Introduction to statistics with R

Anastasiia Zharinova,

Senior Economic Manager,

Economics and Strategic Analysis team, NHS England

November 2022

NHS England and NHS Improvement

What this workshop is about?

2.5 hours with 10 min break, mix of theory and practice with the focus on descriptive statistics and basics of inferential statistics

1. Investigating numeric, time series and categorical data
2. Distributions
3. Relationships between variables
4. Hypotheses testing

Please note
Statistics is a huge subject and is usually being studied in the university for years
This is just a selection of methods/tools that I used a lot during my analytical career
Coding styles are very different and different people have different preferences

Materials:

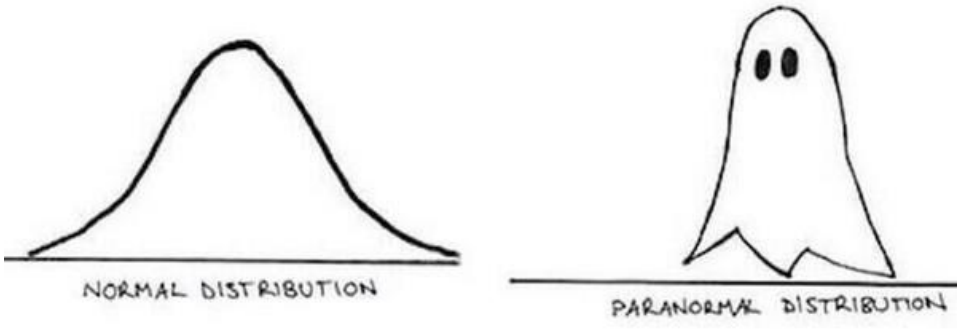
Main coursebook - Learning statistics with R: A tutorial for psychology students and other beginners by Danielle Navarro

Data – NHS Datasets package

Additional reading - An Introduction to Statistical Learning with Applications in R

Why is it important?

1. Statistics is everywhere
2. Before doing more advanced modelling, we need to understand the data
3. Even simple statistical techniques and one line of code can give us insights to the dataset
4. …or its quality
5. Understanding statistics can help us judge the messages
6. … and be more careful when we share ours
7. Statistics helps presenting data in a manner that is clear to everyone
8. And it can be fun!



Part 1. Understanding the data

Descriptive statistics

Summary statistic that quantitatively describes or summarizes features from a collection of information. In healthcare analytics, this usually includes:

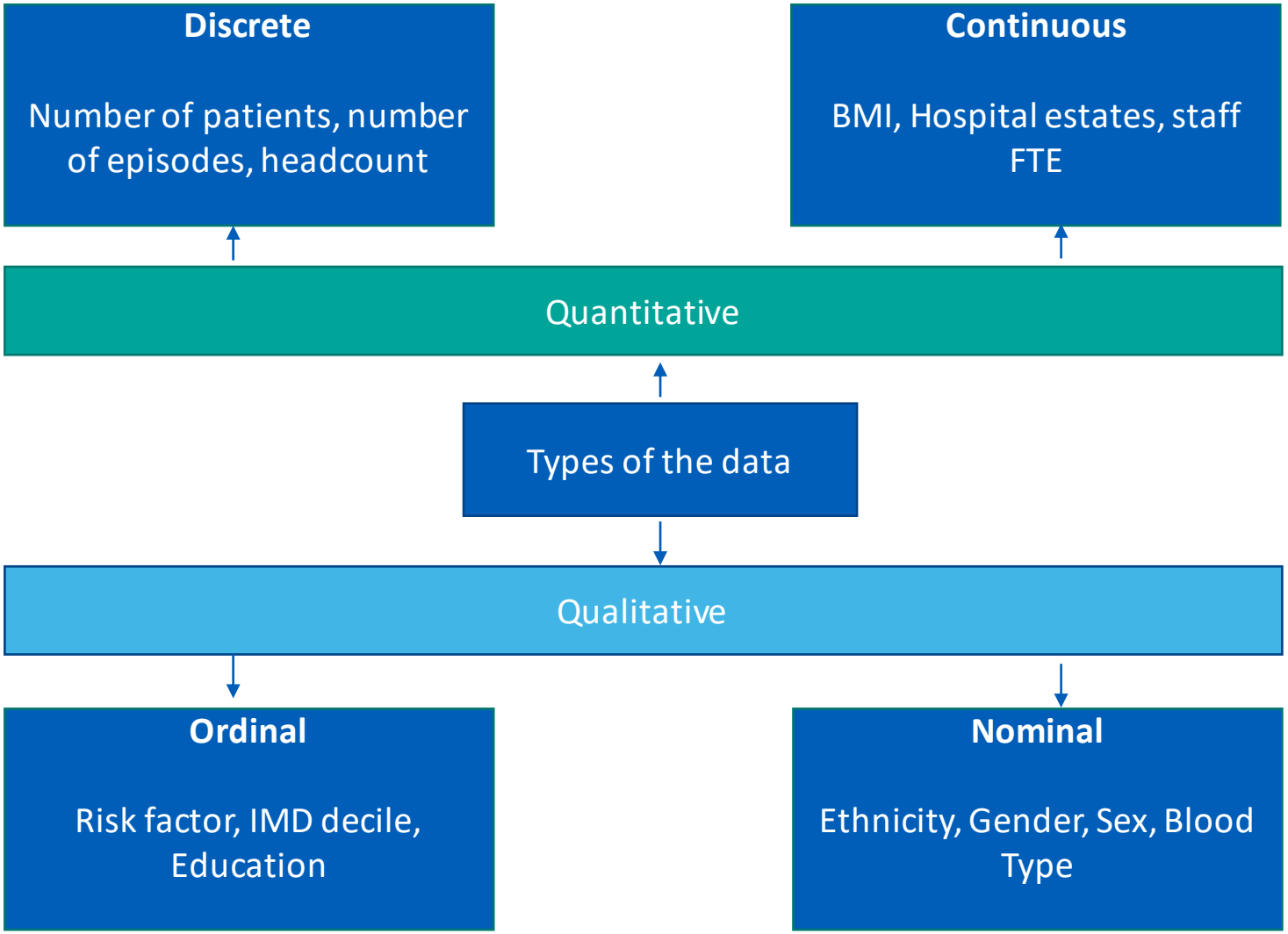
1. Understanding univariate data
    - Central tendency
    - Measures of variability
    - Distribution, skewness and kurtosis
2. Understanding Bivariate and multivariate data
    - Correlation/dependence
    - Cross-tabulations and contingency tables

