Introduction to Statistical Inference

Marta Bofill Roig

March 9, 2025

GRBIO

UPC

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Outline

Introduction

Probability concepts

Statistics concepts

Wrap up with examples

Welcome! Let's introduce ourselves

About me...

- Serra Húnter Fellow (tenure-track lecturer) at the Department of Statistics and Operations Research.
- Member of the GRBIO (Research group on Biostatistics and Bioinformatics).
- Research interests: Clinical trials and Survival analysis.

About you...

- Background
- Interests
- Goals that you expect to achieve with the course

Probability concepts

Random variables

Most scientific questions revolve around the understanding of phenomena

- Why do we have seasons?
- What will be the weather tomorrow?
- How does the age of patients affect their blood pressure?

To **understand** phenomena we need to **measure** them.

Random Variables (cont'd)

- A **variable** is a quantification of a phenomenon
  - i.e., we assign a numeric value to an observable event
- We can have two types of phenomena **Deterministic** and **Stochastic**
  - **Deterministic**: future values of the phenomenon can be exactly predicted from current conditions
  - **Stochastic**: future values of the phenomenon cannot be exactly predicted from current conditions

Random Variables (cont'd)

Some examples

- Why do we have seasons?
  - the earth circles around the sun on a tilted axis
  - this movement remains the same every year

    → Deterministic phenomenon
- How the age of patients affects their blood pressure?
  - as you age, the vascular system changes
  - but we cannot predict the exact blood pressure of a person based on his/her age

    → Stochastic phenomenon

Random Variables (cont'd)

A **random variable** is a numeric quantification of a stochastic phenomenon

Examples

- Blood pressure
  - Phenomenon: circulating blood pressures the walls of blood vessels
  - Random variable: the numeric value in mm Hg of this pressure
- Asthma attack
  - Phenomenon: tightening of muscles around the airways
  - Random variable: '1' occurrence of this phenomenon, and '0' otherwise

Random Variables (cont'd)

We have two types of random variables: **Continuous** and **Discrete**.

- **Continuous** random variables take an uncountable number of possible values
  - E.g., cholesterol levels, BMI, weight, ...
- **Discrete** random variables take a countable number of possible values
  - death from cancer, 'yes' or 'no' (**dichotomous**)
  - none, mild, moderate and severe symptoms (**ordinal**)
  - patients' race (**nominal**)
  - the number of asthma attacks in a period (**count**)

Random Variables (cont'd)

Notation:

- We typically denote random variables with uppercase Latin letters, e.g., Y, X, T, etc.
- With the same corresponding lowercase letter we denote the realizations (i.e., observed values) of these random variables, e.g., y, x, t, etc.

Example:

Let X denote the random variable describing the blood pressure phenomenon, the specific value we observe when we measure the blood pressure is denoted by x.

Random Variables (cont'd)

- The different types of random variables contain different amount of information

$$\text{dichotomous} < \text{ordinal/nominal} < \text{count} < \text{continuous}$$

- I know more about blood pressure if I know its exact value than only knowing that it was below, e.g., 140 mm Hg
  - a blood pressure of 138 mm Hg is different than 110 mm Hg, even though both are below the limit

Note:

This is why, in general, it is not a good idea to categorize continuous variables → by categorizing we lose information!

Distribution Functions

Our aim is to **understand** stochastic phenomena.

- Challenging because we cannot exactly predict them.

Despite the random nature of stochastic phenomena, often there are patterns in randomness,

- i.e., some values of random variables are more probable than others

Distribution Functions (cont'd)

Probability is a numerical description of how likely an event is to occur.

Properties:

- it is constrained between 0 and 1
- 0 indicates impossibility of the event
- 1 indicates certainty

Distribution Functions (cont'd)

We give probabilities of occurrence to all different possible values of a random variable (that correspond to different possible outcomes of the phenomenon under study). This collection of probabilities define the **probability distribution** of the random variable.

