Fundamentals of Statistical Testing

Lecture 1

Dr Milan Valášek
24 January 2022

UNIVERSITY OF SUSSEX

Housekeeping

- Welcome from Jennifer and Milan
- Familiarise yourself with Canvas and AnD website
- Come to practicals and bring laptops
- Help desk sessions from this week (sign up on Canvas)
- More info in this week's practical

- **It won't be easy so do put in the hours!**

Overview

- Recap on distributions
- More about the normal distribution
- Sampling
- Sampling distribution
- Standard error
- Central Limit Theorem

Objectives

After this lecture you will understand

- that there exist mathematical functions that describe different distributions
- what makes the normal distribution normal and what are its properties
- how random fluctuations affect sampling and parameter estimates
- the function of the sampling distribution and the standard error
- the Central Limit Theorem

With this knowledge you'll build a solid foundation for understanding all the statistics we will be learning in this programme!

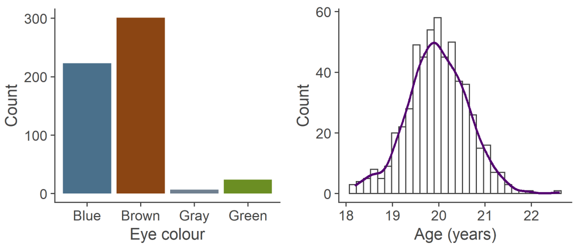
It's all Greek to me!

- μ is the *population* mean
- $\bar{x}$ is the *sample* mean
- $\hat{\mu}$ is the **estimate** of the *population* mean
- Same with *SD*: σ, s, and $\hat{\sigma}$
- Greek is for populations, Latin is for samples, hat is for population estimates

Recap on distributions

- Numerically speaking, the number of observations per each value of a variable
- Which values occur more often and which less often
- The shape formed by the bars of a bar chart/histogram

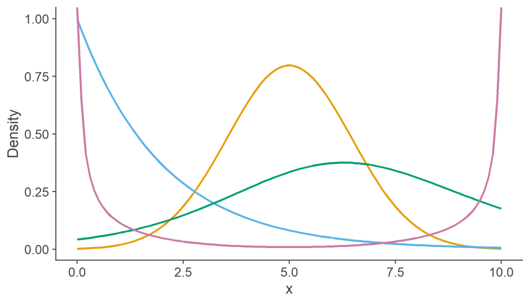
Plot | R code



Known distributions

- Some shapes are "algebraically tractable", *e.g.*, there is a maths formula to draw the line
- We can use them for statistics

Plot | R code

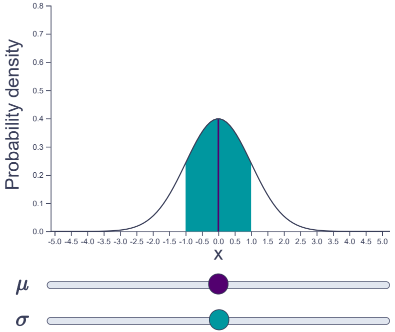


The normal distribution

- AKA Gaussian distribution, The bell curve
- The one you **need to** understand
- Symmetrical and bell-shaped
- *Not every* symmetrical bell-shaped distribution is normal!
- It's also about the proportions
  - The normal distribution has fixed proportions and is a function of two parameters, μ (mean) and σ (or *SD*; standard deviation)

The normal distribution

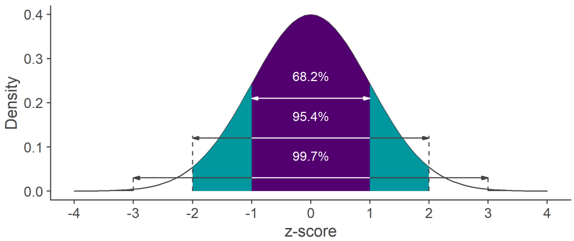
- Peak/centre of the distribution is its mean (also mode and median)
- Changing mean (**centring**) shifts the curve left/right
- *SD* determines steepness of the curve (small σ = steep curve)
- Changing *SD* is also known as **scaling**



Area below the normal curve

- No matter the particular shape of the given normal distribution, the proportions with respect to *SD* are the same
  - ~**68.2%** of the area below the curve is within **±1** ***SD*** from the mean
  - ~**95.4%** of the area below the curve is within **±2** ***SD*** from the mean
  - ~**99.7%** of the area below the curve is within **±3** ***SD*** from the mean
- We can calculate the proportion of the area with respect to any two points

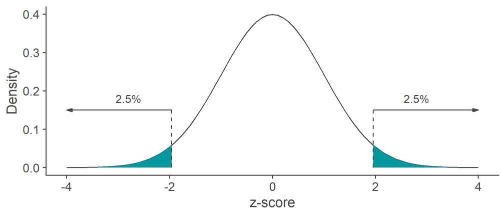
Plot | R code



Area below the normal curve

- Say we want to know the number of *SD*s from the mean beyond which lie the outer 5% of the distribution