All of the above can be done visually or numerically.

While descriptive statistics focuses on observed data, inferential statistics is using data analysis to infer properties of an underlying distribution of probability

In R, there is a range of packages that produce summary statistics, such as Hmisc, **pastecs**, psych . However, base R has a number of helpful functions, too.

Pre-requisite: types of the data



Measures of central tendency

Mean

Average value
3, 5, 8, 11, 15 – mean is 8.4

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Trimmed mean

Average value if we drop an extreme value
3, 5, 8, 11, 15, 100 – mean is 23.7

If we discard the largest 10% of the observations
3, 5, 8, 11, 15, ~~100~~ - mean is 8.4

Median

The middle value
3, 5, 8, 11, 15 – median is 8

3, 5, 8, 11, 15, 18 – median is the middle between 8 and 11=9.5

Mode

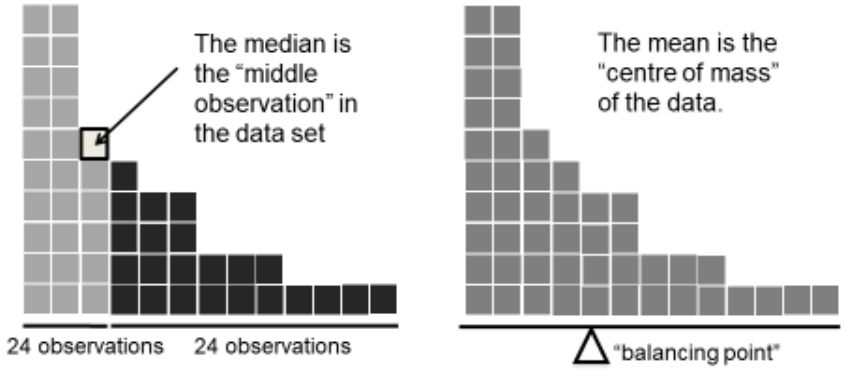
Most common value
3, 5, 8, 8, 11, 15 – mode is 8

Can also be used on qualitative data

Measures of central tendency

Mean or median?

- If your data is ordinal, median is more useful than mean
- If your data is discrete or continuous, mean and median both could work. However, mean can be more helpful if you have low number of observations.
- If your data is nominal, neither are informative



What should I use for proportions and ratios?

- Both median and mean are appropriate
- And both can be very useful. E.g. if we are looking at recovery rate of elective care across England – how much elective activity providers do now compared with pre-covid levels
  - On average, providers recovery is at 105%
  - Half of providers reached 104% recovery

How do I know if I need trimmed mean?