Example:

- We toss a fair coin and denote by X the random variable of the possible outcomes

$$X = \begin{cases} 0, & \text{if tails} \\ 1, & \text{if head} \end{cases}$$

- The distribution of X

$$P(X = x) = \begin{cases} P(X = 0) = 0.5, & \text{if tails} \\ P(X = 1) = 0.5, & \text{if head} \end{cases}$$

Distribution Functions (cont'd)

We have some key functions to describe distributions

The **cumulative distribution function** (CDF) denotes the probability that a random variable X takes a value less or equal to x

$$F_X(x) = P(X \leq x)$$

Properties:

- it is constrained between 0 and 1
- it is an increasing function of x
- it is defined for both continuous and discrete random variables

Distribution Functions (cont'd)

The **probability mass function** denotes the probability that a discrete random variable X takes a particular value x

$$p_X(x) = P(X = x)$$

Properties:

- it is constrained between 0 and 1
- the sum of ht e probabilities for all possible values of X is one

$$\begin{aligned} \sum_x p_X(x) &= P(X=0) + P(X=1) + P(X=2) + \dots \\ &= 1 \end{aligned}$$

Distribution Functions (cont'd)

The **probability density function** (PDF) is used to specify the probability that a continuous random variable X takes values in particular interval (a, b)

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

Properties:

- it is non-negative
- the integral over all possible values of X is one

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1$$

Distribution Functions (cont'd)

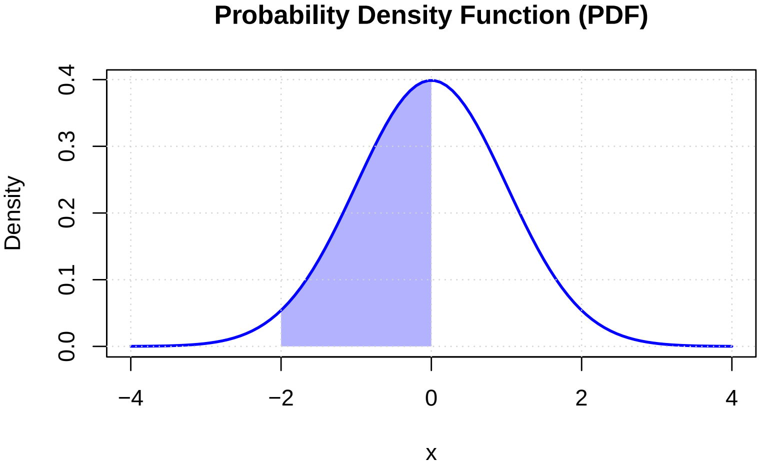
For continuous random variables the cumulative distribution function and the probability density function are linked via

$$f_X(x) = \frac{dF_X(x)}{dx}$$

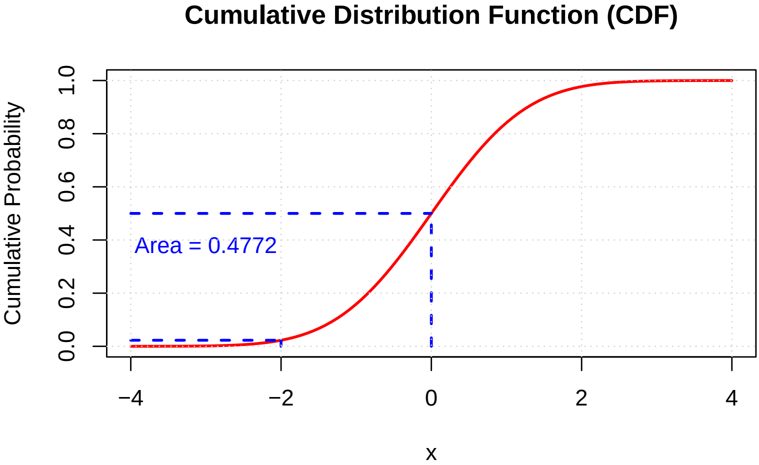
and

$$F_X(x) = \int_{-\infty}^{x} f_X(s)ds$$

Distribution Functions (cont'd)



Distribution Functions (cont'd)



Expectation and Quantiles

We have seen how we can describe the whole distribution of random variables using the

- probability mass function
- probability desity function
- cumulative distribution function

Most often we would like to summarize the distribution by some representative quantities.

