TESA Data Management Workshop

Data Management in Clinical Research

Management of clinical trial data
Data Analysis

Arsénio Nhacolo
Joe Brew

Blantyre, Malawi, July 09-12, 2019

Outline I

Day 2

- Data coding
- Data sharing (with exercise)
- Data processing/cleaning
- Descriptive statistics
- Exploratory data analysis
- Confirmatory data analysis

Data coding

- "Coding" is the act of assigning a (usually numeric) value to a categorical concept.
- Example: Female = 1, Male = 2, Other = 3, Unknown = 4
- Example: Aged 0-5 = 1, Aged 6-18 = 2, Aged 19-45 = 3, Aged 46+ = 4, Unknown = 98

Advantages of coding

- Saves physical space on paper CRFs
- Saves significant time in paper-to-digital data entry
- Forces categorization (not necessarily good)
- Forces a priori thinking about meaningful categorization
- Saves hard-drive space

Do you need to "code"?

- Lots of data capture is now digital
- Categorization is not necesarilly good
- Hard-drive space is rarely a limiting issue
- Coding means one more layer between the data and understanding it

If you choose to code...

- You should have comprehensive data dictionaries: both machine- and human readable
- Your "levels" should make ordenal/notional sense
- Your categories/codes should be tested prior to deployment

Why a ”machine-readable” codebook?

- Automated joins vs. manual recoding
- Easier to subset, search, etc.
- Easier to build downstream software / data products with

Data sharing

- Identifying vs non-identifying information
- Health vs non-health data
- Raw vs processed
- Individual vs aggregated

Data sharing via aggregation

- What is aggregation?
- How do we aggregate?
- A practical exercise

The verbs

- filter
- mutate
- select
- arrange
- mutate
- group_by
- summarise

Data processing

Statistical analysis

Intro

Scientific method

- Competing hypotheses about nature (H_0, H_1, H_2, H_3,. . .)
- Design a study and collect data
- Data give evidence in support of some of the hypotheses more than others
- Science is a process of discarding hypotheses that are inconsistent with observation (data)

Statistical analysis

Biostatistics overview

Statistics is the scientific use of data to describe and/or draw inferences about phenomena.

Biostatistics is the application of statistical reasoning and methods to the solution of biological, medical and public health problems.

How to properly produce and collect data is studied in experimental design and sampling theory.

Organisation and description of data is the subject area of descriptive statistics.

How to draw conclusions from data is the subject of statistical inference.

Role of biostatistics

- Ask questions (generate hypotheses)
- Design and conduct studies to answer questions (generate evidence by making observations, collecting data)
- Describe the observations (descriptive statistics)
- Evaluate the data, assess the strength of the evidence in favour of/against hypotheses (statistical inference)

Statistical analysis

Biostatistics overview – EDA

Exploratory data analysis (EDA)

- Descriptive statistics
- First step of data analysis
- Use of numerical and graphical methods to:
  - give an overview of data;
  - visualize distributions/patterns/variation and relationships;
  - detect anomalies (missing values, outliers, etc.);
  - assess the assumptions/decide on appropriate methods for confirmatory analysis;
  - generate hypothesis for future research.
- EDA can also use confirmatory analysis (statistical inference) methods, as long as their use lies in building a hypothesis.

Confirmatory data analysis (CDA)

- Inferential statistics
- Use of estimation and test of hypotheses methods to draw conclusions about a larger population from a sample:
  - estimate unknown population parameters based on sample statistics;
  - assess the strength of evidence for competing hypotheses;
  - compare populations based on their samples;
  - make predictions.
- CDA uses data to conjecture what is true or likely to be true

Statistical reasoning

- Natural laws do not perfectly predict phenomena/events
- There is a natural variation that leads to uncertainty about an event
- Despite variation, there are important patterns that can be discovered

Variables

Variable is a characteristic taking on different values (e.g., gender, age, blood pressure)

Random variable is a variable whose values are partly due to chance.