- Some NHS data is tidier than other
- Some observations are unrealistic
- Sometimes we have to drop such outliers anyway, so if we do, we technically calculate 'trimmed mean'

Measures of variability

1. **Range –** difference between max and min
2. **Max to mean ratio** – ratio of maximum value to the minimum value
3. **Interquartile range** - difference between the 25th quartile* and the 75th quartile*
4. **Mean (average) absolute deviation** – mean of absolute deviations from the mean - $(X) = \frac{1}{N}\sum_{i=1}^{N}|X_i - \bar{X}|$
5. **Median absolute deviation** - median of absolute deviations from the median $median(|x_i - m|)$
6. **Variance** – squared deviation of a variable from the mean $\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$
7. **Standard deviation** - square root of the variance $s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$
8. **Coefficient of variation (relative standard deviation)**** – ratio of a standard deviation to a mean $\widehat{c_v} = \frac{s}{\bar{x}}$

*Quartiles and quantiles

Quantile (percentile) - cut points dividing the range into continuous intervals
Quartile is a quantile which divides the number of data points into four parts, or quarters, of more-or-less equal size

** In the same way, we can calculate coefficients of mead and median deviations, but those are less common

Measures of variability

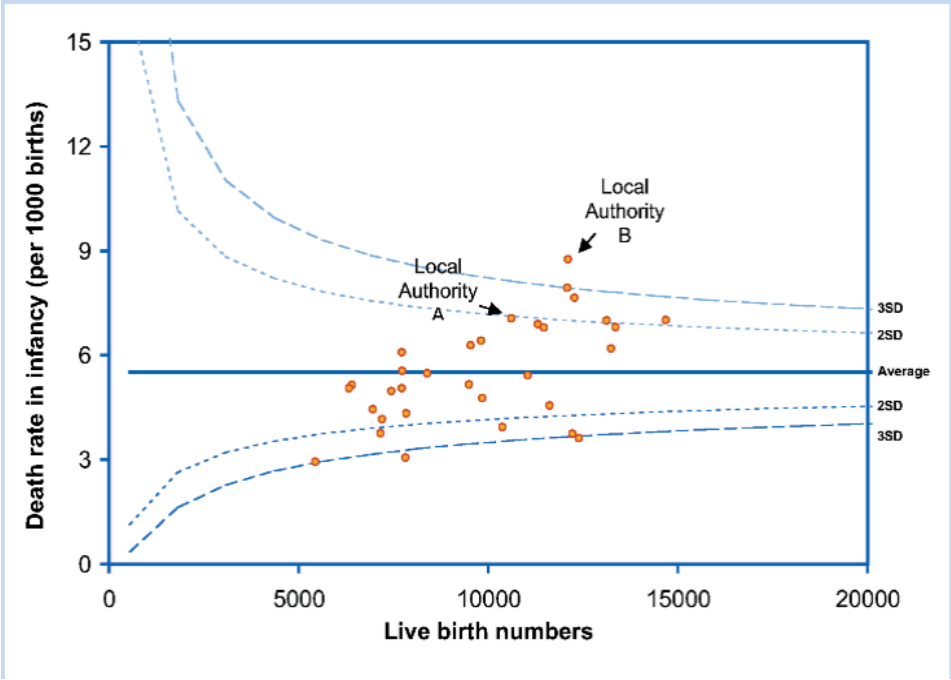
Measure	Pros	Cons	When to use
Range	Gives full spread of the data	Sensitive to outliers	To spot data quality
Max to mean ratio	Helpful with rates	Not easy to interpret	When comparing variation across different metrics
Interquartile range	Works for skewed data	Could be difficult to explain	In summary statistics
Mean absolute deviation	Easy to interpret	Not accurate if the data skewed	When data is normally distributed
Median absolute deviation	Easy to interpret	Not very common	When data is not normally distributed
Variance	In-built in most statistical packages	Does not use the same units as data, so difficult to interpret	When using variance for further modelling
Standard deviation	Expressed in the same units as data, so easy to interpret	Not as helpful if the mean is misleading	In summary statistics for the metric of interest
Coefficient of variation	Variation in a context of the mean, can be used to compare different metrics	Misleading for very low numbers (where mean is close to 0)	When understanding variations across different metrics