Expectation and Quantiles (cont'd)

The **expected value** of a random variable X is the mean of its distribution

$$E(X) = \begin{cases} \sum_x x P(X = x), & \text{if } X \text{ is discrete} \\ \int x f_X(x) dx, & \text{if } X \text{ is continuous} \end{cases}$$

- the expected value is a **weighted** average of the random variable's values
- weighted because different values have different probabilities of occurrence
- (for continuous random variables different **intervals of values** have different probabilities)

Expectation and Quantiles (cont'd)

Example: We are interested in the severity of complications after a surgery

- we denote by X the random variable of the possible outcomes

$$X = \begin{cases} 0, & \text{if no complications} \\ 1, & \text{if mild complications} \\ 2, & \text{if moderate complications} \\ 3, & \text{if severe complications} \end{cases}$$

- the distribution of X

$$P(X = x) = \begin{cases} P(X = 0) = 0.4, & \text{if no complications} \\ P(X = 1) = 0.3, & \text{if mild complications} \\ P(X = 2) = 0.1, & \text{if moderate complications} \\ P(X = 3) = 0.2, & \text{if severe complications} \end{cases}$$

Expectation and Quantiles (cont'd)

- What is the expected values of X?

$$\begin{aligned} E(X) &= \sum_x xP(X=x) \\ &= 0\cdot P(X=0)+1\cdot P(X=1)+2\cdot P(X=2)+3\cdot P(X=3) \\ &= 0\cdot 0.4+1\cdot 0.3+2\cdot 0.1+3\cdot 0.2 \\ &= 1.1 \end{aligned}$$

Expectation and Quantiles (cont'd)

- The expected value is a **location** measure of the distribution of a random variable X
- The expected value gives us some information where the mean of the distribution is located
- Another set of useful location measures is the **quantiles** of the distribution

Expectation and Quantiles (cont'd)

The k-th **quantile** of a random variable X is the value below which a given $k\%$ of values in its distribution fall

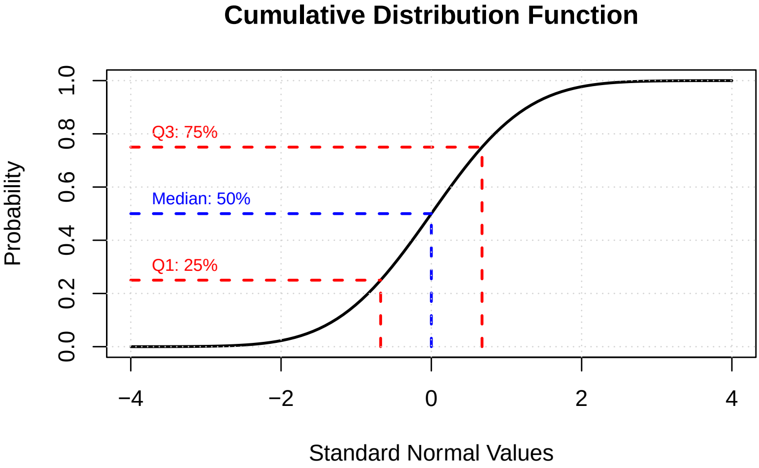
$$Q_X(k) = \{x : F_X(x) = k\},\ 0 \leq k \leq 1$$

i.e., the value x for which the CDF equals k.

The most-used quantiles are:

- **median**: the value x below which 50% of the observations fall
- **1st quartile**: the value x below which 25% of the observations fall
- **3rd quartile**: the value x below which 75% of the observations fall

Expectation and Quantiles (cont'd)



Variance

- The expected value and the quantiles give us information about the location of the distribution of a random variable.
- Another important quantity is the spread of the distribution, i.e., how far away are the values of a random variable located from each other.

