STAT115: Introduction to Biostatistics

University of Otago
Ōtākou Whakaihu Waka

Lecture 10: Introduction to Statistical Modelling

Outline

- Populations and parameters
- Samples and statistics
- Estimation of parameters
- Introduce the normal distribution

Big picture

- We might be interested in the cholesterol levels of male heart attack patients
  - e.g. what is the mean cholesterol level for such patients?
- How could we find the mean cholesterol level for such patients?

Big picture

- We might be interested in the cholesterol levels of male heart attack patients
  - e.g. what is the mean cholesterol level for such patients?
- How could we find the mean cholesterol level for such patients?
- Problem: question refers to a population – let's say all male NZ heart attack patients since 1980
  - We cannot answer it unless we measure every individual in the population
  - Typically impossible (or at least impractical)
- Formulate a statistical model
  - Use a sample to tell us about the population

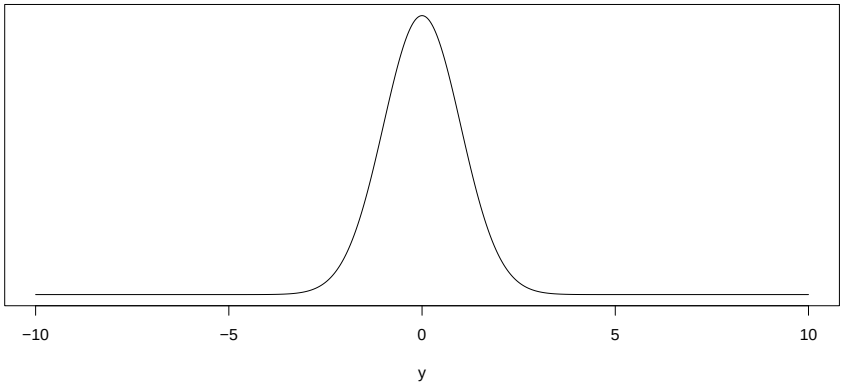
Statistical model

- Idea: assume the population values follow some distribution
  - The distribution tell us how the values vary in the population
- The distribution has unknown parameters
  - Parameter: any quantity that describes a population
- It is the parameter(s) that are of interest
  - Tell us about the population
- Abstract concepts
  - Introduce an example to make the idea more concrete
- We have seen probability distributions in simple 'generic' cases
  - Introduce specific case: normal distribution

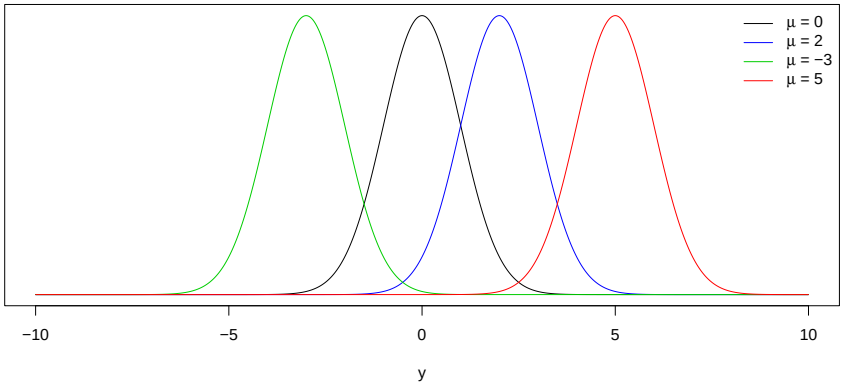
Normal model

- Assume that the data are normally distributed
  - We might also say we are using a normal model
- The normal distribution is sometimes called:
  - Bell-shaped curve
  - Gaussian model
- Described by two parameters:
  - Mean μ (Greek letter mu)
  - Standard deviation σ (Greek letter sigma)
    - Often refer to the variance σ^2 instead of the standard deviation
- We will spend some time familiarizing ourselves with the normal distribution

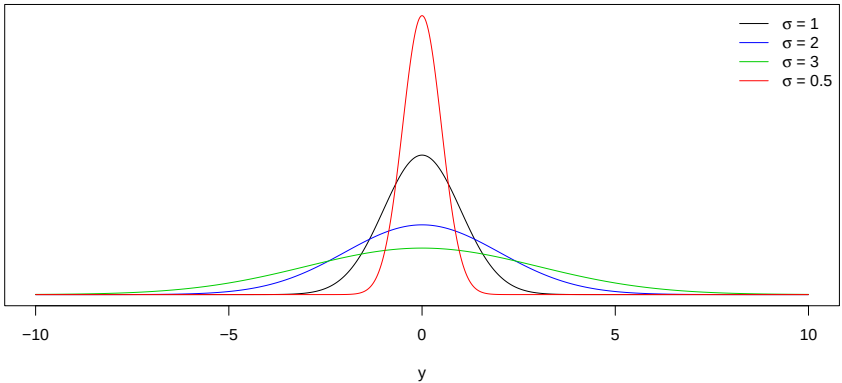
Probability density function (pdf) of normal distribution: $\mu = 0$, $\sigma = 1$