Some visual presentations of variability will be considered in the next chapter

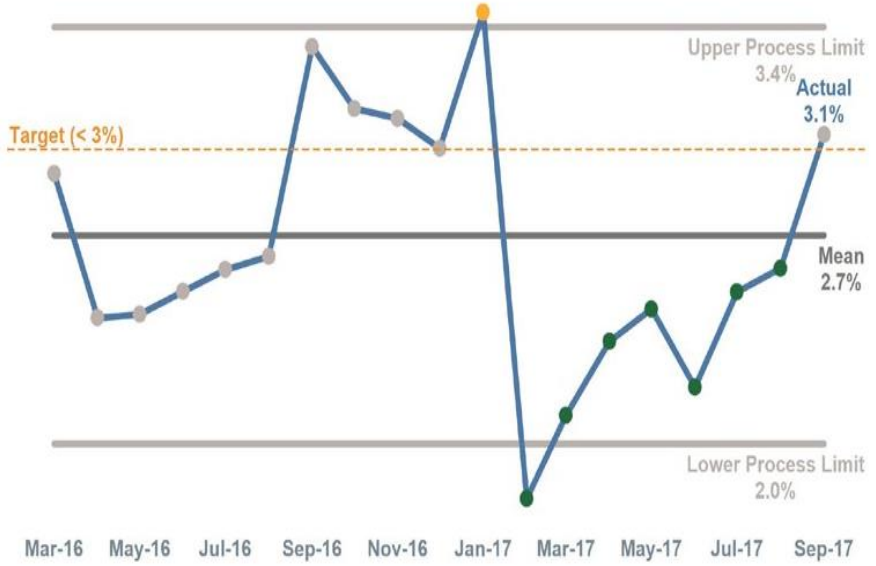
Statistical process control (SPC) methods

SPS is the use of statistical techniques to control a process or production method. Two most common methods that are used in the NHS to understand what is 'different' and what is the 'norm'.

Funnel plots



Statistical process control charts



https://www.england.nhs.uk/publication/making-data-count/

Part 2. Probability and distributions (30 min)

Context

Inferential statistics – applying inferences in the data - is built on probability theory. Probabilities are relatively common in healthcare analytics and epidemiology, for example, in the risk stratification or calculating a probability of an event. Other statistical tools, such as forecasting and regression modelling are also built on probabilities.

There are two approaches to probability:

Frequentist	Bayesian
Assigns probability to the data as a long-run frequency	Assigns probability to hypotheses
No prior probability	Some prior probability/belief
Confidence interval: over the infinite sample size taken from population, 95% of these contain true population value	Credible interval: 95% probability that the population value is within the limits of an interval
Hypotheses being tested using p value and significance levels	Hypotheses being tested using Bayes factor

We will focus on frequentist view in the next 2 sections.

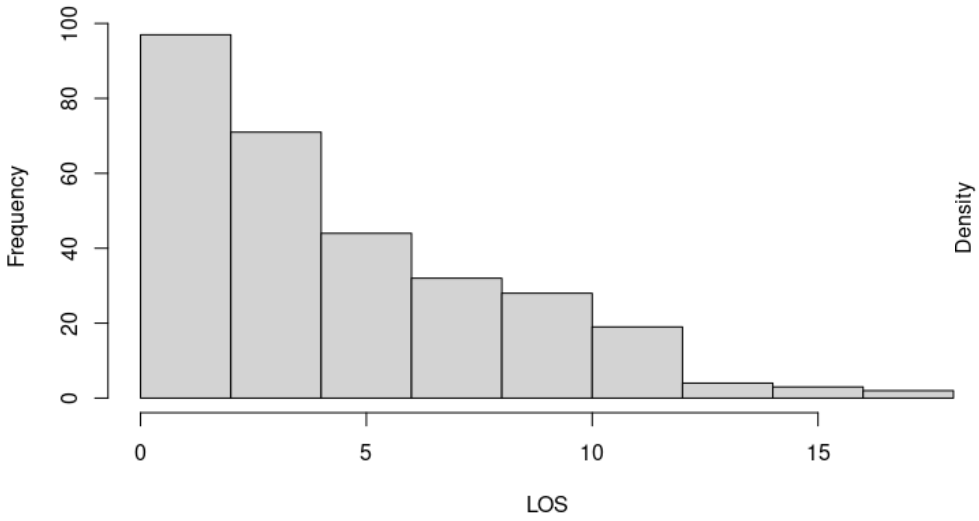
https://medium.com/analytics-vidhya/a-short-story-on-bayesian-vs-frequentist-statistics-27f55ae56253