Variance (cont'd)

The **variance** of a random variable X measures how far its values are spread out from their mean (i.e., expected value)

$$Var(X) = E[(X - E(X))^2] = \begin{cases} \sum_x \{x - E(X)\}^2 P(X = x), & \text{if } X \text{ is discrete} \\ \int \{x - E(X)\}^2 f_X(x) dx, & \text{if } X \text{ is continuous} \end{cases}$$

Variance (cont'd)

Notes:

- the variance is always positive
- the reason why we compute squared differences is because otherwise the positive and negative differences would cancel out
- the fact that we calculate square differences means that the variance is on the squared scale of the random variable
- this is why most often we also calculate the **standard deviation**

$$sd(X) = \sqrt{Var(X)}$$

Covariance and Correlation

- The variance measures how far a set of numbers is spread out for a single random variable.

- However, often we are interested in measuring the spread of pairs of random variables, and how these spreads are associated with each other, e.g.,

  - how are changes in blood pressure associated with changes in age?

  - how are changes in cholesterol associated with changes in BMI?

Covariance and Correlation (cont'd)

The **covariance** is a measure of how much two random variables change together

$$cov(X,Y) = E[(X - E(X)) \cdot (Y - E(Y))]$$

Covariance and Correlation (cont'd)

Notes:

- it can be positive or negative
  - if the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values, the covariance is positive
  - in the opposite case, the covariance is negative
- the magnitude of the covariance is not easy to interpret because it depends on the magnitudes of the variables
- the variance is a special case of the covariance, $Var(X) = cov(X, X)$

Covariance and Correlation (cont'd)

The **correlation** is a standardized version of the covariance and is a measure of the linear correlation (dependence) between two variables

$$corr(X,Y) = \frac{cov(X,Y)}{sd(X)sd(Y)}$$

Properties:

- it is constrained between -1 and 1
- 0 means no correlation between the two variables
- -1 means perfect negative correlation
- 1 mean perfect positive correlation

Covariance and Correlation (cont'd)

Notes:

- The correlation is a measure of **linear** association.
- Two variables may have zero linear correlation but still be (strongly) associated.

Standard Distributions

- We have seen that the distribution of a random variable describes in a generic manner the probability of certain events.
- Often we use distributions that place some restrictions on their shape
  - these distributions have **parameters** that control key quantities of the distribution,
  - typically, the mean and variance of the distribution,
  - parameters are typically denoted with Greek letters.

Standard Distributions (cont'd)

The **normal (Gaussian) distribution** has a probability density function given by the equation:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$$

Notes:

- We write $X \sim N(\mu, \sigma^2)$ to denote that X follows the normal distribution.
- The parameter μ denotes the mean of the distribution (i.e., $E(X) = \mu$).
- The parameter σ^2 denotes the variance of the distribution (i.e., $var(X) = \sigma^2$).
- It is a symmetric distribution around μ, so the median is also μ.
- (Note that here π is not a parameter but the constant $\pi = 3.1415...$).

Standard Distributions (cont'd)

- It describes phenomena for which the majority of the observations is located around the mean, and as we get further away from the mean, fewer and fewer observations are to be found.
- It plays a very central role in the analysis of continuous random variables.

Standard Distributions (cont'd)

The **binomial distribution** has a probability mass function given by the equation

$$p_X(x) = \binom{N}{x} \pi^x (1-\pi)^{N-x}$$

Notes:

- We write $X \sim Bin(N, \pi)$ to denote that X follows the binomial distribution.
- It describes the number of 'successes' out of N independent trials, where the probability of success of each trial is π.
- The mean is $E(X) = N\pi$ and the variance is $var(X) = N\pi(1-\pi)$.
- (Note: the first term is the binomial coefficient giving the number of ways x successes can be distributed in N trials)

Standard Distributions (cont'd)

Notes:

- When we have one trial, i.e., $N = 1$ we get the **Bernoulli distribution**.
- It plays a very central role in the analysis of dichotomous and ordinal random variables.

Standard Distributions (cont'd)

The **Poisson distribution** has a probability mass function given by the equation

$$p_X(x) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

Notes:

- We write $X \sim Pois(\lambda)$ to denote that X follows the Poisson distribution.
- It describes the number of 'events' that occur in a given period of time.
- The mean and variance are equal to λ, i.e., $E(X) = var(X) = \lambda$.
- It plays a very central role in the analysis of count variables.
- (Note: $x!$ is the factorial, e.g., $5! = 1 \times 2 \times 3 \times 4 \times 5$).