Pdf of normal distribution: different μ



Pdf of normal distribution: different σ



Model for total cholesterol in heart attack patients

- We assume that total cholesterol (in mmoles per litre) follow a normal distribution
  - This is an assumption about the population of all male heart attack patients
  - Parameters μ and σ are unknown
    - μ: mean cholesterol at a population level
    - σ: standard deviation of cholesterol at a population level
- Typically use Greek letters for parameters
  - Here we are using μ and σ

Populations and samples

- Big idea: use a sample (and statistics) to estimate parameters
  - The estimate is an educated guess at the parameter value
- We have total cholesterol measurements (in mmole per litre) from 32 male heart attack patients in Auckland (cf. lecture 2)
- How could we use this sample to estimate μ?

Populations and samples

- Big idea: use a sample (and statistics) to estimate parameters
  - The estimate is an educated guess at the parameter value
- We have total cholesterol measurements (in mmole per litre) from 32 male heart attack patients in Auckland (cf. lecture 2)
- How could we use this sample to estimate μ?
- The sample mean $\bar{y}$ could be used to estimate the population mean μ
  - The sample mean $\bar{y}$ is an example of a statistic
  - Statistic: any quantity computed from values in a sample

That's easy ... are we done?

- Our example: finding a 'suitable' statistic is straightforward
  - Not always the case
- Let's imagine a more complex study:
  - Do cholesterol levels change with age? Is there an association between cholesterol and heart volume?
    - What statistic(s) should we use for that?
- Later in semester we will think more about general strategies for finding suitable statistics (estimators)

Model fitting

- For our model, we have
  - $\hat{\mu} = \bar{y}$
  - $\hat{\sigma} = s$
  - The population std deviation (σ) is being estimated by the sample std deviation (s)
- We have used the hat symbol $\hat{}$ to represent that we are estimating a parameter
  - $\hat{\mu}$ is said "mu-hat"
  - $\hat{\mu} = \bar{y}$: the parameter μ is being estimated by $\bar{y}$ (a statistic)

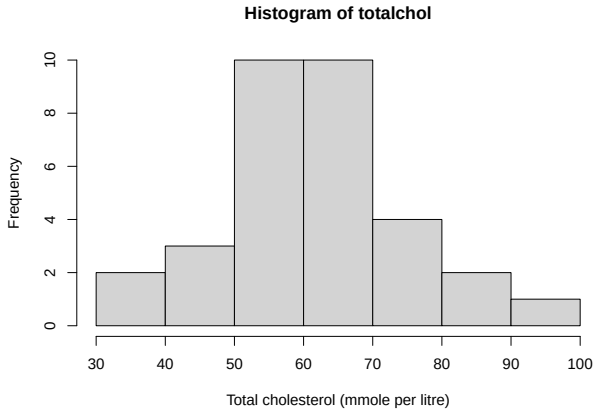
Fitted model

- Look at the fitted model (graphically)
  - (Normal) model at the estimated parameter values
- Compare the fitted model to the data
  - Load the Auckland heart attack patient data into R (for the next few slides)
  - Apply R coding ideas we learned about earlier

```
nzheart = read.csv('../data/nzheart.csv')
totalchol = nzheart$Chol
hist(totalchol,xlab="Total cholesterol (mmole per litre)")
```

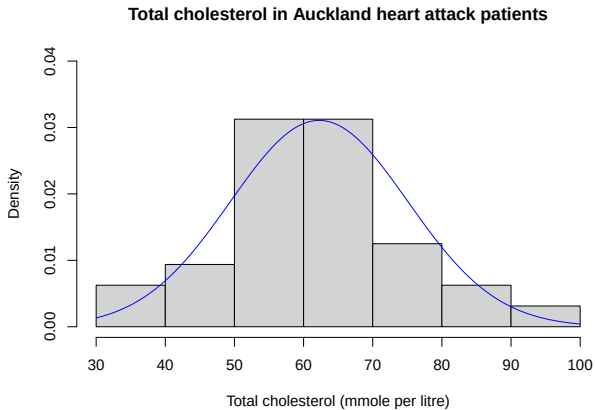
Fitted model

Basic plot



Fitted model

Embellished plot



https://mathstatfiles.otago.ac.nz/STAT115/covidplot.r

Statistical models

- Common mistakes:
  - Believing that μ is 62 mmole per litre (the sample mean)
    - This is $\bar{y}$, the estimate of μ
    - It is impossible to know the value of μ
  - Believing that the model is absolutely true
    - Hope the model chosen is a reasonable approximation to reality
    - We should check this
- Checking the model fit
  - Looking to see if the model and the data are 'out of sync'
    - Plot: normal appears to describe the data reasonably well
  - Think more about model fit later in course (regression)

Statistical models

Normal and non-normal models

- In this example we have used a normal model for the data
- Seems reasonable enough for cholesterol measurements
  - Continuous values (in principle)
  - Distribution appears reasonably close to symmetric
- Not all data looks like this!
  - Different types of data (yes/no, count, categories, time, space, ...)
  - Different characteristics (e.g. income)
  - Different complexity
- Options...
  - Transform the data so it is more easily modelled
  - Use a probability distribution with different characteristics ...