Frequency and probability distributions

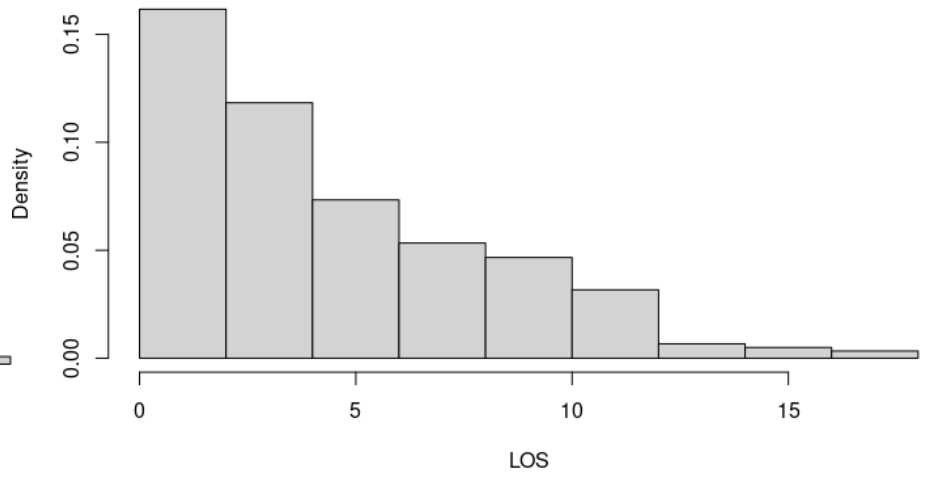
In frequentist view, each event has a frequency. As a result, we can visualise frequency of an even in a form of histogram

If probabilities of events add up to 1 (law of total probability), we can identify probability distribution.

Frequency distribution
(distribution of frequency)

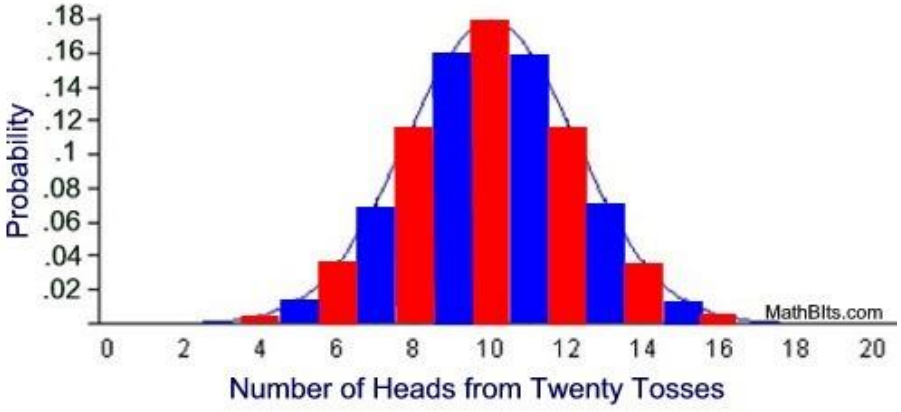
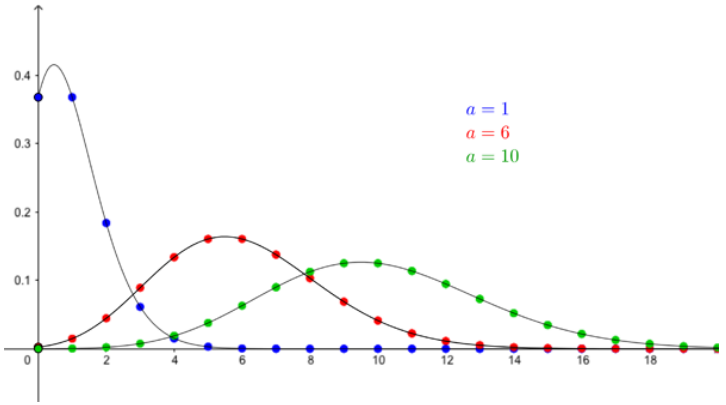


Probability distribution
(distribution of probability)