Statistics concepts

Population and Sample

The aim is to **understand** a phenomena

Research questions:

- will the new treatment for hypertension work better than the standard one?
- which are risk factors for IC admission due to COVID-19?
- are genetic factors related to the onset of breast cancer?
- ...

Population and Sample (cont'd)

These questions are generically formulated: *For which specific patients are we talking about?*

Research questions: will the new treatment for hypertension work better than the standard one?

- ... patients older than 50 years old
- ... who had blood pressure higher than 160 mm Hg for two consecutive days
- ... no family history of hypertension

Population and Sample (cont'd)

The **target population** is the precise definition of the total group of individuals for whom we want to draw conclusions.

- This is achieved by formulating the **inclusion criteria** for the study

Ideally, we collect data from the whole population (i.e., from all subjects), and proceed to analyze them:

- data are actually the realizations from the random variables of interest,
- e.g., blood pressure measurements

Population and Sample (cont'd)

However, the problem is that it is infeasible to collect data from all subjects in the population:

- simply because the population contains too many subjects

To proceed, we work with a sample (i.e., a small subset) from the population.

A well-chosen sample will contain most of the information about a particular population characteristic.

Population and Sample (cont'd)

When all subjects from the population have the same chance to be included in the sample we obtain a **random sample**.

- Such a sample is guaranteed to provide us with valid statements about the target population.

However, most often, we cannot take a random sample but rather a "convenience" sample, e.g.,

- subjects from a hospital's registry
- subjects from a specific area
- ...

Population and Sample (cont'd)

Problems with non-random samples

- (academic) hospital patients are not the same as the ones seen in the community
- patients who return questionnaires are different than those who do not
- ...

These problems can lead to **sampling bias**, i.e., the results of the analysis can be wrongly attributed

Population and Sample (cont'd)

To be able to make generalizations from our "convenience" sample, we want it to be sufficiently **representative** of the target population.

A **representative sample** is a group of subjects from the target population that adequately replicates the population according to whatever characteristic or quality is under study.

A representative sample parallels key variables and characteristics of the larger population, e.g., sex, age, education level, socioeconomic status, etc.

Statistical inference refers to the use of statistics to draw conclusions about an unknown aspect of a population based on a random sample.

Estimation and Sampling Variability

The fact that we can only work with a (representative) sample from our target population causes an important complication.

Sampling Error: There will be a difference between the characteristic we measure in the sample and the same characteristic in the population.

Note: a larger sample size reduces the likelihood of sampling errors and increases the likelihood that the sample accurately reflects the target population.

Estimation and Sampling Variability (cont'd)

Sampling variability is the variability in the analysis results caused by the fact that we work with the sample and not the whole population.

- We often work with a particular study/sample.
- However, this is just one sample from our target population.
- In principle, we could take many different samples from the population.
- Each sample would yield different results.

Estimation and Sampling Variability (cont'd)

Let's return to our research questions:

- will the new treatment for hypertension work better than the standard one?

But at whom is this question targeted?

- our specific sample
- or our target population?

If we want to draw some conclusions from the sample at hand regarding the population, we need to quantify and account for the sampling variability.

Estimation and Sampling Variability (cont'd)

But how can we say anything about the variability from different samples, given that we have only a single sample at hand?

Under some assumptions and the statistical theory, we can determine the magnitude of sampling variability of samples such as our own, but based only on the data in our single sample.

Estimation and Sampling Variability (cont'd)

Example:

- what is the average blood pressure for a specific group of patients?
- let's formulate this question more precisely
    - the characteristics of the group define our target population
    - we denote by X the random variable describing the blood pressure values
    - this random variable will follow a distribution, with mean denoted by a parameter μ
    - our aim is to estimate μ

The **estimand** is the parameter of the target population we wish to estimate from a sample.

Estimation and Sampling Variability (cont'd)

Example:

- let's assume that we will obtain a representative sample from this population of size n (i.e., the number of patients in our sample)
- Important: We think generically, we do not have the sample yet! – if I will get a sample, how can I use it to estimate μ?