Classification of variables by role

- Explanatory or independent variable – variable affecting or causing the outcome/response
- Response or dependent variable – variable that is affected or caused (outcome measure)

We're about to start some analysis

Let's get ready

- Open RStudio
- Download the .csv file from databrew.cc/data/mozambique.csv onto your Desktop
- Read the data into R by running the following

```
> library(tidyverse)
> df <- read_csv('~/Desktop/mozambique.csv')
```

Statistical analysis

Biostatistics overview – Variables

Classification of variables by measurement scale

- Quantitative (numerical, amount)
  - Discrete (e.g., number of children)
  - Continuous (e.g., weight)
- Qualitative (categorical, attribute)
  - Nominal (e.g., race, gender, religion)
  - Ordinal (e.g., education level)

Qualitative and quantitative variables

Let's take a glimpse at the data

```
> glimpse(df)
> nrow(df)
> ncol(df)
```

Exploratory data analysis

Exploratory data analysis

EDA is generally cross-classified in two ways:

- non-graphical or graphical
- univariate or multivariate (usually bivariate)

Non-graphical methods involve calculation of summary statistics, while graphical methods summarize the data in a diagrammatic or pictorial way.

Univariate methods look at one variable (data column) at a time, while multivariate methods look at two or more variables at a time to explore relationships.

Univariate non-graphical EDA

Usual goals:

- understand the distribution
- detect outliers

For categorical variables

Tabulation of frequency and relative frequency of each category is the commonly used univariate non-graphical EDA.
E.g.:

Gender	Count	Proportion(%)
Female	89	55.28
Male	72	44.72
Total	161	100.00

Univariate non-graphical EDA for a quantitative variable

```
> min(df$year)
> max(df$year)
> range(df$year)
> length(df$year)
```

Exploratory data analysis

Univariate non-graphical EDA

Univariate non-graphical EDA for a categorical variable

```
> df %>%
+   group_by(event_type) %>%
+   summarize(total = n())
```

Exploratory data analysis

Univariate non-graphical EDA

Univariate non-graphical EDA for a categorical variable

```
> df %>%
+   group_by(event_type) %>%
+   summarize(total = n()) %>%
+   arrange(total) %>%
+   mutate(percent = total / sum(total) * 100)
```

For quantitative variables

Preliminary assessments, using the observed sample, about the population distribution characteristics of the variable:

- Central tendency – typical or middle values
  - Mean – the sum of all of the data values (x_i) divided by the number of values (n). Is the most common measure of central tendency.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  - Median – the middle value (or the average of the two middle values in case of even number of values) after all of the values are put in a ordered list. It is preferred over the mean for skewed distribution or when there are outliers.
  - Mode – the most likely or frequently occurring value. It is rarely used.

Exploratory data analysis

Univariate non-graphical EDA

For quantitative variables

Preliminary assessments, using the observed sample, about the population distribution characteristics of the variable:

```
> mean(df$fatalities)

[1] 0.959381

> median(df$fatalities)

[1] 0
```

- Spread – indicator of how far away from the centre we are still likely to find data values
  - Variance – the mean of the squared differences of the observations and the corresponding mean.

    $$\text{Population: } \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}, \text{ Sample: } s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

  - Standard deviation (SD) – the square root of the variance. It is more interpretable than variance, as it has the same units as the original data.
  - Interquartile range (IQR) – quartiles are 3 values dividing the ordered data in 4 equally sized parts. One quarter of the data fall below the 1^{st} quartile (Q_1), one half fall below the 2^{nd} quartile ($Q_2 = Median$), and three fourths below the 3^{rd} quartile (Q_3). $IQR = Q_3 - Q_1$. More spread out data imply higher IQR.
  - Range – is $maximum - minimum$

Exploratory data analysis

Univariate non-graphical EDA: Variance

```
> # Variance:  the mean of the squared differences of the
> x <- df$fatalities
> mean(abs(mean(x) - x)^2)

[1] 8.234327
```