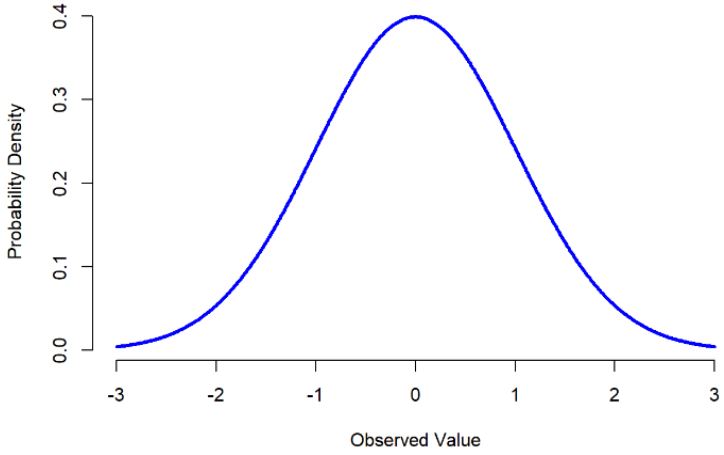
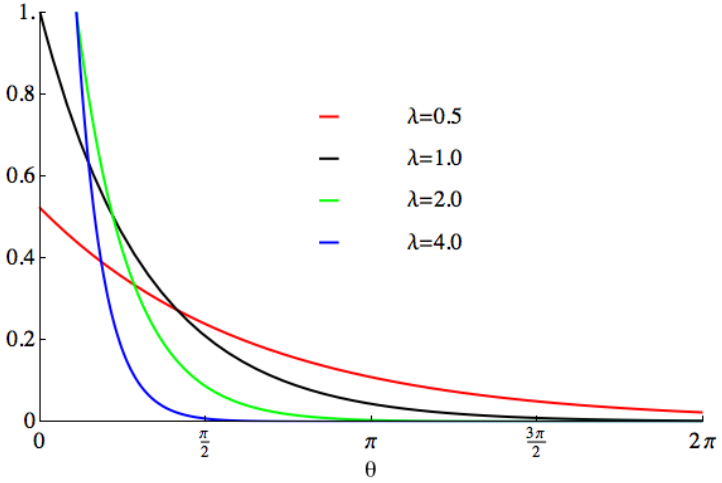


How do we know the distribution of our data? We can identify visually or do statistical tests, e.g Shapiro-Wilk's. More information in workshop on fitting the distributions - https://github.com/semanzi/fitting_distributions_with_R_NHSR_2021

Most common types of distributions - discrete

	Binomial distribution	Poisson distribution
Definition	the distribution of binary data from a finite sample	the distribution of discrete number of events from an finite sample
Formula	$P(X) = {}_nC_x p^x (1-p)^{n-x}$	$P(x) = \frac{e^{-\lambda} * \lambda^x}{x!}$
Example	Flipping a coin 20 times	Distribution of a number of GP calls
Visualisation		

Most common types of distributions - continuous

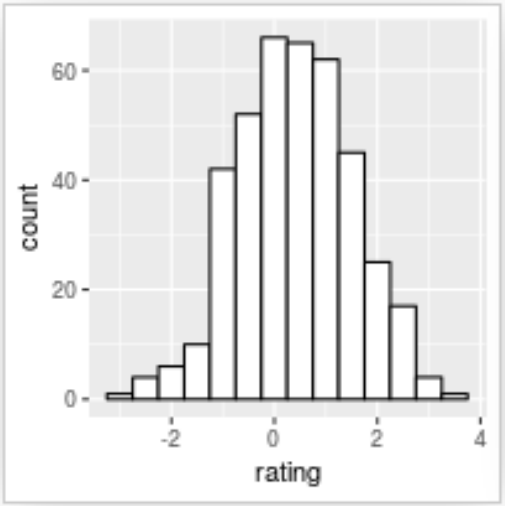
	Normal distribution	Exponential distribution
Definition	Continuous distribution of a random values. Also known as bell curve or Gaussian distribution	Continuous distribution of a random values that is the best represented as a function of exponenta
Formula	$Z = \frac{X - \mu}{\alpha}$	$f(x) = \begin{cases} \lambda e^{-\lambda * x}, x \geq 0 \\ 0, x < 0 \end{cases}$
Example	Birth weight	GP call duration
Visualisation		

Ways to plot a distribution

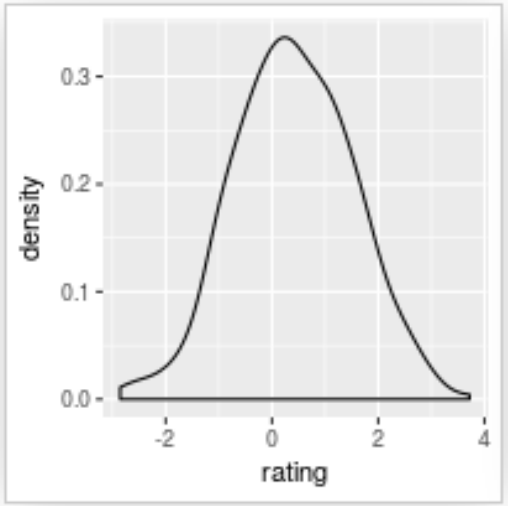
As seen in earlier slides, distributions can be plotted as:

1. Histogram - frequencies of values of a variable bucketed into ranges. Shape is affected by the number of buckets aka bins.
2. Density line – probabilities of a values. A better representation as does not depend on number of bins.