Estimation and Sampling Variability (cont'd)

Example:

- each subject in the sample has a random variable X_i describing his/her blood pressure levels
- we could then estimate μ using the sample mean

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

The **estimator** is a rule for estimating the parameter in the population using the data we will collect in a sample.

Estimation and Sampling Variability (cont'd)

Example:

- when we have available specific values x_i from a **realized sample**, we calculate the realized value of the sample mean

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

where x_i denotes the blood pressure measurements for patient i

The estimate of a particular population characteristic we obtain from a specific sample using an estimator is called the **point estimate**.

Estimation and Sampling Variability (cont'd)

But what would be the variability of this point estimate in all different samples of size n from this target population?

$$se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of blood pressure values in our population.

Estimation and Sampling Variability (cont'd)

The **standard error** is the standard deviation of the results from all possible samples from the target population.

- the collection of the possible results from all different samples is called the **sampling distribution**, and
- the **standard error** is the standard deviation of this distribution.

Summary

- Inference requires working from our data, through study sample and study population, to a target population.
- Problems and biases can crop up at each stage of this path.
- The best way to proceed from sample to study population is to have drawn a random sample.
- A population can be thought of as a group of individuals, but also as providing the probability distribution for a random observation drawn from that population.
- Populations can be summarised using parameters that mirror the summary statistics of sample data.

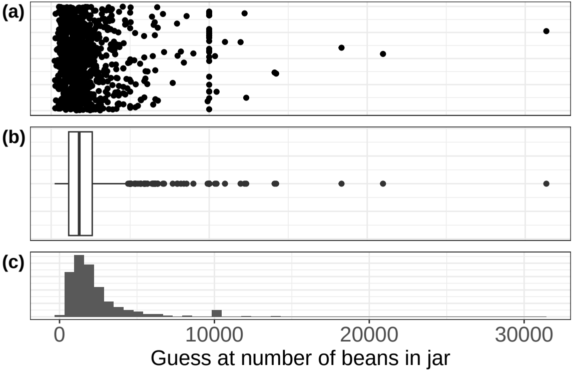
Wrap up with examples

Example 1: How many jelly beans are in this jar[1]?

Nine hundred and fifteen people provided their guesses ($n = 915$ responses).

[1] Not exactly this one, but similar. Experiment at https://sms.cam.ac.uk/media/1187673/

Three ways of presenting the pattern of the respondents



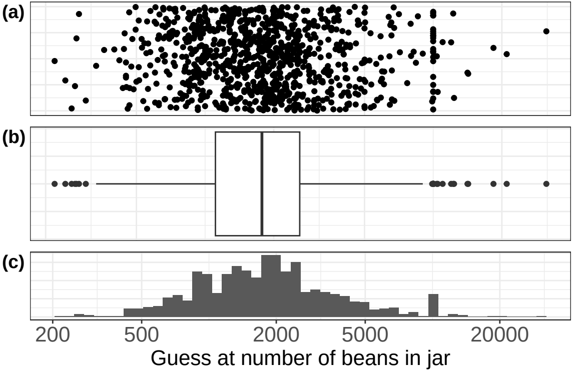
(a) **strip-chart** or **dot-diagram**: shows each data-point as a dot but each one is given a dandom jitter

(b) **box-and-whisker plot**: summarizes some essential features of the data distribution (box: Q1, Median, Q3; whisker: *min* and *max* excluding outliers)

(c) **histogram**: counts how many data-points lie in each of a set of intervals - it gives a very rough idea of the shape of the distribution

Data distribution is highly **skewed**, meaning it is **not symmetric** around some central value, and has a **long right-hand tail**.

Can we present the data in a more informative way?



- We could transform the data in a way that reduces the impact of the extremes, say by plotting it on what is called a **logarithmic scale**.
- Now the space between 100 and 1000 is the same as the space between 1000 and 10000.
- Now the figure shows a clearer pattern, with a fairly symmetric distribution and no extreme outliers.