Exploratory data analysis

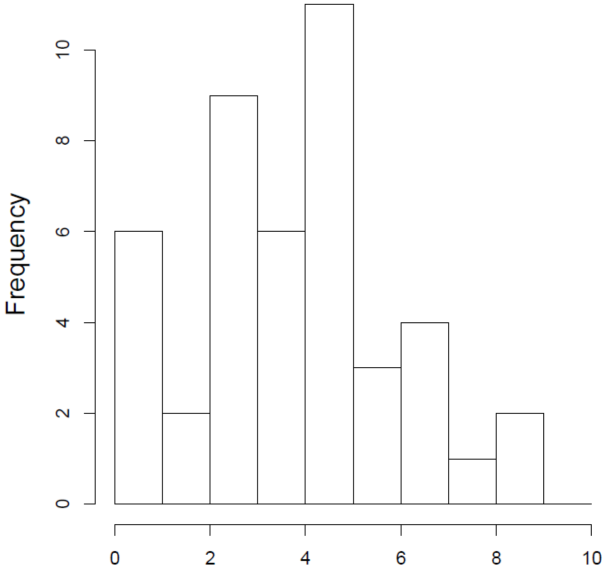
Univariate non-graphical EDA: Standard deviation, interquartile range, range

```
> # Variance:  the mean of the squared differences of the
> x <- df$fatalities
> # Standard deviation
> sd(x)

[1] 2.87094

> # Quantile
> quantile(x, probs = c(0.25, 0.75))

25% 75%
  0   1

> # Range
> range(x)

[1]  0 58
```

Univariate graphical EDA

Histogram

- A bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases in each category, for qualitative variables, or in each bin (a range of values), for quantitative variables.
- One of the best ways to quickly learn a lot about your data, including central tendency, spread, modality, shape and outliers.


Frequency
0
2
4
6
8
10
0
2
4
6
8
10

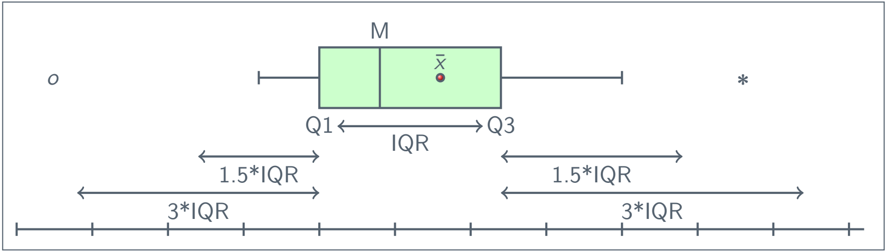
Univariate graphical EDA

Histogram

```
> hist(df$fatalities)
```

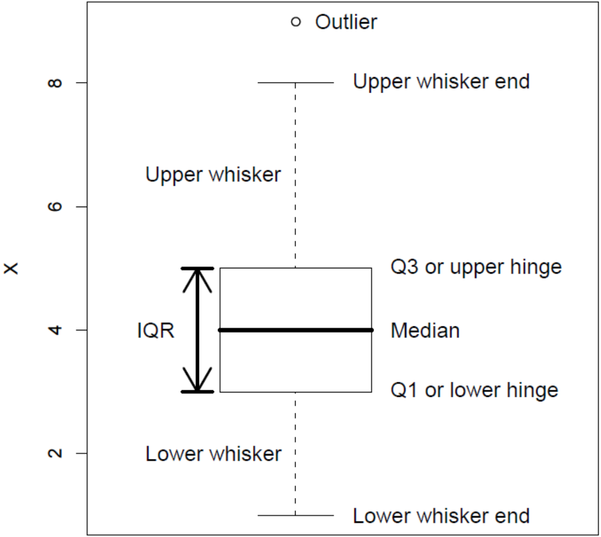
Boxplot

- Also known as box and whisker plot
- Shows robust measures of location (central tendency) and spread
- Provide information about symmetry and outliers



Exploratory data analysis

Univariate graphical EDA – Boxplot



Exploratory data analysis

Univariate graphical EDA – Boxplot

Boxplot

```
> boxplot(df$fatalities)
```

Quantile-quantile plot

- Graphical method for assessing if a set of data came from some theoretical distribution
- Created by plotting two sets of quantiles against one another
- If both sets of quantiles came from the same distribution, the points form a roughly straight line
- Normal Q-Q plot is the frequently used

Exploratory data analysis

Univariate graphical EDA – QQ plot

Multivariate non-graphical EDA

Generally used to show the relationship between two or more variables in the form of either cross-tabulation or statistics.