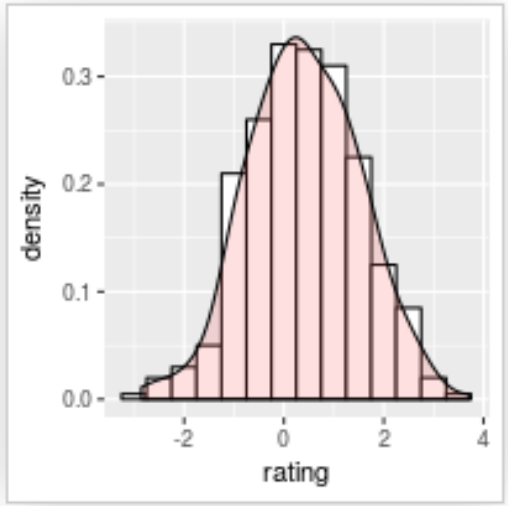
Histogram



Density line



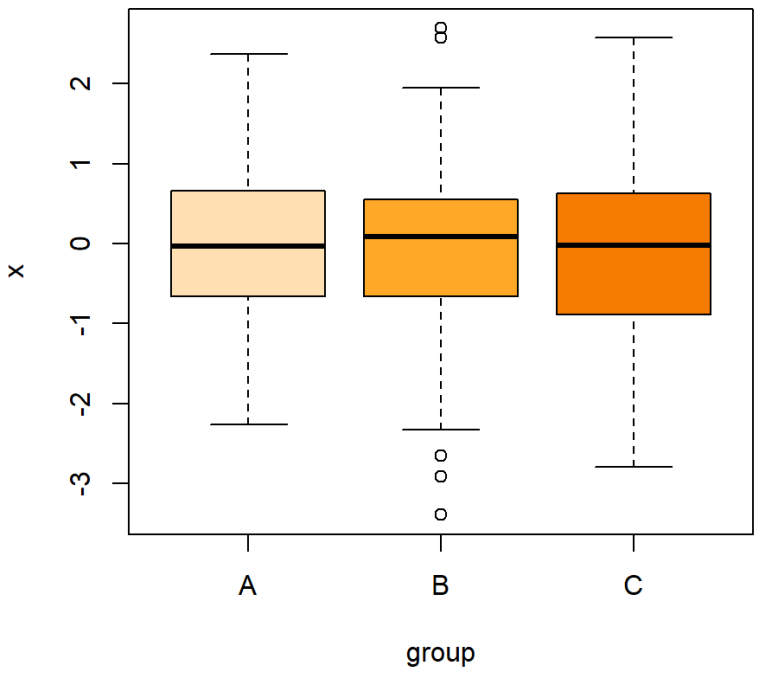
Histogram with density overlay



Ways to plot a distribution

Boxplots

Boxplots are one of the most common ways to plot distribution as they include median, IQR, max, min and outliers



Violin plots

Violin plots are the combination of boxplot and density line

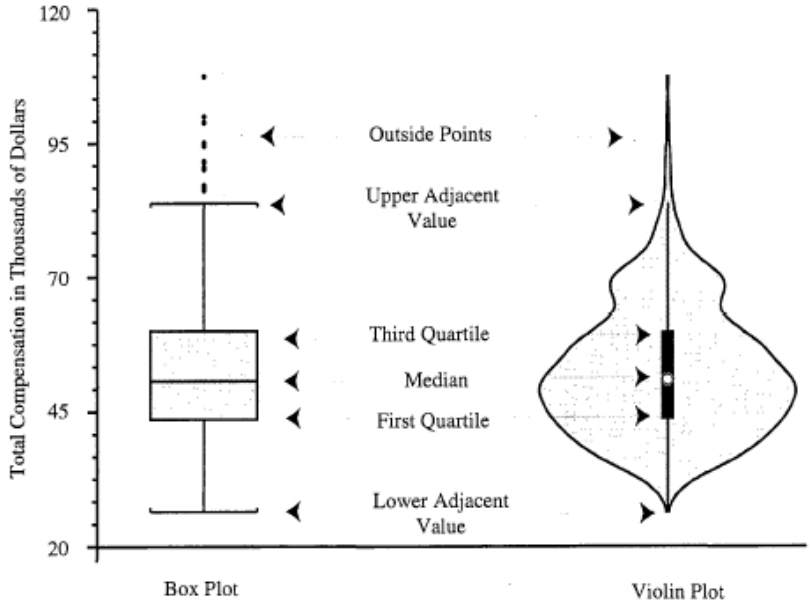


Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.

Part 3. Relationship between variables and hypotheses testing (40 min)

Context

One of the most common metrics when analysis two variables is covariance. Covariance is a measure of the joint variability of two random variables. However, covariance is not as informative – positive covariance means that variables are dependent on each other, but the magnitude of the effect cannot be estimated.

$$\operatorname{cov}(X, Y) = \sum_{i=1}^{N} \frac{(x_i - \overline{x})(y_i - \overline{y})}{N}.$$

Instead, we use correlation, where both sign and value can be used. E.g. Pearson correlation coefficient is built on covariance.

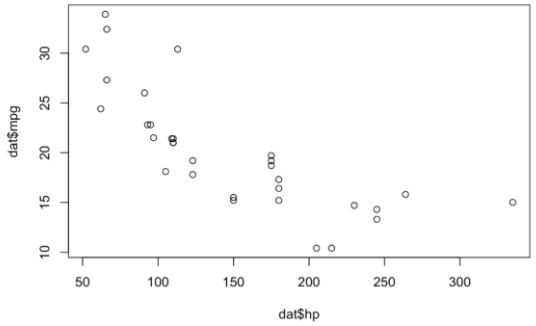
$$r_{xy} = \frac{\mathrm{Cov}(x,y)}{S_x S_y}$$

It is not always clear what is strong and what is weak correlation. Overall, statisticians somewhat agreed that:

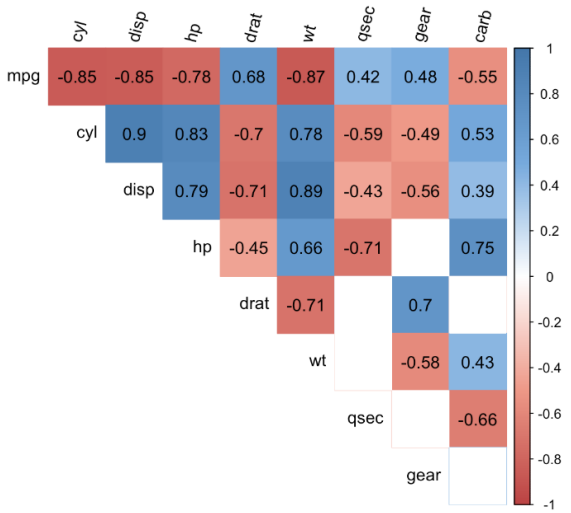
- Absolute correlation 0.9 – very strong
- Absolute correlation between 0.7 and 0.9 – strong
- Absolute correlation between 0.4 and 0.7 – moderate
- Otherwise - weak

Correlation between variables

	Empirical	Visual
2 vars	**Correlation metrics:** 1. Pearson correlation – for quantitative continuous variables. Can only capture linear relationships 2. Spearson correlation – for the qualitive ordinal variables. Can capture both linear and non-linear relationships	**Scatterplot**
n vars	**Correlation matrix**	**Correlogram**




```
##        mpg   cyl  disp    hp  drat    wt  qsec  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00 -0.21 -0.66
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66  0.27  1.00
```



https://statsandr.com/blog/correlation-coefficient-and-correlation-test-in-r/

Hypotheses testing

In the same way researchers work with hypotheses, statisticians can test their beliefs about statistical patterns, such as probabilities, correlation, variable taking a specific value.

H0 – null hypothesis

H1 – alternative hypothesis

To test hypotheses, we use various statistical test to calculate t-statistics () and p-value(). We can then reject or accept hypothesis at a specific confidence level, e.g. 95%

For example, if we want to check correlation

H0: true correlation is equal to 0

H1: true correlation is not equal to 0

However, analysts will know that correlation does not imply causation. Regression modelling workshop from NHS-R Senior Fellow Chris Mainey - https://github.com/chrismainey/Regression_Modelling_NHSR

Comparing values

Hypotheses testing can also be useful when we want to compare values:

- Between different groups. Is there a statistically significant difference in hospitals' performance?
- In time. Did population become healthier after implemented intervention?

There are two most common tests:

1. Student's T test.
2. Welch's T test

Both of the tests assume normality of distributions and independence. However, Student's t-test also implies equal variation.

Their principle is the same as any other statistical test and is using hypotheses testing.

H0: true difference in mean is equal to 0

H1: true difference in means is not equal to 0