Summary statistics

Measures of the **location** of the data distribution:

- **Mean**, the sum of the numbers divided by the number of guesses: 2409
- **Median**, the middle value when the numbers are put in order: 1775
- **Mode**, the most common value: 10000

Describing the **spread** of the data distribution:

- **Range**, minimum and maximum values: from 219 to 31337
- **Inter-quantile range**, this is the distance between the 25-th and 75-th quantiles of the data: from 1110 to 2599.5
- **Standard deviation**: 2422.2

Example 2: Child heart surgery[2]

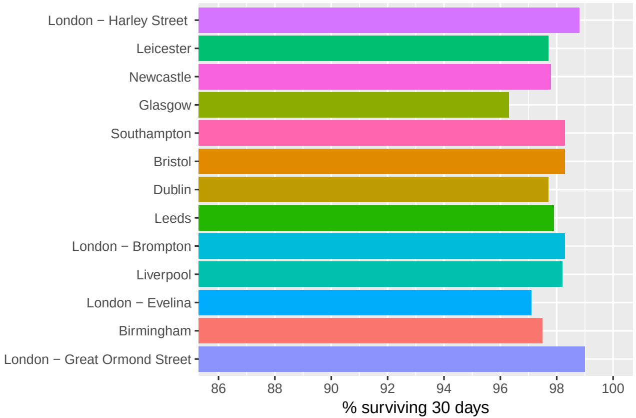
Data from thirteen hospitals from 2012 to 2015 in UK and Ireland

Hospital	Operations	Survivors	Deaths	Surviving (%)	Dying (%)
London (Harley Str.)	418	413	5	98.8	1.2
Leicester	607	593	14	97.7	2.3
Newcastle	668	653	15	97.8	2.2
Glasgow	760	733	27	96.3	3.7
Southampton	829	815	14	98.3	1.7
Bristol	835	821	14	98.3	1.7
Dublin	983	960	23	97.7	2.3
Leeds	1038	1016	22	97.9	2.1
London (Brompton)	1094	1075	19	98.3	1.7
Liverpool	1132	1112	20	98.2	1.8
London (Evelina)	1220	1185	35	97.1	2.9
Birmingham	1457	1421	36	97.5	2.5
London (Great Ormond Str.)	1892	1873	19	99.0	1.0

Deaths were counted if they occurred within 30 days of surgery; *Surviving %* denotes percentage of surviving 30 days after surgery; *Dying %* denotes percentage of dying within 30 days after surgery.

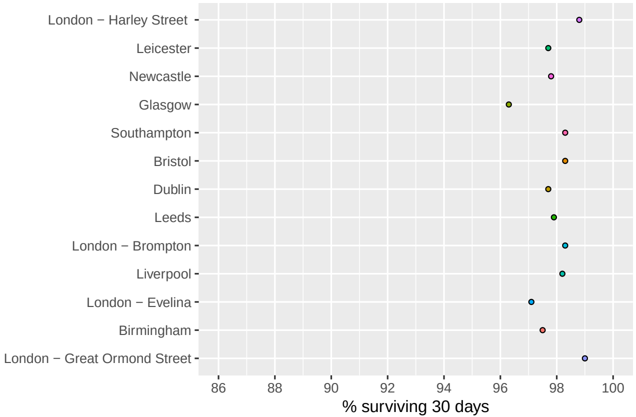
[2] Source `https://www.childrensheartsurgery.info/`

Summarising the survival rates for thirteen hospitals



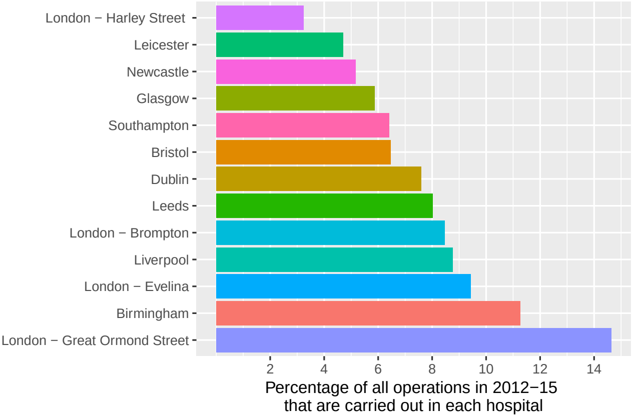
- Horizontal **bar-chart** of 30-day survival rates
- The choice of the start of the horizontal azis (here 85%) can have a crucial effect on the impression given by the plot
  - If the axis starts at 0%, all the hospitals will look indistinguishable
  - If the axis started at 95%, the differences would look misleadingly dramatic

Summarising the survival rates for thirteen hospitals (cont'd)



- Bar-chart of 30-day survival rates for thirteen hospitals represented as a **dot-plot**
- Here the non-zero axis is less important as, unlike a bar, it is not directly connected to the data-point.