Cross-tabulation

- For two variables, a two-way table with column headings that match the levels of one variable and row headings that match the levels of the other variable, filled with the counts of all subjects that share a pair of levels
- Depending on the goals, row percentages, column percentages and/or cell percentages are also included
- For three or more variables make separate two-way tables for two variables at each level of a third variable

Multivariate non-graphical EDA

```
> df %>%
+   group_by(admin1, event_type) %>%
+   summarise(n = n())
# A tibble: 59 x 3
# Groups:   admin1 [11]
   admin1       event_type                     n
   <chr>        <chr>                      <int>
 1 Cabo Delgado Battles                       31
 2 Cabo Delgado Protests                      15
 3 Cabo Delgado Riots                         14
 4 Cabo Delgado Strategic developments        14
 5 Cabo Delgado Violence against civilians   134
 6 Gaza         Battles                        3
 7 Gaza         Protests                       1
 8 Gaza         Riots                         13
```

Multivariate non-graphical EDA

Adding percentages

```
> df %>%
+   group_by(admin1, event_type) %>%
+   summarise(n = n()) %>%
+   group_by(admin1) %>%
+   mutate(p = n / sum(n) * 100)
```

```
# A tibble: 59 x 4
# Groups:   admin1 [11]
   admin1       event_type                     n     p
   <chr>        <chr>                      <int> <dbl>
 1 Cabo Delgado Battles                       31 14.9 
 2 Cabo Delgado Protests                      15  7.21
 3 Cabo Delgado Riots                         14  6.73
 4 Cabo Delgado Strategic developments        14  6.73
 5 Cabo Delgado Violence against civilians   134 64.4 
```

Exploratory data analysis

Multivariate non-graphical EDA – Cross-tabulation

gender	smoke
Female	No
Female	Yes
Female	No
Male	Yes
Male	Yes
Female	Yes
Male	Yes
Female	No
Female	No
Male	Yes
Female	No

	Smoking n(%)		
Gender	No	Yes	Total
Female	81(83.5)	16(16.5)	97(100)
Male	28(44.4)	35(55.6)	63(100)
Total	109(68.1)	51(31.9)	160(100)

Exploratory data analysis

Multivariate non-graphical EDA – Univariate statistics by category

Univariate statistics by category

- For one categorical variable (usually explanatory) and one quantitative variable (usually outcome)
- Produce univariate non-graphical statistics for the quantitative variables separately for each level of the categorical variable
- Compare the statistics across levels of the categorical variable

Univariate statistics by category

```
> df %>%
+   group_by(admin1) %>%
+   summarise(maxfat = max(fatalities)) %>%
+   arrange(desc(maxfat))
```

```
# A tibble: 11 x 2
   admin1       maxfat
   <chr>         <dbl>
 1 Sofala           58
 2 Cabo Delgado     26
 3 Manica           24
 4 Zambezia         10
 5 Maputo            7
 6 Nampula           7
 7 Inhambane         6
 8 Maputo City       4
```

Correlation and covariance

For two quantitative variables, the commonly used statistics of interest are the sample covariance and/or sample correlation.

- Covariance
  - a measure of how much two variables (X and Y) co-vary, i.e., how much and in what direction does one variable change when the other changes

$$\mathrm{Cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

  - Positive covariance suggests that when one variables is above the mean the other will probably also be above the mean, and vice versa
  - Negative covariance suggest that when one variable is above its mean, the other is below its mean.
  - Covariance near zero suggest that the two variables vary independently of each other

Exploratory data analysis

Multivariate non-graphical EDA – Correlation and covariance

Covariance

```
> cov(df$latitude, df$fatalities)
[1] 2.540272
```

- Correlation
  - Easier to interpret than covariance
  - Ranges from -1 to 1, and often denoted by r

$$r = r_{x,y} = \mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{s_x s_y}$$

  - -1 is “perfect” negative linear correlation and 1 “perfect” positive
  - 0 indicates that X and Y are uncorrelated
  - independence implies zero correlation, but the reverse is not necessarily true

Exploratory data analysis

Multivariate non-graphical EDA – Correlation and covariance

Correlation

```
> cor(df$latitude, df$fatalities)
```

Covariance and correlation matrices

When there are many quantitative variables the most common non-graphical EDA technique is to calculate all of the pairwise covariances and/or correlations and assemble them into a matrix.