Plot | R code




```
qnorm(p = .025, mean = 0, sd = 1) # lower cut-off
## [1] -1.959964
qnorm(p = .975, mean = 0, sd = 1) # upper cut-off
## [1] 1.959964
```

Critical values

- If *SD* is known, we can calculate the cut-off point (critical value) for **any proportion** of normally distributed data

```
qnorm(p = .005, mean = 0, sd = 1) # lowest .5%
```

```
## [1] -2.575829
```

```
qnorm(p = .995, mean = 0, sd = 1) # highest .5%
```

```
## [1] 2.575829
```

```
# most extreme 40% / bulk 60%
qnorm(p = .2, mean = 0, sd = 1)
```

```
## [1] -0.8416212
```

```
qnorm(p = .8, mean = 0, sd = 1)
```

```
## [1] 0.8416212
```

- Other known distributions have different cut-offs but the principle is the same

Sampling from distributions

- Collecting data on a variable = randomly sampling from distribution
- The underlying distribution is often assumed to be normal
- Some variables might come from other distributions
    - Reaction times: *log-normal* distribution
    - Number of annual casualties due to horse kicks: *Poisson* distribution
    - Passes/fails on an exam: *binomial* distribution

Sampling from distributions

- Samples from the same population differ from one another

```
# draw a sample of 10 from a normally distributed
# population with mean 100 and sd 15
rnorm(n = 6, mean = 100, sd = 15)
```

```
## [1] 101.61958  80.95560  89.62080  96.04378 106.40106  86.21514
```

```
# repeat
rnorm(6, 100, 15)
```

```
## [1]  80.31573 107.63193  85.82520  99.95288  93.55956  74.73945
```

Sampling from distributions

- Statistics ($\bar{x}$, s, *etc.*) of two samples will be different
- **Sample** statistic (*e.g.*, $\bar{x}$) will likely differ from the **population** parameter (*e.g.*, μ)

```r
sample1 <- rnorm(50, 100, 15)
sample2 <- rnorm(50, 100, 15)
```

```r
mean(sample1)
```

```
## [1] 98.56429
```

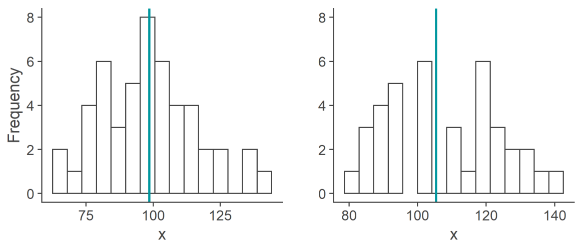
```r
mean(sample2)
```

```
## [1] 105.4175
```

Sampling from distributions

- Statistics ($\bar{x}$, s, *etc.*) of two samples will be different
- **Sample** statistic (*e.g.*, $\bar{x}$) will likely differ from the **population** parameter (*e.g.*, μ)

Plot | R code



Sampling distribution

- If we took all possible samples of a given size (say *N = 50*) from the population and each time calculated $\bar{x}$, the means would have their own distribution

- This is the **sampling distribution** of the mean

  - Approximately **normal**

  - Centred around the **true population mean**, μ

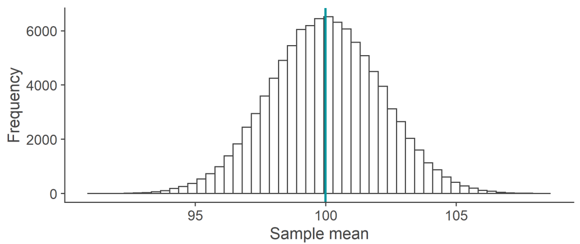
- Every statistic has its own sampling distribution (not all normal though!)

Sampling distribution

```
x_bar <- replicate(100000, mean(rnorm(50, 100, 15)))
mean(x_bar)
```

```
## [1] 99.99395
```

Plot | R code



Standard error

- Standard deviation of the sampling distribution is the **standard error**

```
sd(x_bar)
```

```
## [1] 2.122072
```

- Sampling distribution of the mean is *approximately normal*: *~68.2% of means of samples* of size 50 from this population will be *within ±2.12 of the true mean*

Standard error

- Standard error can be **estimated** from any of the samples

$$\widehat{SE} = \frac{SD}{\sqrt{N}}$$

```
samp <- rnorm(50, 100, 15)
sd(samp)/sqrt(length(samp))
```

```
## [1] 1.872102
```

```
# underestimate compared to actual SE
sd(x_bar)
```

```
## [1] 2.122072
```

- If ~68.2% of sample means lie within ±1.87, then there's a ~68.2% probability that $\bar{x}$ will be within ±1.87 of μ

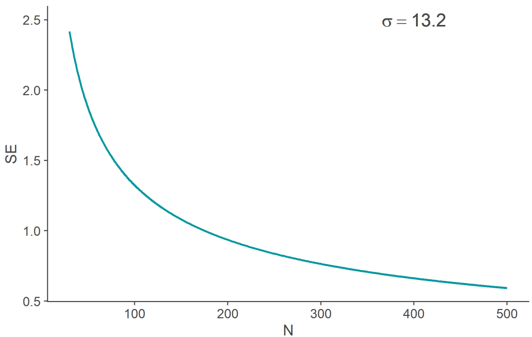
```
mean(samp)
```

```
## [1] 98.22903
```

