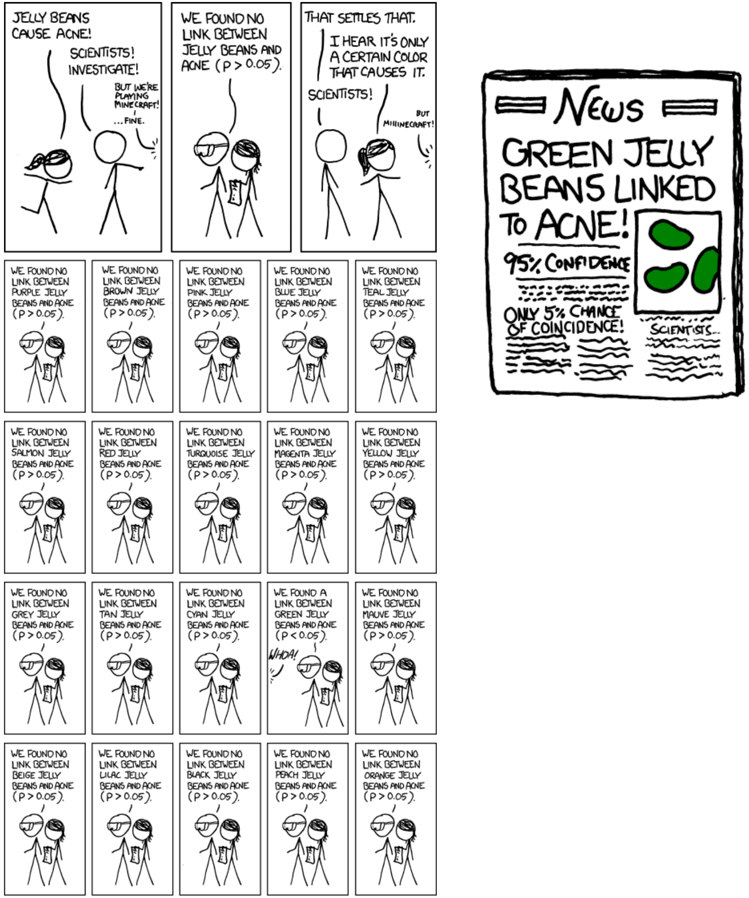
Based on t-distribution. Their goal is assessing the statistical significance of the difference between two sample means. There are three main types of t-test:

- An Independent Samples t-test compares the means for two groups.
- A Paired sample t-test compares means from the same group at different times (say, one year apart).
- A One sample t-test tests the mean of a single group against a known mean.

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

Multiple Testing

if you have 20 predictor variables and one outcome variable, all randomly generated, the odds are pretty good that at least one predictor will (falsely) turn out to be statistically significant if you do a series of 20 significance tests at the alpha = 0.05 level.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

https://www.explainxkcd.com/wiki/index.php/882:_Significant

Multiple Testing

Adjustment procedures in statistics can compensate for this by setting the bar for statistical significance more stringently than it would be set for a single hypothesis test.

- One such procedure, the Bonferroni adjustment, simply divides the alpha by the number of comparisons.
- Another, used in comparing multiple group means, is Tukey's "honest significant difference," or Tukey's HSD. This test applies to the maximum difference among group means, comparing it to a benchmark based on the t-distribution

ANOVA

Suppose that, instead of an A/B test, we had a comparison of multiple groups, say A/B/C/D, each with numeric data: e.g web pages visits.

Instead of worrying about all the different comparisons between individual pages we could possibly make, we can do a single overall test that addresses the question, “Could all the pages have the same underlying stickiness, and the differences among them be due to the random way in which a common set of session times got allocated among the four pages?”

The procedure used to test this is ANOVA.

ANOVA (2)

The basis for it can be seen in the following resampling procedure:

1. Combine all the data together in a single box.
2. Shuffle and draw out four resamples of five values each.
3. Record the mean of each of the four groups.
4. Record the variance among the four group means.
5. Repeat steps 2–4 many (say, 1,000) times.

What proportion of the time did the resampled variance exceed the observed variance? This is the p-value.

Just like the t-test can be used instead of a permutation test for comparing the mean of two groups, there is a statistical test for ANOVA based on the F-statistic. The F-statistic is based on the ratio of the variance across group means (i.e., the treatment effect) to the variance due to residual error

Chi-Square Test

The chi-square test is used with count data to test how well it fits some expected distribution. The most common use of the chi-square statistic in statistical practice is with r × c contingency tables, to assess whether the null hypothesis of independence among variables is reasonable.

The classical way is based on the X-statistic and Pearson residual R

$$R = \frac{Observed - Expected}{\sqrt{Expected}} ; X = \sum_i^r \sum_j^c R^2$$

The X-statistic follows a chi-square distribution with $v = (r - 1) \times (c - 1)$ degrees of freddom, from which the p-value can be computed

Chi-Square Test

There is also a resempling-based version of this test

1. Constitute a box containing the sum of the values in each row
2. Shuffle, take c separate samples of 1,000, and count.
3. Find the squared differences between the shuffled counts and the expected counts and sum them.
4. Repeat steps 2 and 3, say, 1,000 times.
5. How often does the resampled sum of squared deviations exceed the observed? That’s the p-value.

Power and Sample Size

Power is the probability of detecting a specified effect size with specified sample characteristics (size and variability). Usually set to 80% .

The most common use of power calculations is to estimate how big a sample you will need.

For calculating power or required sample size, there are four moving parts:

- Sample size
- Effect size you want to detect
- Significance
- Power

Specify any three of them, and the fourth can be calculated

Power and Sample Size

The statsmodels package contains several methods for power calculation.

Here, we use proportion_effectsize to calculate the effect size and TTestIndPower to solve for the sample size:

```
effect_size = sm.stats.proportion_effectsize(0.0121, 0.011)
analysis = sm.stats.TTestIndPower()
result = analysis.solve_power(effect_size=effect_size,
alpha=0.05, power=0.8, alternative='larger')
print('Sample Size: %.3f' % result)
--
Sample Size: 116602.393
```