Note

$$\text{Cov}(X, X) = Var(X) = s_x^2 \text{ and } r_{x,x} = Cor(X, X) = 1$$

$$\begin{array}{c|ccc} & X & Y & Z \\ \hline X & 9.3 & -4.5 & 2.5 \\ Y & -4.5 & 5.5 & -1.5 \\ Z & 2.5 & -1.5 & 3.7 \end{array} \qquad \begin{array}{c|ccc} & X & Y & Z \\ \hline X & 1 & -0.6 & 0.4 \\ Y & -0.6 & 1 & -0.3 \\ Z & 0.4 & -0.3 & 1 \end{array}$$

Covariance matrix   Correlation matrix

Exploratory data analysis

Multivariate non-graphical EDA – Correlation and covariance

Correlation and covariance matrices

```
> m <- as.matrix(df %>% select(year, fatalities, latitude),
> cov(m)
> cor(m)
```

Multivariate graphical EDA

Few useful techniques for two categorical variables, the commonly used is a grouped bar plot with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

Univariate graphs by category

- Plots of a quantitative variable (usually dependent) for each category of a qualitative variable (usually explanatory)
- Side-by-side boxplots are the most commonly used for examining the:
  - relationship between a categorical variable and a quantitative variable
  - distribution of the quantitative variable at each level of the categorical variable

Exploratory data analysis

Multivariate graphical EDA

Barplot

```
> gr <- df %>%
+   group_by(admin1) %>%
+   summarise(deaths = sum(fatalities))
> barplot(gr$deaths)
```

Barplot

```
> gr <- df %>%
+   group_by(admin1) %>%
+   summarise(deaths = sum(fatalities))
> barplot(gr$deaths, names.arg = gr$admin1, las = 3)
```

Barplot

```
> gr <- df %>%
+   group_by(admin1) %>%
+   summarise(deaths = sum(fatalities))
> barplot(gr$deaths, names.arg = gr$admin1, las = 3)
```

Exploratory data analysis

Multivariate graphical EDA

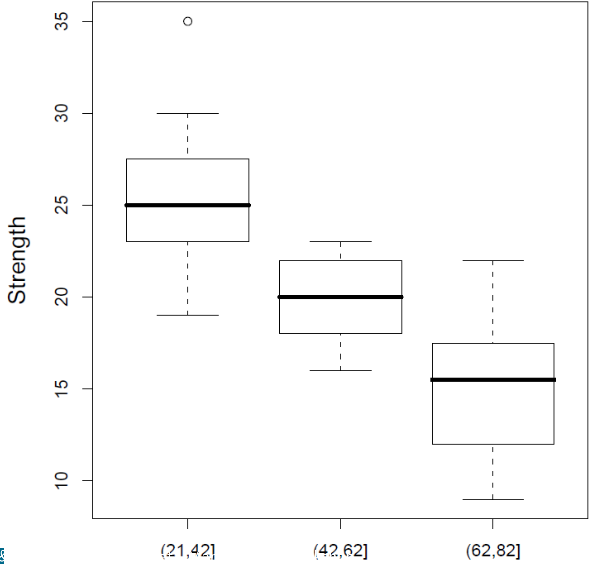
Barplot

```
> gr <- df %>%
+   group_by(admin1) %>%
+   summarise(deaths = sum(fatalities))
> ggplot(data = gr,
+        aes(x = admin1,
+            y = deaths)) +
+   geom_bar(stat = 'identity') +
+   theme(axis.text.x = element_text(angle = 90))
```

Grouped barplot

```
> gr <- df %>%
+   group_by(admin1, event_type) %>%
+   summarise(deaths = sum(fatalities))
> ggplot(data = gr,
+        aes(x = admin1,
+            y = deaths,
+            group = event_type,
+            fill = event_type)) +
+   geom_bar(stat = 'identity') +
+   theme(axis.text.x = element_text(angle = 90))
```

Multivariate graphical EDA – Side-by-side boxplot



Exploratory data analysis

Multivariate graphical EDA

Side by side boxplot

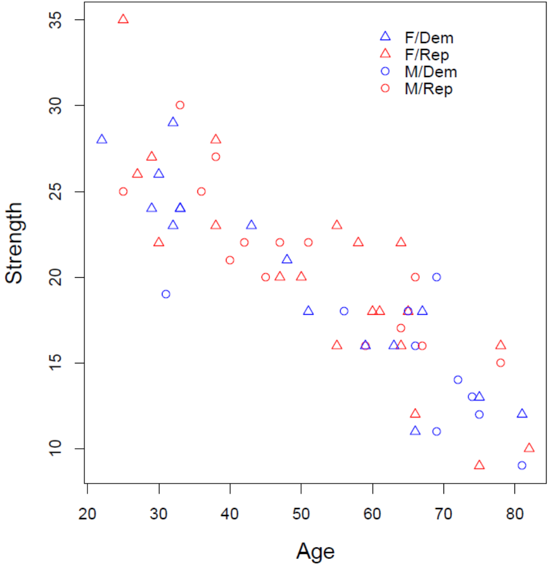
```
> boxplot(fatalities ~ admin1, data = df)
```

Scatterplot

- For two quantitative variables, one variable on the x-axis (usually the explanatory), the other on the y-axis (usually the outcome) and a point for each observation.
- One or two additional categorical variables can be accommodated by encoding the additional information in the symbol type and/or colour.