Standard error

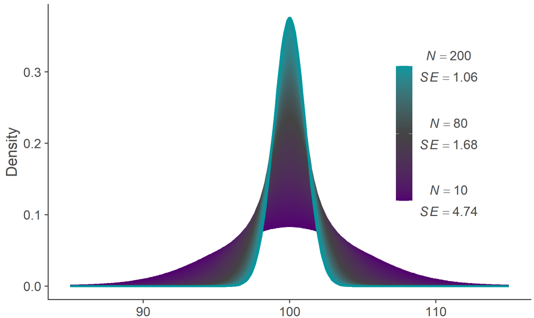
- *SE* is calculated using *N*: there's a relationship between the two

Plot | R code



Standard error

- That is why *larger samples are more reliable*!



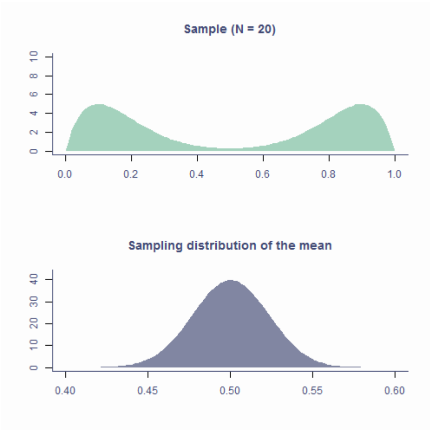
Standard error

- Allows us to gauge the *resampling accuracy* of parameter estimate (*e.g.*, $\hat{\mu}$) in sample
- The smaller the *SE*, the more confident we can be that the parameter estimate ($\hat{\mu}$) in our sample is close to those in other samples of the same size
- We don't particularly care about our specific sample: we care about the population!

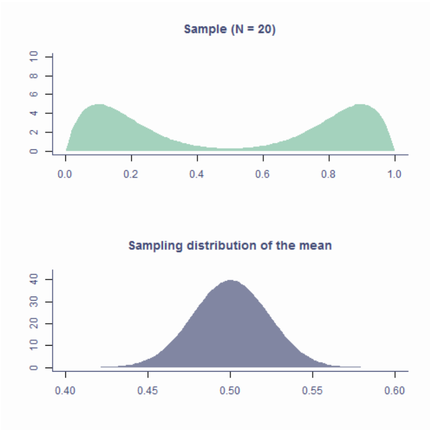
The Central Limit Theorem

- Sampling distribution of the mean is *approximately normal*
- True no matter the shape of the population distribution!
- This is the Central Limit Theorem
    - *"Central"* as in *"really important"* because, well, it is!

CLT in action



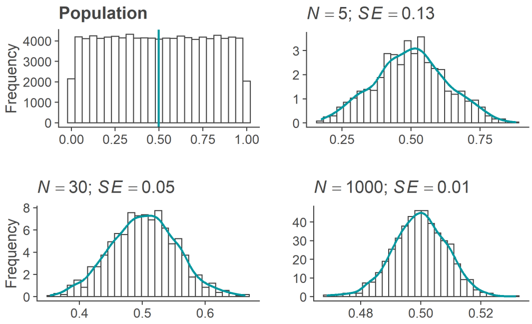
CLT in action



Approximately normal

- As N gets larger, the sampling distribution of $\bar{x}$ tends towards a normal distribution with **mean =** μ and $SD = \frac{\sigma}{\sqrt{N}}$

Plot | R code



Take-home message

- *Distribution* is the number of observations per each value of a variable
- There are many mathematically well-described distributions
  - Normal (Gaussian) distribution is one of them
- Each has a formula allowing the calculation of the probability of drawing an arbitrary range of values

Take-home message

- Normal distribution is
  - **continuous**
  - **unimodal**
  - **symmetrical**
  - **bell-shaped**
  - *it's the right proportions that make a distribution normal!*
- In a normal distribution it is true that
  - **~68.2%** of the data is within **±1** ***SD*** from the mean
  - **~95.4%** of the data is within **±2** ***SD*** from the mean
  - **~99.7%** of the data is within **±3** ***SD*** from the mean
- Every known distribution has its own *critical values*

Take-home message

- Statistics of random samples differ from parameters of a population
- As *N* gets bigger, sample statistics approaches population parameters
- Distribution of sample parameters is the **sampling distribution**
- **Standard error** of a parameter estimate is the *SD* of its sampling distribution
    - Provides *margin of error* for estimated parameter
    - The larger the sample, the less the estimate varies from sample to sample

Take-home message

- **Central Limit Theorem**
  - Really important!
  - Sampling distribution of the mean tends to normal even if population distribution is not normal
- Understanding distributions, sampling distributions, standard errors, and CLT it *most of what you need* to understand all the stats techniques we will cover