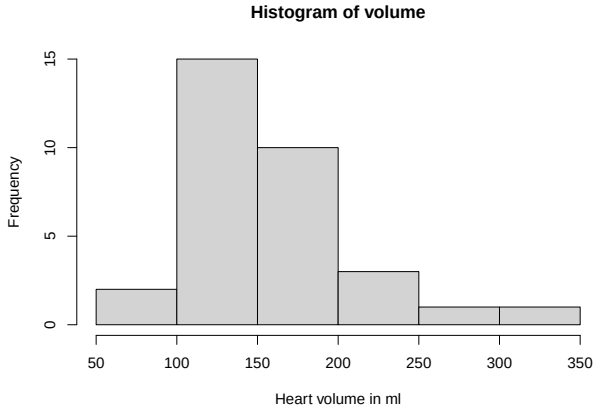
Transformations

- Let's look at heart volume from the Auckland heart attack patient data
- The data are continuous, so a normal distribution is not out of the question
- But does a normal distribution model seem appropriate?

```
nzheart = read.csv('../data/nzheart.csv')
volume = nzheart$Vol
hist(volume,xlab="Heart volume in ml")
```

Transformations

Histogram of heart volumes (raw scale)



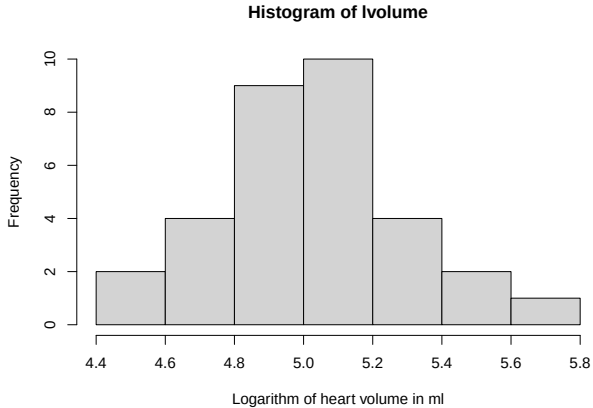
Log-transformations

- Physical data (particularly when non-negative) often have a skewed (non-symmetric) distribution
- Taking the logarithm can sometimes make the data look more normally distributed
- I.e. work with variable $\log(y)$ rather than y.
  - Here and throughout course, $\log$ is assumed to be natural (base e) logarithm
  - Sometimes indicated by $\ln$ on calculators
  - R uses natural logarithm by default

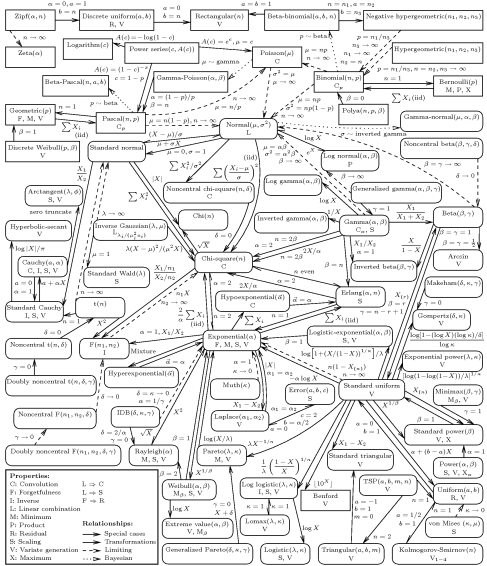
```
lvolume = log(volume)
hist(lvolume,xlab="Logarithm of heart volume in ml")
```

Transformations

Histogram of logairthms of heart volumes



Other Distributions for Modelling



Looking forward I

- We will be working with a normal model for a few weeks
  - Look more at the normal distribution
    - Use it to describe (and model) data
    - Want to understand it as much as possible
- Explore a strategy for estimating parameters
  - Barely scratch the surface
  - Cover in more depth in higher level courses STAT 270, 370, 371
- Explore 'extensions' to normal model (e.g. regression)
- Explore models for other types of data: yes/no data

Looking forward II

- What does our estimate tell us about the parameter?
  - We have an estimate of μ from the sample of size 68
  - How 'close' to the true value of μ is it likely to be?
- Is the estimate likely to be better / worse if it were from:
  - A sample of size 6?
  - A sample of size 600?
- Explore how to determine how precise/uncertain the estimate is
- Also important to consider how the data were collected
  - e.g. does hospital in question get 'typical' heart attack patients?
  - We'll come back to this later in the semester
- Use the model for prediction (regression)

Summary

- Introduction to statistical modeling
  - Fit a normal model
  - Estimated the parameter μ with the statistic $\bar{y}$
- Next: get a better understanding of the normal distribution