Exploratory data analysis

Multivariate graphical EDA – Scatterplot



Scatterplot

```
> plot(df$year, df$fatalities)
> plot(df$latitude, df$fatalities)
> ggplot(data = df,
+        aes(x = latitude,
+            y = fatalities)) +
+   geom_point()
```

Exploratory data analysis
Advice

You should always perform EDA before any CDA

Don't limit yourself to the methods described here, use whatever techniques necessary to:

- become more familiar with the data
- check for obvious mistakes
- learn about variable distributions
- learn about relationships between variables

Confirmatory data analysis

Confirmatory data analysis (CDA)

Based on statistical inference (the process of generalizing results and making predictions about a population based on a sample).

Population is the set all of subjects that are under investigation or whose properties are to be determined.

- A descriptive measure computed from data of an entire population is called parameter, e.g., population mean (μ), variance (σ^2).

Sample is a portion (subset) of a population.

- A descriptive measure computed from data of a sample is called statistic, e.g., sample mean ($\bar{x}$), variance (s^2).

Confirmatory data analysis (CDA)

Let's get some new data.

```
> # www.databrew.cc/planes.csv
> planes <- read_csv('~/Desktop/planes.csv')
```

Confirmatory data analysis

Intro – Probability

Tools of probability are required to make inferences about an unknown truth based on sample data.

Probability measures the uncertainty associated with occurrence of outcomes or events.

- E.g., in a population composed of 200 professionals, of which 30 are medical doctors (MD), 70 data managers (DM) and 100 lab technicians (LT), the probability of randomly selecting (assuming a person has only one profession):
  - a medical doctor is $\text{P}(MD) = \frac{30}{200} = 0.15 = 15\%$
  - a lab technicians is $\text{P}(LT) = \frac{100}{200} = 0.5 = 50\%$
  - a MD or LT is $\text{P}(MD \cup LT) = \frac{30+100}{200} = \frac{130}{200} = 0.65 = 65\%$
  - a professional is $\text{P}(MD \cup DM \cup LT) = \frac{30+70+100}{200} = 1 = 100\%$
  - a statistician is $0/200 = 0$

Confirmatory data analysis

Intro – Probability

```
> # Define flight distance
> distance <- 500
> numerator <- sum(planes$fatalities_00_14)
> denominator <- sum(planes$avail_seat_km_per_week) / dista
> numerator/denominator
```

Probability distribution

Theoretical distribution are used to describe variables of interest.

Discrete distributions

A probability mass function (PMF) p assigns to each possible realisation x of a discrete random variable X the probability $p(x)$, i.e. $P(X = x)$. The commonly distributions:

- Binomial – X is the number of successes in n independent trials, each of which has success probability p.
- Multinomial – each of n trials can result in one of k different values which occur with probabilities $(p_1, p_2, \ldots, p_k)$, where $p_1 + p_2 + \ldots + p_k = 1$, then $X = (X_1, X_2, \ldots, X_k)$.
- Possion – X represents counts which have no theoretical upper limit, e.g., number of car accidents occurring in a day.

Continuous distributions

Since for continuous random variables, $P(X = x) = 0$, PMF is useless.

Distribution is now specified by representing probabilities as areas under a curve of the probability density function (PDF) of X, i.e.,

$$P(a < X \leq b) = \int_a^b f(x)dx$$

The commonly distributions:

- Normal (Gaussian)
  - symmetric, bell-shaped PDF curve, characterized by mean (μ) and variance (σ^2), its support are all real numbers
  - any X normally distributed can be converted to the standard normal Z (i.e., with $\mu = 0$ and $\sigma = 1$) by

$$z = \frac{x - \mu}{\sigma}$$

- t
  - similar in shape to the normal, with *degrees of freedom* (df $= n - 1$), tends to normal as n increases
  - arises commonly when evaluating how far a sample mean is from a population mean when the standard deviation of the sampling distribution is unknown and estimated from the data
- Chi-square (χ^2)
  - df, support: positive real numbers
  - a χ^2 distribution with df equal to m, commonly arises from the sum of squares of m independent standard normal (Z) random variables
- F
  - numerator and denominator df (v,w), support: positive real numbers
  - if X and Y are two independent χ^2 random variables with v and w df, then $\frac{X/v}{Y/w}$ defines a new random variable that follows the F-distribution with v and w df.

Confirmatory data analysis

Intro – Probability distribution

Example Normal distribution

Normal distribution

```
> data <- rnorm(n = 1000)
> hist(data)
```

Population, sample and sampling distributions

Distribution describes the central tendency, spread and shape of the:

- N population values – population distribution
- n values sampled from the population – distribution of the sample
- summary statistics of all possible samples of size n taken from the population – sampling distribution

Common sampling distributions

The most frequently used sampling distribution are of the:

- sample mean
- sample proportion
- difference between two sample means of samples from two populations
- difference between two sample proportions of samples from two populations

Confirmatory data analysis

Intro – Probability distribution

Sampling distribution of the mean

- The sample mean is the statistic of interest.
- Distribution describes probability associated with every possible sample mean.
- The probability of the observed sample means vary, with some having high and other low.

Central limit theorem (CLT)

Given a population of any continuous distribution with mean μ and standard deviation σ, the sampling distribution of the sample mean $(\bar{X})$ is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$, when the sample size (n) is large.

Statistical inference methods

Inferential methods grouped into estimation and hypotheses testing.

Estimation

- Point estimation
  - A sample sample statistic is an estimator of a population parameter, the value of the statistic for a particular sample is an estimate
  - Point estimate is the best estimate of the true population parameter based on data, and it should ideally have minimal
    - bias – the difference between its expected value and the true parameter
    - standard error (SE) – standard deviation of the sampling distribution, calculated as $\sigma/\sqrt{n}$ or, in case of unknown σ, $s/\sqrt{n}$, for proportion $s = \sqrt{\hat{p}(1-\hat{p})}$

- Interval estimation
  - Confidence interval (CI) is an interval surrounding point estimate that expresses the uncertainty or variability associated with the estimate. It should ideally have
    - high confidence level $(1 - \alpha)$ – the probability that the interval contains the true population parameter value. The level of $(1 - \alpha) \times 100\%$ with $\alpha = 0.05$, i.e., 95% is the most commonly used
    - low width (δ) – the difference between the point estimate and one of the bounds of the CI, generally calculated as $z_{\alpha/2}\text{SE}$ or, in case of unknown σ $t_{\alpha/2,df}\text{SE}$

Confirmatory data analysis

Statistical inference methods – Estimation

```
> frangos <- read_csv('~/Desktop/frangos.csv')
> plot(frangos$days, frangos$grams)
> mean(frangos$grams[frangos$days >= 9 &
+                      frangos$days <= 11])
```

Commonly used point estimates

Single population

- The population mean is estimated by the sample mean
- The population proportion is estimated by the sample proportion

Two independent populations

- The difference between two population means is estimated by the difference between two sample means
- The difference between two population proportions is estimated by the difference between two sample proportions

Confidence interval (CI) estimation

A 95% CI for a true but unknown population parameter is

$$\left[\text{point estimate} - z_{0.05/2} \times \text{SE}; \text{point estimate} + z_{0.05/2} \times \text{SE}\right]$$

Interpretation

- If we would repeatedly draw random samples of the same size from the same population, 95% of the intervals would include the true population parameter.
- We are 95% confident that the interval contains the true population parameter.

Sample size for estimation

To calculate the minimum required sample size for estimation:

1. Specify the desired confidence level: $1 - \alpha$
2. Specify the desired width of the CI: $\delta = z_{\alpha/2}\sigma/\sqrt{n}$
3. Assume a value for the population standard deviation σ (from previous studies or conduct a pilot study)
4. Solve for n, i.e.,

$$n = \left(\frac{z_{\alpha/2}\sigma}{\delta}\right)^2$$

Hypothesis testing

Rationale

- We want to make a decision concerning a population by examining a sample
- We collect data and check how likely or unlikely is the observed data if the hypothesis is true
- We may reject or not reject the hypothesis based on the data

Hypothesis testing

Rationale

- We want to make a decision concerning a population by examining a sample
- We collect data and check how likely or unlikely is the observed data if the hypothesis is true
- We may reject or not reject the hypothesis based on the data

Steps

- Select the probability model for the observed data (distribution)
- Set up a null hypothesis (H_0) and alternative hypothesis (H_1)
  - two-sided test
  - one-sided test
- Select a test statistic (Z, t)
- Choose the significance level (α), $\alpha = 0.05$ is commonly used; decision rule and critical region (rejection region)
- Compute the value of the test statistic from the observed data
- Make a statistical decision and conclusion

(illustration of critical region)

Statistical decision

- Reject H_0 in favour of H_1 because the test statistic is very unlikely under H_0 (test statistic falls in the rejection region; p-value $\leq \alpha$)
- Fail to reject H_0 because the test statistic is likely under H_0 (test statistic falls in the acceptance region; p-value $> \alpha$)

Confirmatory data analysis

Statistical inference methods – Hypothesis testing

```
> # Hypothesis = chickens fed by corn are bigger than chick
> fit <- lm(grams ~ diet + days, data = frangos %>% filter(
> summary(fit)
```

Possible errors

- Type I error –reject a true H_0
  - $\alpha =$ probability of Type I error
- Type II error – fail to reject a false H_0
  - $\beta =$ probability of Type II error
  - $1 - \beta =$ Power $=$ probability of rejecting a false H_0

Confirmatory data analysis

Statistical inference methods – Hypothesis testing

p-value

The probability of obtaining the test statistic that is equal to or more extreme (against H_0) than the observe one when H_0 is true.

- p-value $\leq \alpha$: statistically significant result (reject H_0)
- p-value $> \alpha$: not stat. significant result (fail to reject H_0)
- The final conclusion should also take in the consideration the clinical/practical significance!

Test statistic

Generally it is calculated as

$$\text{test statistic} = \frac{\text{sample statistic} - \text{hypothesized value}}{\text{standard error of the sample statistic}}$$

One group

Test statistic for mean

$H_0 : \mu = \mu_0$ versus $H_0 : \mu \neq \mu_0$

- If σ is known:
$$z = \frac{(\bar{x} - \mu_0)}{\frac{\sigma}{\sqrt{n}}}$$
- If σ is unknown:
$$t = \frac{(\bar{x} - \mu_0)}{\frac{s}{\sqrt{n}}}$$

Test statistic for proportion

$H_0 : p = p_0$ versus $H_0 : p \neq p_0$

-
$$z = \frac{(\hat{p} - p_0)}{\sqrt{\frac{(\hat{p} - p_0)}{n}}}$$

Two groups

Means

$H_0 : \mu_1 - \mu_1 = 0$ versus $H_0 : \mu_1 - \mu_2 \neq 0$

- Calculate test statistic (z if population variances are known in both groups, and t otherwise):

$$t = \frac{(\bar{x}_1 - \bar{x}_2 - 0)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \text{ replace } s \text{ by } \sigma \text{ for z test}$$

  degrees of freedom for t: $\text{df}{=}(n_1 - 1) + (n_2 - 1)$

- Calculate the $(1 - \alpha) \times 100\%$ CI for the true difference in population means:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- The p-value and CI lead to the same conclusion, i.e.,
  - p-value $\leq \alpha$ implies that the $(1 - \alpha) \times 100\%$ CI does not

- The *p*-value (from the test statistic) and CI lead to the same conclusion, i.e.,
    - *p*-value $\leq \alpha$ implies that the $(1-\alpha)\times 100\%$ CI does not contain the H_0 value (0), and vice-versa

More on statistical inference methods

- There a lots of other inferential methods applicable for lots of different scenarios not covered here.
- Regression models allow for more complex analysis: assess and quantify the effects (main and interactions) of many explanatory variables on an outcome variable simultaneously, make outcome prediction based on values of explanatory variables.
- The sample size calculation for hypothesis testing tends to be more complex for complex inference methods, in some cases there are no mathematical formulae (simulations are used instead), but in all the sample size is estimated such that at significance level α, the method used for inference will have the power of $1 - \beta$ to detect the effect of a specified size.

Framework – Frequentist vs Bayesian

The inference methods describe here fit into Frequentist framework

Frequentist framework

...

Bayesian framework

...

Thank you!