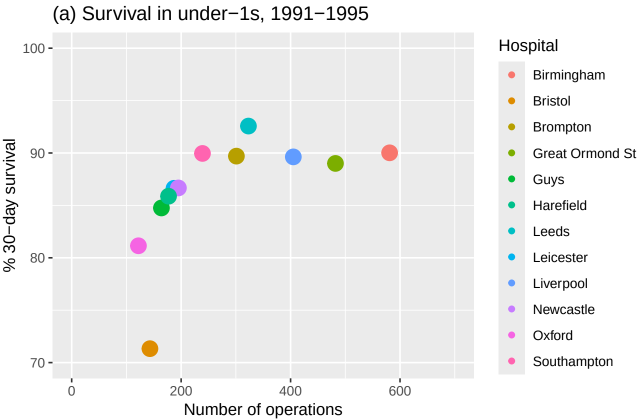
Summarising the number of operations per hospital



- Horizontal bar chart of the proportions being treated in each hospital.

Describing relationships between variables

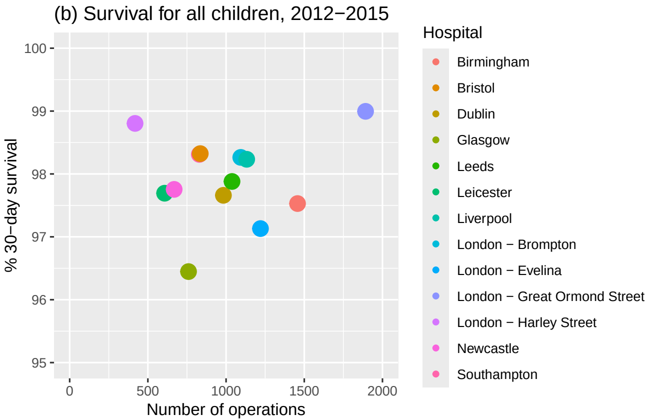
Do busier hospitals have higher survival rates?



- Scatter-plot of survival rates against number of operations in child heart surgery in 1991-1995
- Estimated correlation: 0.59

Describing relationships between variables (cont'd)

Do busier hospitals have higher survival rates?



- Scatter-plot of survival rates against number of operations in child heart surgery in 2012-201
- Estimated correlation: 0.17

Four common features of a good data visualisation[3]

1. It contains reliable information.

2. The design has been chosen so that relevant patterns become noticeable.

3. It is presented in an attractive manner, but appearance should not get in the way of honesty, clarity and depth.

4. When appropriate, it is organised in a way that enables some exploration.

[3] According to `https://www.albertocairo.com/`

Summary

- A variety of statistics can be used to summarise the empirical distribution of data, including measures of location and spread.
- Skewed data distributions are common, and some summary statistics are very sensitive to outlying values.
- Data summaries always hide some detail and care is required so that important information is not lost.
- Consider transformations to better reveal patterns, and use the eye to detect patterns, outliers, similarities and clusters.
- When exploring data, a primary aim is to find factors that explain the overall variation.

References and other materials

1. D.R. Cox, E. J. Snell. Applied Statistics - Principles and Examples. Chapman & Hall, 1981.

2. Spiegelhalter, David. The art of statistics: Learning from data. Penguin UK, 2019.

3. James, G., Witten, D., Hastie, T., Tibshirani, R. An introduction to statistical learning. Springer, 2013.

4. Teaching couses from `https://www4.stat.ncsu.edu/~davidian/`

5. Teaching couses from `https://www.drizopoulos.com/`

Thank you for the attention!

GRBIO

UPC

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH