

Introduction (Module 1)



Statistics (MAST20005) &
Elements of Statistics
(MAST90058)

School of Mathematics and Statistics
University of Melbourne

Semester 2, 2020

Aims of this module

- Brief information about this subject
- Brief revision of some prerequisite knowledge (probability)
- Introduce some basic elements of statistics, data analysis and visualisation

Outline

Subject information

Review of probability

Descriptive statistics

Basic data visualisations

What is statistics?

Let's see some examples...



[Bureau Home](#) > [Australia](#) > [Victoria](#) > [Forecasts](#) > Melbourne Forecast

Melbourne Forecast



[View the current warnings for Victoria](#)

Forecast issued at 5:05 am EST on Tuesday 17 July 2018.

Forecast for the rest of Tuesday



Max **16**

Showers developing. Windy.

interval
estimate

Possible rainfall: **1 to 3 mm**

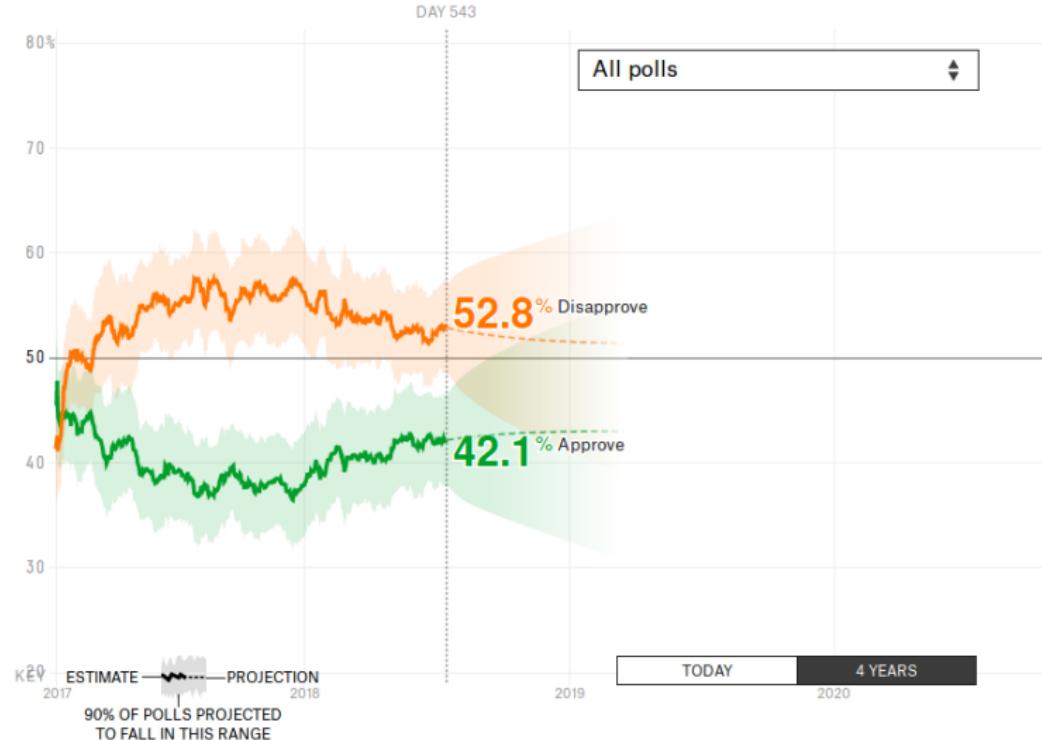
Chance of any rain: **80%**

Melbourne area

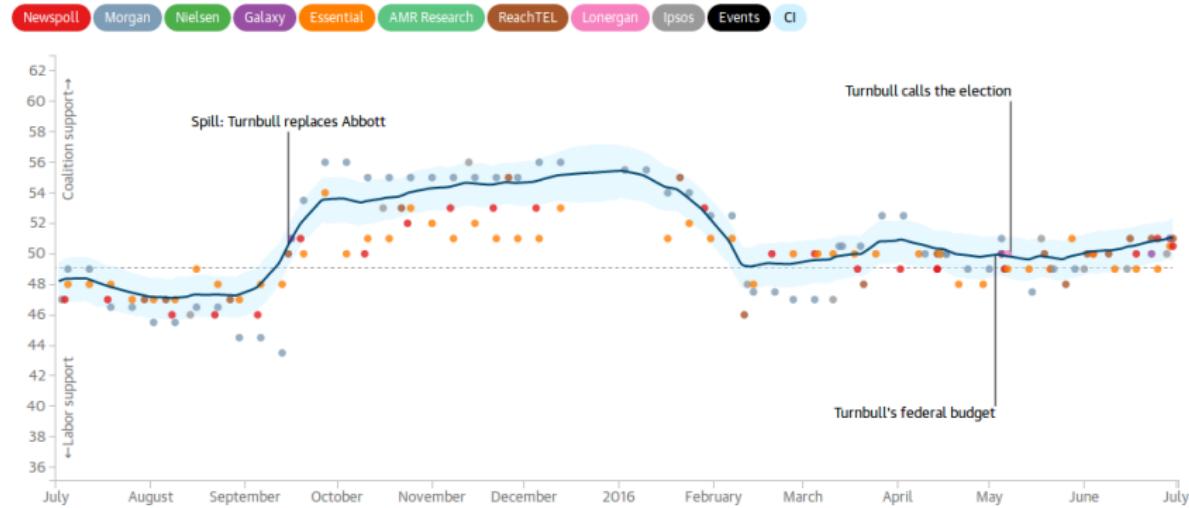
Sunny morning. High (80%) chance of showers during this afternoon and evening and the chance of hail and thunder. Winds northerly 25 to 40 km/h increasing to 40 to 55 km/h this morning then turning west to northwesterly 30 to 45 km/h in the late afternoon.

How unpopular is Donald Trump?

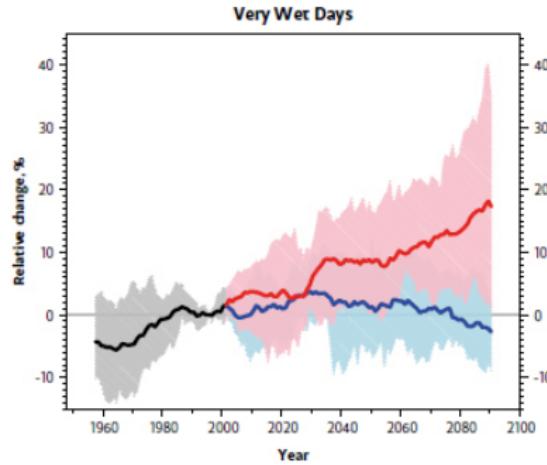
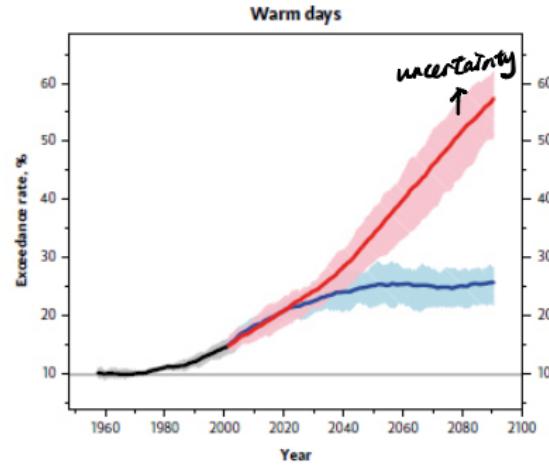
An updating calculation of the president's approval rating, accounting for each poll's quality, recency, sample size and partisan lean. [How this works »](#)



Poll of polls: who will win the Australian election?

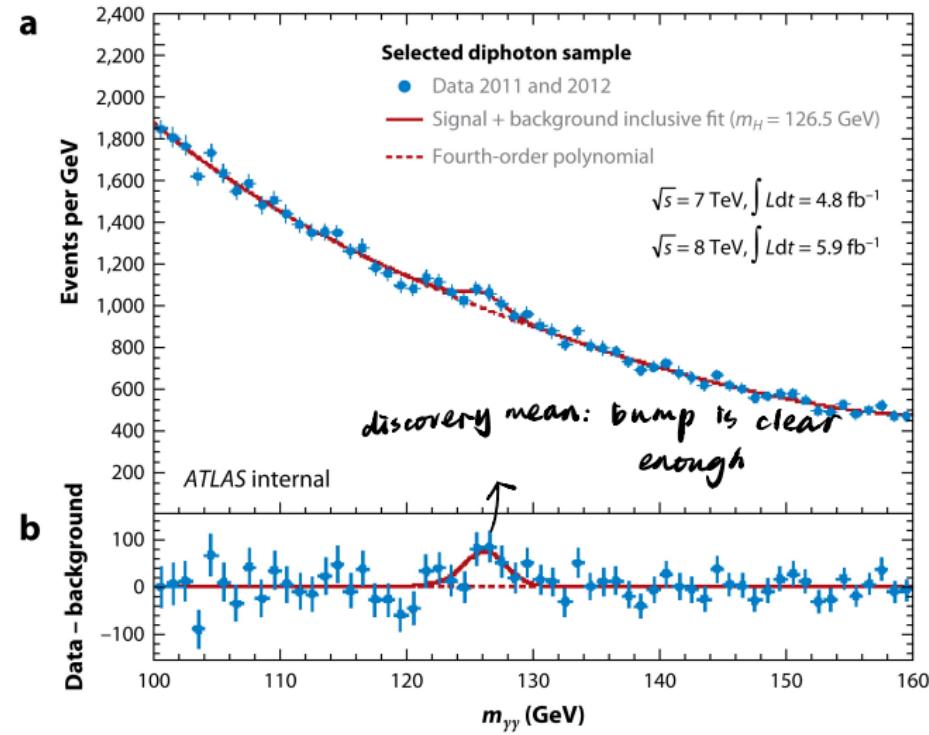


Climate change modelling



* assess the uncertainty

Discovery of the Higgs Boson (the ‘God Particle’)

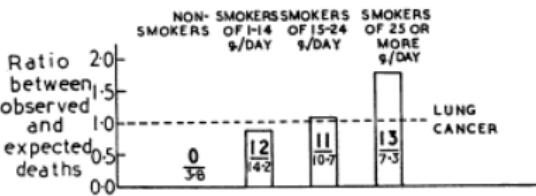


Smoking leads to lung cancer

TABLE IV.—*Standardized Death Rate Per Annum Per 1,000 Men Aged 35 Years and Above in Relation to the Most Recent Amount of Tobacco Smoked*

Cause of Death	No. of Deaths Recorded	Death Rates of Non-smokers	Death Rates of Men Smoking a Daily Average of			Death Rate of All Men
			1 g.-	15 g.-	25 g.+	
Lung cancer .. .	36*	0.00	0.48	0.67	1.14	0.66
Other cancers .. .	92	2.32	1.41	1.50	1.91	1.65
Respiratory disease (other than cancer)	54	0.86	0.88	1.01	0.77	0.94
Coronary thrombosis ..	235	3.89	3.91	4.71	5.15	4.27
Other cardiovascular diseases	126*	2.23	2.07	1.58	2.78	2.14
Other diseases .. .	247	4.27	4.67	3.91	4.52	4.36
All causes .. .	789	13.61	13.42	13.38	16.30	14.00

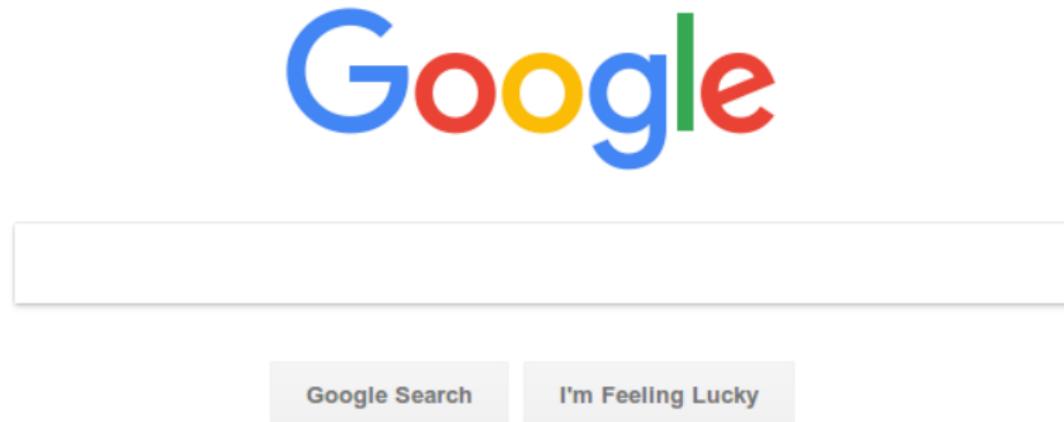
* 1 case in which lung cancer was recorded as a contributory but not a direct cause of death has been entered in both groups.



Richard Doll & A. Bradford Hill

"The Mortality of Doctors in Relation to Their Smoking Habits"
British Medical Journal (1954)

A/B testing for websites



41 shades of blue?!

Example: Zillow Real Estate Data

The screenshot shows the Zillow website interface. At the top, there are navigation links for Buy, Rent, Sell, Mortgages, Agent Finder, and More. The search bar indicates "San Francisco Bay Area". Below the search bar, there are filters for Listing Type (any), Any Price, 0+ Beds, Home Type, and More. There are also buttons for Save Search and Saved Homes (0).

A message at the top left says "Only showing 500 homes. Zoom in, or use filters to narrow your search." To the right, it displays "Bay Area CA Real Estate" with "9,876 homes for sale. 142 unmapped." Below this, there are tabs for Homes for You, Newest, Cheapest, and More.

The main area features a map of the San Francisco Bay Area with numerous red dots representing property locations. On the right side, there is a grid of six property listings:

- House for Sale: \$490,000**
2558 Craig Ct, Castro Valley, CA
Keller Williams Realty
- Condo for Sale: \$595,000**
77 Carl St, San Francisco, CA
- House for Sale: \$29,500,000**
2820 Scott St, San Francisco, CA
Golden Gate Sotheby's International Realty
- Condo for Sale: \$287,274**
1160 Mission St, UNIT 709, San Francisco, CA
Searcina Real Estate
- Open Sat. 1-4pm**
\$178,550 (New!)
- Interior view**

THE INDEPENDENT



(Ireland, €1) 70p
Thursday 7 June 2007
www.independent.co.uk
• 18MBPS-440

Bipolar disorder
Also known as manic depression, it affects 100 million people around the world

Hypertension
High blood pressure affects 16 million people in Britain. Can lead to stroke, heart disease and kidney failure

Type 1 diabetes
Diabetic condition in which sufferers have to inject insulin. Affects 350,000 people in UK

Type 2 diabetes
Almost 2 million Britons are affected by this late-onset disease, which is linked with the growing obesity epidemic

Coronary heart disease
The most frequent cause of death in Britain, with 100,000 victims every year. By 2020, it will be the biggest killer in the world

Rheumatoid arthritis
Nearly 400,000 people in Britain are afflicted with this auto-immune disease of the joints

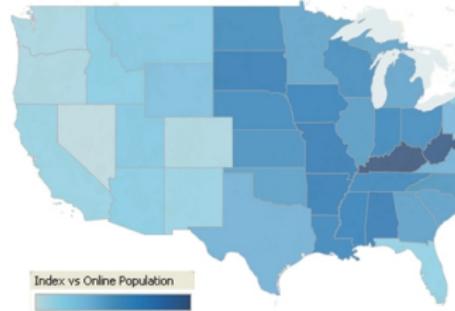
Crohn's disease
Up to 60,000 people are affected by this debilitating bowel condition which can cause distress and pain for a lifetime

THE GENETIC REVOLUTION

DISCOVERY OF GENES RESPONSIBLE FOR SEVEN OF THE MOST COMMON ILLNESSES OFFERS HOPE TO MILLIONS OF SUFFERERS

FULL STORY, PAGE 2

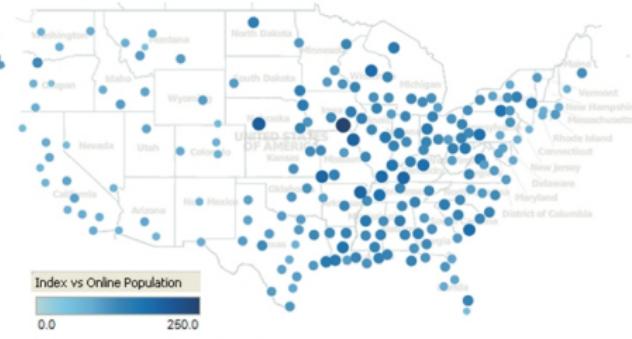
Web analytics



2011 Black Friday Shoppers by State

Week ending November 26, 2011, compared with the Online Population

State	Visits Share	Visits Share	Index
	Black Friday 2011 Shoppers	Online Population	
1 California	10.05%	12.45%	81
2 Texas	7.93%	7.72%	103
3 New York	5.23%	6.08%	86
4 Illinois	4.88%	4.34%	112
5 Florida	4.88%	5.92%	82
6 Ohio	4.49%	3.66%	123
7 Pennsylvania	4.03%	4.01%	100
8 Georgia	3.57%	3.23%	111
9 Michigan	3.51%	3.23%	109
10 North Carolina	3.35%	2.92%	115

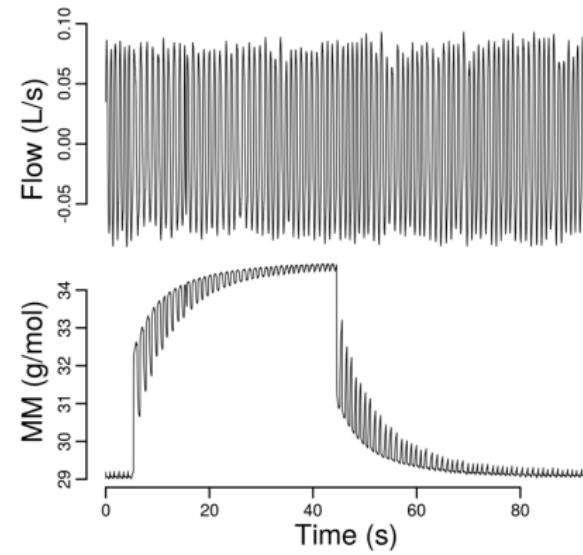


2011 Black Friday Shoppers by DMA

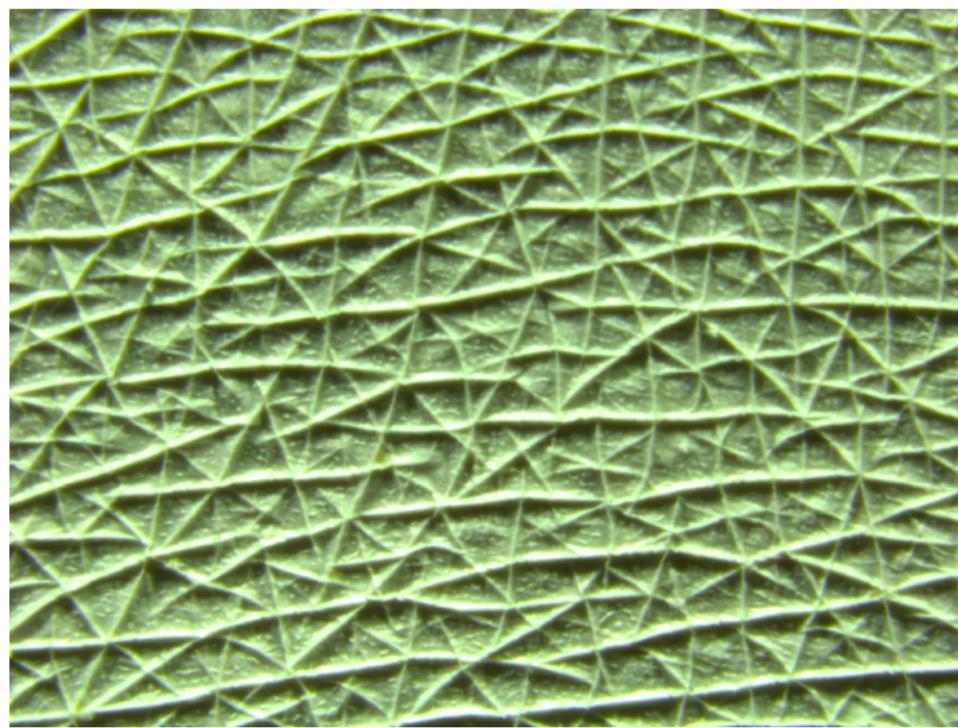
Week ending November 26, 2011, compared with the Online Population

DMA	Visits Share	Visits Share	Index
	Black Friday 2011 Shoppers	Online Population	
1 New York, NY	4.44%	6.51%	68
2 Los Angeles, CA	4.01%	4.91%	82
3 Chicago, IL	3.01%	3.04%	99
4 Dallas-Ft. Worth, TX	2.29%	2.21%	104
5 Atlanta, GA	2.12%	2.07%	102
6 Philadelphia, PA	1.98%	2.58%	77
7 Houston, TX	1.90%	1.84%	103
8 Boston et al, MA-NH	1.74%	2.10%	83
9 Washington et al, DC-MD	1.72%	2.04%	84
10 Minneapolis-St. Paul, MN	1.61%	1.51%	107

Lung testing in infants



Skin texture image analysis





1 February 2014 Last updated at 08:47



A statistically modelled wedding

By Ruth Alexander
BBC News



TEDDY TAN

It's the classic dilemma for anyone planning a wedding. The list of friends and family you'd like to invite is seemingly endless. Your budget is not, and neither is the venue's floor space. Could one couple have found a solution via statistical modelling?

Damjan Vukcevic and Joan Ko, planning their wedding in Melbourne, Australia, were struggling to draw up an invitation list of family and friends in places as far-flung as Serbia, Taiwan, the UK and the US.

In today's
Magazine

The world's first
church-mosque-
synagogue?
[The face of British](#)

Goals of statistics

- Answer questions using **data**
- Evaluate **evidence**
- Optimise **study design**
- Make **decisions**

And, importantly:

- Clarify **assumptions**
- Quantify **uncertainty**

Why study statistics?

“The best thing about being a statistician is that you get to play in everyone’s backyard.”

—John W. Tukey (1915–2000)

“I keep saying the sexy job in the next ten years will be statisticians... The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it’s going to be a hugely important skill in the next decades...”

—Hal Varian, Google’s Chief Economist, Jan 2009

The best job

U.S. News [Best Business Jobs](#) in 2020:

1. Statistician
2. Medical and Health Services Manager
3. Mathematician

CareerCast (recruitment website) [Best Jobs of 2019](#):

1. Data Scientist
2. Statistician
3. University Professor

Subject overview

Statistics (MAST20005), Elements of Statistics (MAST90058)

These subjects introduce the basic elements of **statistical modelling**, **statistical computation** and **data analysis**. They demonstrate that many commonly used statistical procedures arise as applications of a common theory. They are an entry point to further study of both **mathematical** and **applied** statistics, as well as broader data science.

Students will develop the ability to **fit statistical models to data**, **estimate parameters of interest** and **test hypotheses**. Both classical and Bayesian approaches will be covered. The importance of the underlying **mathematical theory of statistics** and the use of **modern statistical software** will be emphasised.

Joint teaching

MAST20005 and MAST90058 share the same lectures but have separate tutorials and lab classes. The teaching and assessment material for both subjects will overlap significantly.

Subject website (LMS)

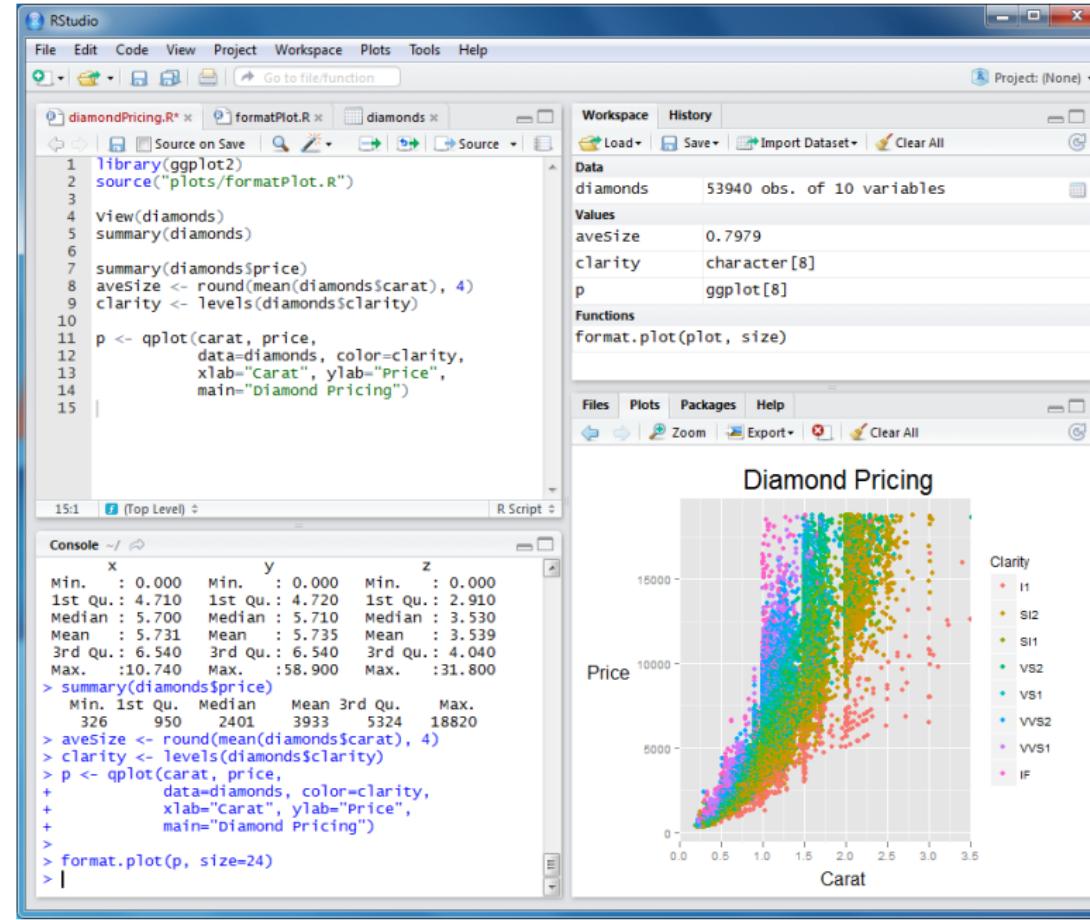
- Full information is on the subject website, available through the Learning Management System (LMS).
- Only a brief overview is covered in these notes. Please read all of the info on the LMS as well.
- New material (e.g. problem sets, assignments, solutions) and announcements will appear regularly on the LMS.

Subject structure

- **Lectures:** Three 1-hour lectures per week. Lecture notes/slides will appear on the LMS.
- **Tutorials:** One 1-hour tutorial per week (starting in week 2). Tutorial problems and solutions will appear on the LMS.
- **Computer lab classes:** One 1-hour lab per week (starting in week 2), immediately following the tutorial. Lab notes, exercises and solutions will appear on the LMS.

Computing

- This subject introduces basic statistical computing and programming skills.
- We make extensive use of the [R statistical software environment](#).
- Knowledge of R will be **essential** for some of the tutorial problems, assignment questions and will also be examined.
- We will use the [RStudio](#) program as a convenient interface with R.



Textbook

R. Hogg, E. Tanis, and D. Zimmerman. *Probability and Statistical Inference*. 9th Edition, Pearson, 2015.

- This subject is based on Chapters 6–9.
- Some of the teaching material is taken from the textbook.
- This textbook is being **phased out** for this subject.
- There are **important differences** between the subject content and the textbook. We will point many of these out, but please ask if unsure.

Assessment

- 3 assignments (20%)
 1. Hand out at the start of week 4, due at the end of week 5
 2. Hand out at the start of week 7, due at the end of week 8
 3. Hand out at the start of week 10, due at the end of week 11
- 45-minute computer lab test held in week 12 (10%)
- 3-hour written examination in the examination period (70%)

Plagiarism declaration

- Everyone must complete the **Plagiarism Declaration Form**
- Do this on the LMS
- Do this ASAP!

Staff contacts

Subject coordinator / Lecturer (stream 2)

Dr Tingjin Chu <tingjin.chu@unimelb.edu.au>

Lecturer (stream 1)

Dr Damjan Vukcevic <damjan.vukcevic@unimelb.edu.au>

Tutorial coordinator

Ms Martina Hoffmann <martina.hoffmann@unimelb.edu.au>

See the LMS for details of consultation hours

Online discussion forum (Piazza)

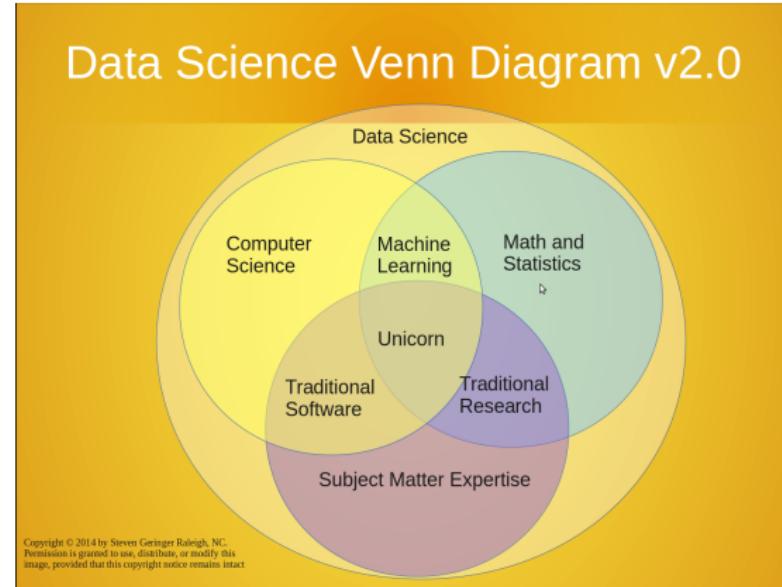
- Access via the LMS
- Post any general questions on the forum
- Do **not** send them by email to staff
- You can answer each others' questions
- Staff will also help to answer questions

Student representatives

Student representatives assist the teaching staff to ensure good communication and feedback from students.

See the LMS to find the contact details of your representatives.

What is Data Science?



Data science is a 'team sport'

How to succeed in statistics / data science?

- Get experience with **real data**
- Develop your computational skills, **learn R**
- Understand the **mathematical theory**
- Collaborate with others, **use Piazza**

This subject is challenging

- It is mathematical
 - Manipulating equations
 - Calculus
 - Probability
 - Proofs
- But the ‘real’ world also matters
 - Context can ‘trump’ mathematics
 - More than one correct answer
 - Often uncertain about the answer

Diversity

In 2017: **341 students**

- 60% Bachelor of Commerce
 - 24% Bachelor of Science
 - 6% Master of Science (Bioinformatics)
 - 10% 8 other degrees/categories
-

What are your strengths and weaknesses?

Get extra help

- Your classmates
- Piazza
- Textbooks
- [Consultation hours](#)
- [Oasis](#)

Homework

1. Complete plagiarism declaration on the [LMS](#)
2. Log in to Piazza
3. Install [RStudio](#) on your computer
4. Start reading lab notes for week 2 (long!)

Tips

The best way to learn statistics is by **solving problems** and '**getting your hands dirty**' with data.

We encourage you to attend all lectures, tutorial and computer labs to get as much practice and feedback as possible.

Good luck!

Outline

Subject information

Review of probability

Descriptive statistics

Basic data visualisations

Why probability?

- It forms the mathematical foundation for statistical models and procedures
- Let's review what we know already...

Random variables (notation)

- Random variables (rvs) are denoted by uppercase letters: X, Y, Z , etc.
- Outcomes, or realisations, of random variables are denoted by corresponding lowercase letters: x, y, z , etc.

Distribution functions

- The cumulative distribution function (cdf) of X is

$$F(x) = \Pr(X \leq x), \quad -\infty < x < \infty$$

- If X is a continuous rv then it has a probability density function (pdf), $f(x)$, that satisfies

$$f(x) = F'(x) = \frac{d}{dx} F(x)$$

$$F(x) = \int_{-\infty}^x f(t) dt$$

- If X is a discrete rv then it has a probability mass function (pmf),

$$p(x) = \Pr(X = x), \quad x \in \Omega$$

where Ω is a discrete set, e.g. $\Omega = \{1, 2, \dots\}$.

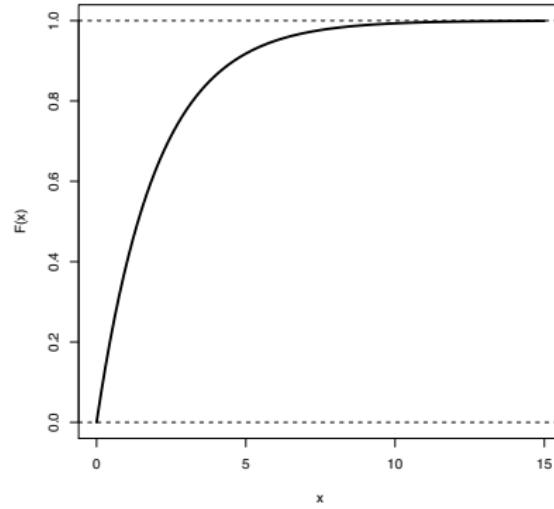
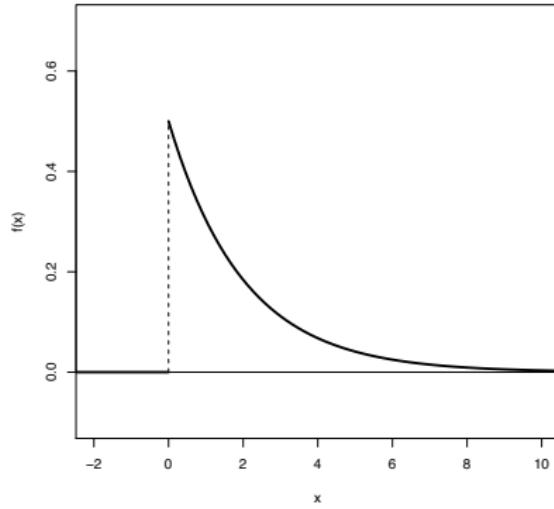
- $\Pr(X > x) = 1 - F(x)$ is called a tail probability of X
- $F(x)$ increases to 1 as $x \rightarrow \infty$ and decreases to 0 as $x \rightarrow -\infty$
- If the rv has a certain distribution with pdf f (or pmf p), we write

$$\underbrace{X \sim f}_{\text{(or)}} \quad \underbrace{X \sim p}_{\text{(or)}}$$

Example: Unemployment duration

A large group of individuals have recently lost their jobs. Let X denote the length of time (in months) that any particular individual will stay unemployed. It was found that this was well-described by the following pdf:

$$f(x) = \begin{cases} \frac{1}{2}e^{-x/2}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$



Clearly, $f(x) \geq 0$ for any x and the total area under the pdf is:

$$\Pr(-\infty < X < \infty) = \int_0^{\infty} \frac{1}{2}e^{-x/2} dx = \frac{1}{2} \left[-2e^{-x/2} \right]_0^{\infty} = 1.$$

The probability that a person in the population finds a new job within 3 months is:

$$\Pr(0 \leq X \leq 3) = \int_0^3 \frac{1}{2}e^{-x/2} dx = \frac{1}{2} \left[-2e^{-x/2} \right]_0^3 = 0.7769.$$

Example: Received calls

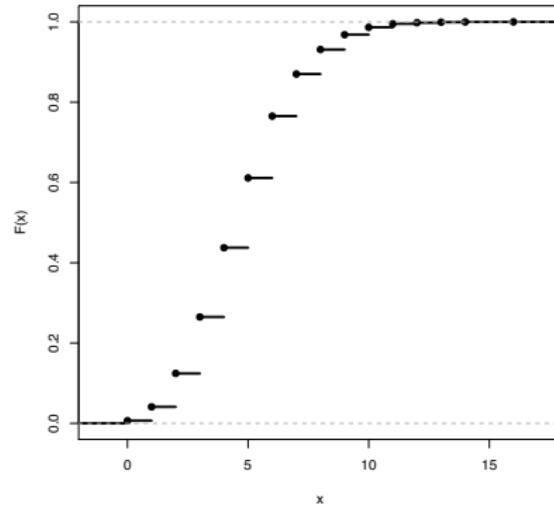
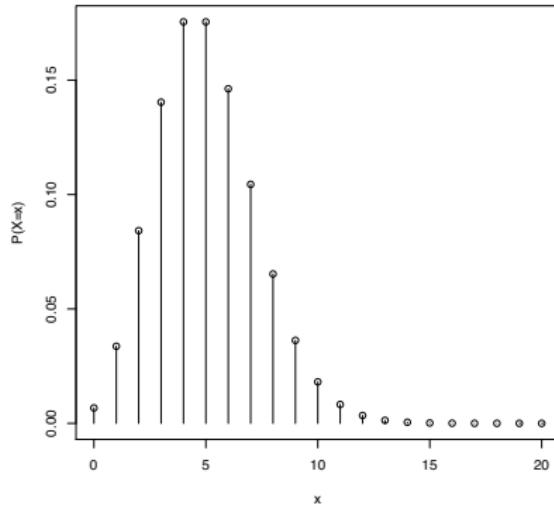
The number of calls received by an office in a given day, X , is well-represented by a pmf with the following expression:

$$p(x) = \frac{e^{-5} 5^x}{x!}, \quad x \in \{0, 1, 2, \dots\},$$

where $x! = 1 \cdot 2 \cdots (x - 1) \cdot x$ and $0! = 1$. For example,

$$\Pr(X = 1) = e^{-5} 5 = 0.03368$$

$$\Pr(X = 3) = \frac{e^{-5} 5^3}{3 \cdot 2 \cdot 1} = 0.1403$$



To show that $p(x)$ is a pmf we need to show

$$\sum_{x=0}^{\infty} p(x) = p(0) + p(1) + p(2) + \dots = 1.$$

Since the Taylor series expansion of e^z is $\sum_{i=0}^{\infty} z^i/i!$, we can write

$$\sum_{i=0}^{\infty} \frac{e^{-5} 5^x}{x!} = e^{-5} \sum_{i=0}^{\infty} \frac{5^x}{x!} = e^5 e^{-5} = 1.$$

Moments and variance

- The **expected value** (or the **first moment**) of a rv is denoted by $\mathbb{E}(X)$ and

$$\mathbb{E}(X) = \sum_{x=-\infty}^{\infty} x p(x) \quad (\text{discrete rv})$$

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx \quad (\text{continuous rv})$$

- Higher moments**, $\mu_k = \mathbb{E}(X^k)$, for $k \geq 1$, can be obtained by

$$\mathbb{E}(X^k) = \sum_{x=-\infty}^{\infty} x^k p(x) \quad (\text{discrete rv})$$

$$\mathbb{E}(X^k) = \int_{-\infty}^{\infty} x^k f(x) dx \quad (\text{continuous rv})$$

- More generally for a function $g(x)$ we can compute

$$\mathbb{E}(g(X)) = \sum_{x=-\infty}^{\infty} g(x)p(x) \quad (\text{discrete rv})$$

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x) dx \quad (\text{continuous rv})$$

Letting $g(x) = x^k$ gives the moments.

- The **variance** of X is defined by

$$\text{var}(X) = \mathbb{E} \left\{ (X - \mathbb{E}(X))^2 \right\}$$

and the **standard deviation** of X is $\text{sd}(X) = \sqrt{\text{var}(X)}$

- “Computational” formula: $\text{var}(X) = \mathbb{E}(X^2) - \{\mathbb{E}(X)^2\}$

Basic properties of expectation and variance

- For any rv X and constant c ,

$$\mathbb{E}(cX) = c\mathbb{E}(X), \quad \text{var}(cX) = c^2 \text{var}(X)$$

- For any two rvs X and Y ,

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

- For any two independent rvs X and Y ,

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$$

- More generally,

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2 \text{cov}(X, Y)$$

where $\text{cov}(X, Y)$ is the **covariance** between X and Y

Covariance

- Definition of covariance:

$$\text{cov}(X, Y) = \mathbb{E} \{(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\}$$

- Specifically, for the continuous case

$$\text{cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mathbb{E}(X))(y - \mathbb{E}(Y))f(x, y) dx dy$$

where $f(x, y)$ is the bivariate pdf for pair (X, Y) .

- “Computational” formula:

$$\begin{aligned}\text{cov}(X, Y) &= \mathbb{E}\{(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\} \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\end{aligned}$$

Correlation

- If $\text{cov}(X, Y) > 0$ then X and Y are **positively correlated**
- If $\text{cov}(X, Y) < 0$ then X and Y are **negatively correlated**
- If $\text{cov}(X, Y) = 0$ then X and Y are **uncorrelated**
- The **correlation** between X and Y is defined as:

$$\rho = \text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\text{sd}(X) \text{sd}(Y)}, \quad -1 \leq \rho \leq 1$$

- When $\rho = \pm 1$ then X and Y are perfectly correlated

Moment generating functions

- A **moment generating function (mgf)** of a rv X is

$$M_X(t) = \mathbb{E}(e^{tX}), \quad t \in (-\infty, \infty)$$

- It enables us to generate moments of X by differentiating at $t = 0$

$$M'_X(0) = \mathbb{E}(X)$$

$$M_X^{(k)}(0) = \mathbb{E}(X^k), \quad k \geq 1$$

- The mgf uniquely determines a distribution. Hence, knowing the mgf is the same as knowing the distribution.
- If X and Y are independent rvs,

$$M_{X+Y}(t) = \mathbb{E}\{e^{t(X+Y)}\} = \mathbb{E}\{e^{tX}\} \mathbb{E}\{e^{tY}\} = M_X(t)M_Y(t)$$

i.e. the mgf of the sum is the product of individual mgfs.

Bernoulli distribution

- X takes on the values 1 (success) or 0 (failure)
- $X \sim \text{Be}(p)$ with pmf

$$p(x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}$$

- Properties:

$$\mathbb{E}(X) = p$$

$$\text{var}(X) = p(1-p)$$

$$M_X(t) = pe^t + 1 - p$$

Binomial distribution

- $X \sim \text{Bi}(n, p)$ with pmf

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, 1, \dots, n\}$$

- Properties:

$$\mathbb{E}(X) = np$$

$$\text{var}(X) = np(1-p)$$

$$M_X(t) = (pe^t + 1 - p)^n$$

Poisson distribution

- $X \sim \text{Pn}(\lambda)$ with pmf

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x \in \{0, 1, \dots\}$$

- Properties:

$$\mathbb{E}(X) = \text{var}(X) = \lambda$$

$$M_X(t) = e^{\lambda(e^t - 1)}$$

- It arises as an approximation to $\text{Bi}(n, p)$. Letting $\lambda = np$ gives

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \approx e^{-\lambda} \frac{\lambda^x}{x!}$$

as $n \rightarrow \infty$ and $p \rightarrow 0$.

Uniform distribution

- $X \sim \text{Unif}(a, b)$ with pdf

$$f(x) = \frac{1}{b-a}, \quad x \in (a, b)$$

- Properties:

$$\mathbb{E}(X) = \frac{(a+b)}{2}$$

$$\text{var}(X) = \frac{(b-a)^2}{12}$$

$$M_X(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}$$

- If $b = 1$ and $a = 0$, this is known as the uniform distribution over the unit interval.

Exponential distribution

- $X \sim \text{Exp}(\lambda)$ with pdf

$$f(x) = \lambda e^{-\lambda x}, \quad x \in [0, \infty)$$

- It approximates “time until first success” for independent $\text{Be}(p)$ trials every Δt units of time with $p = \lambda \Delta t$ and $\Delta t \rightarrow 0$
- Properties:

$$\mathbb{E}(X) = 1/\lambda$$

$$\text{var}(X) = 1/\lambda^2$$

$$M_X(t) = \frac{\lambda}{\lambda - t}$$

- It is famous for being the only continuous distribution with the **memoryless property**:

$$\Pr(X > y + x \mid X > y) = \Pr(X > x), \quad x \geq 0, \quad y \geq 0.$$

Normal distribution

- $X \sim N(\mu, \sigma^2)$ with pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in (-\infty, \infty), \quad \mu \in (-\infty, \infty), \quad \sigma > 0$$

- It is important in applications because of the Central Limit Theorem (CLT)

- Properties:

$$\mathbb{E}(X) = \mu$$

$$\text{var}(X) = \sigma^2$$

$$M_X(t) = e^{t\mu + t^2\sigma^2/2}$$

- When $\mu = 0$ and $\sigma = 1$ we have the **standard normal distribution**.
- If $X \sim N(\mu, \sigma^2)$,

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$



1 Quantiles

Let X be a continuous rv. The p th quantile of its distribution is a number π_p such that $p = \Pr(X \leq \pi_p) = F(\pi_p)$.

In other words, the area under $f(x)$ to the left of π_p is p :

$$p = \int_{-\infty}^{\pi_p} f(x) dx = F(\pi_p)$$

- π_p is also called the $(100p)$ th percentile
- The 50th percentile (0.5 quantile) is the median, denoted by $m = \pi_{0.5}$
- The 25th and 75th percentiles are the first and third quartiles, denoted by $q_1 = \pi_{0.25}$ and $q_3 = \pi_{0.75}$

Example: Weibull distribution

The time X until failure of a certain product has the pdf

$$f(x) = \frac{3x^2}{4} e^{-(x/4)^3}, \quad x \in (0, \infty).$$

The cdf is

$$\text{30\% } F(x) = 1 - e^{-(x/4)^3}, \quad x \in (0, \infty)$$

Then $\pi_{0.3}$ satisfies $0.3 = F(\pi_{0.3})$. Therefore,

$$\begin{aligned} 1 - e^{-(\pi_{0.3}/4)^3} &= 0.3 \\ \Rightarrow \ln(0.7) &= -(\pi_{0.3}/4)^3 \\ \Rightarrow \pi_{0.3} &= -4(\ln 0.7)^{1/3} = 2.84. \end{aligned}$$

Law of Large Numbers (LLN)

Consider a collection X_1, \dots, X_n of independent and identically distributed (**iid**) random variables with $\mathbb{E}(X) = \mu < \infty$, then with probability 1 we have:

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mu, \quad \text{as } n \rightarrow \infty.$$

The LLN ‘guarantees’ that long-run averages behave as we expect them to:

$$\mathbb{E}(X) \approx \frac{1}{n} \sum_{i=1}^n X_i.$$

Central Limit Theorem (CLT)

why this assumption?

Consider a collection X_1, \dots, X_n of iid rvs with $\mathbb{E}(X) = \mu < \infty$ and $\text{var}(X) = \sigma^2 < \infty$. Let,

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

Then

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

follows a $N(0, 1)$ distribution as $n \rightarrow \infty$.

This is an extremely important theorem!

It provides the 'magic' that will make statistical analysis work.

Example

Let X_1, \dots, X_{25} be iid rvs where $X_i \sim \text{Exp}(\lambda = 1/5)$.

Recall that $\mathbb{E}(X) = 5$.

Thus, the LLN implies

$$\bar{X} \rightarrow \mathbb{E}(X) = 5.$$

Moreover, since $\text{var}(X) = 1/\lambda^2 = 25$, we have

$$\bar{X} \approx N\left(\frac{1}{\lambda}, \frac{1}{n\lambda^2}\right) = N\left(5, \frac{5^2}{25}\right)$$

fairly small n is enough
to use CLT.

Is $n = 25$ large enough?

A simulation exercise

Generate $B = 1000$ samples of size n . For each sample compute \bar{x} .
The continuous curve is the normal $N(5, 5^2/n)$ distribution prescribed
by the CLT.

Sample 1: $x_1^{(1)}, \dots, x_n^{(1)} \rightarrow \bar{x}^{(1)}$

Sample 2: $x_1^{(2)}, \dots, x_n^{(2)} \rightarrow \bar{x}^{(2)}$

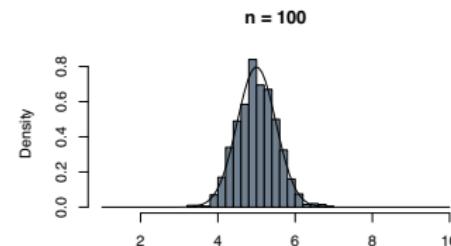
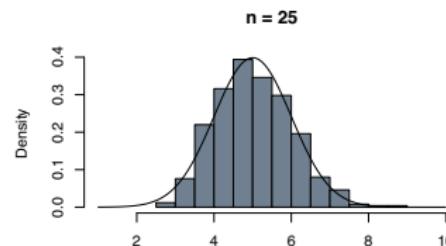
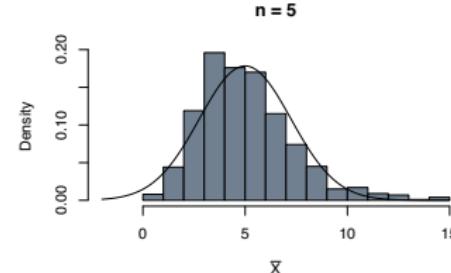
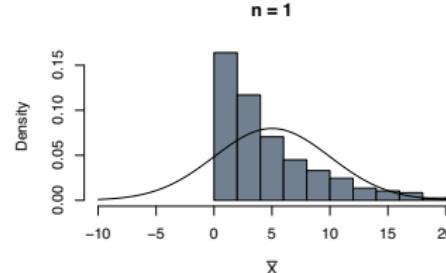
⋮

Sample B: $x_1^{(B)}, \dots, x_n^{(B)} \rightarrow \bar{x}^{(B)}$

Then represent the distribution of $\{\bar{x}^{(b)}, b = 1, \dots, B\}$ by a histogram.

A simulation exercise

The distribution of \bar{X} approaches the theoretical distribution (CLT). Moreover it will be more and more concentrated around μ (LLN). To see this, note that $\text{var}(\bar{X}) = \sigma^2/n \rightarrow 0$ as $n \rightarrow \infty$.



Challenge problem

Let X_1, X_2, \dots, X_{25} be iid rvs with pdf $f(x) = ax^3$ where $0 < x < 2$.

1. What is the value of a ?
2. Calculate $\mathbb{E}(X_1)$ and $\text{var}(X_1)$.
3. What is an approximate value of $\Pr(\bar{X} < 1.5)$?

$$\begin{aligned} 1. \quad & \int_0^2 ax^3 dx = 1 \\ & a \left[\frac{1}{4}x^4 \right]_0^2 = 1 \\ & \Rightarrow a \cdot 4 = 1 \\ & \Rightarrow a = \frac{1}{4} \end{aligned}$$

$$\begin{aligned} 2. \quad & \mathbb{E}(X_1) \\ &= \int_0^2 x \cdot f(x) dx \\ &= \int_0^2 x \cdot \frac{1}{4}x^3 dx \\ &= \frac{1}{4} \int_0^2 x^4 dx \\ &= \frac{1}{4} \left[\frac{1}{5}x^5 \right]_0^2 \\ &= \frac{1}{4} \cdot \left[\frac{32}{5} \right] = \frac{8}{5} \end{aligned} \quad \begin{aligned} & \mathbb{E}(\bar{X}) \\ &= \int_0^2 \bar{x} \cdot f(x) dx \\ &= \int_0^2 \bar{x} \cdot \frac{1}{4}x^3 dx \\ &= \frac{1}{4} \int_0^2 \bar{x} x^3 dx \\ &= \frac{1}{4} \int_0^2 \bar{x} \cdot \frac{1}{4}\bar{x}^3 d\bar{x} \\ &= \frac{1}{16} \int_0^2 \bar{x}^4 d\bar{x} \\ &= \left[\frac{1}{24}\bar{x}^5 \right]_0^2 \\ &= \frac{1}{24} \cdot 32 = \frac{8}{3} \end{aligned}$$

$$\begin{aligned} 3. \quad & \text{from CLT:} \\ & \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1) \\ & \frac{\bar{X} - \frac{8}{5}}{\sigma/5} \rightarrow N(0, 1) \\ & \bar{X} \rightarrow N\left(\frac{8}{5}, \frac{8}{75} \cdot \frac{1}{25}\right) \\ & \Pr(\bar{X} < 1.5) \\ &= \Pr\left(\bar{X}_5 < \frac{1.5 - \frac{8}{5}}{\sqrt{\frac{8}{75} \cdot \frac{1}{25}}}\right) \\ &= \Phi\left(\frac{1.5 - \frac{8}{5}}{\sqrt{\frac{8}{75} \cdot \frac{1}{25}}}\right) = 0.0623 \end{aligned}$$

$$\text{var} = E(x_1^2) - E(x_1)^2 = \frac{8}{3} - \frac{64}{75} = \frac{200-192}{75} = \frac{8}{75}$$

Outline

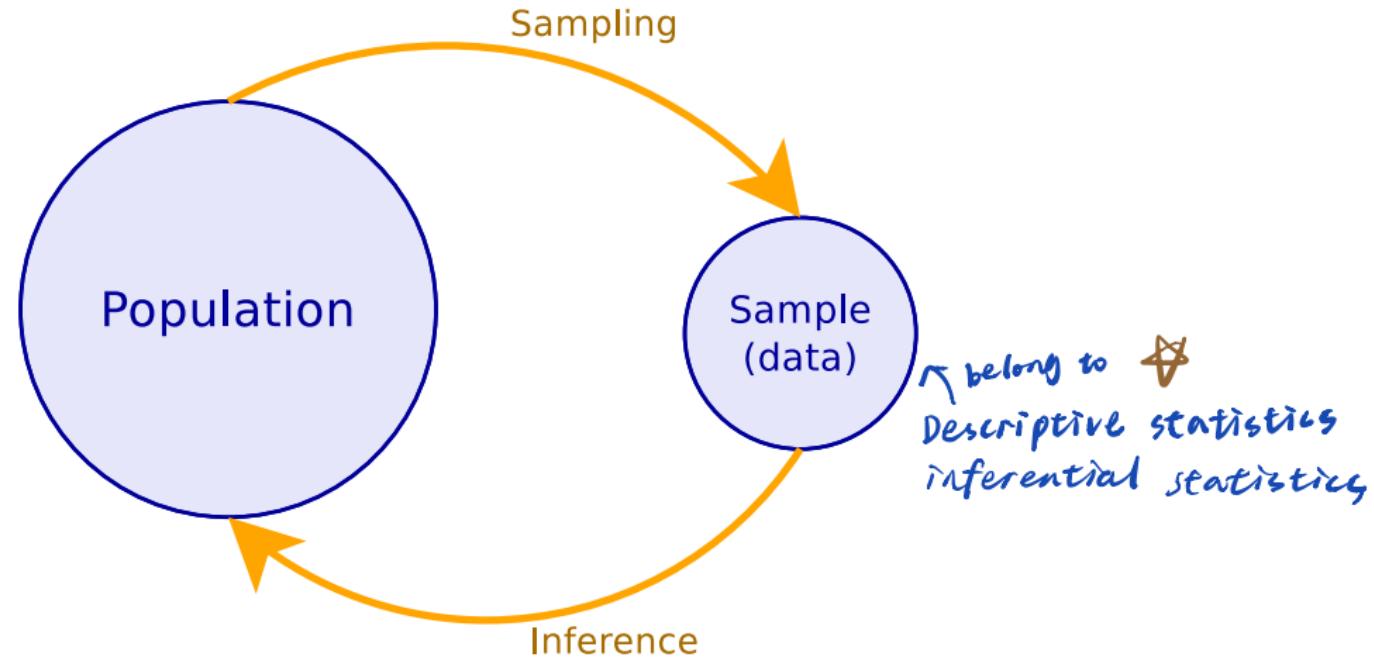
Subject information

Review of probability

Descriptive statistics

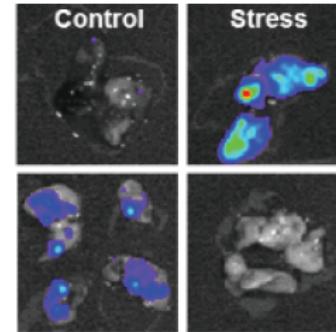
Basic data visualisations

Statistics: the big picture



Example: Stress and cancer

- An experiment gives independent measurements on 10 mice
- Mice are divided in control and stress groups
- The biologist considers two different proteins:
 - Vascular endothelial growth factor C (VEGFC)
 - Prostaglandin-endoperoxide synthase 2 (COX2)



Categorical variable

numerical variable

Mouse	Group	VEGFC	COX2
1	Control	0.96718	14.05901
2	Control	0.51940	6.92926
3	Control	0.73276	0.02799
4	Control	0.96008	6.16924
5	Control	1.25964	7.32697
6	Stress	4.05745	6.45443
7	Stress	2.41335	12.95572
8	Stress	1.52595	13.26786
9	Stress	6.07073	55.03024
10	Stress	5.07592	29.92790

Data & sampling

describe sample as random variable
each unit is randomly selected from all possible ones from population
(think of as hypothetical infinite).

- The **data** are numbers:

$$x_1, \dots, x_n$$

- The model for the data is a **random sample**, that is a sequence of

iid rvs:

$$X_1, X_2, \dots, X_n$$

This model is equivalent to random selection from a hypothetical infinite **population**.

- The goal is to use the data to learn about the distribution of the random variables (and, therefore, the population).

Statistic

r.v format

- A **statistic** $T = \phi(X_1, \dots, X_n)$ is a function of the sample and its realisation is denoted by $t = \phi(x_1, \dots, x_n)$. → **realise format**
- Note: the word “statistic” can also be used to refer to both the realisation, t , as well as the random variable, T . Sometime need to be more specific about which one is meant.
- A statistic has two purposes:
 - Describe or summarise the sample — **descriptive statistics**
 - Estimate the distribution generating the sample — **inferential statistics**
- A statistic can be both descriptive and inferential, it depends on how you wish to use/interpret it (see later)
- We now introduce some commonly used descriptive statistics...

Moment statistics

Sample mean = $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{23.59}{10} = 2.359$

Sample variance = $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 3.98761$

Sample standard deviation = $s = \sqrt{3.98761} = 1.9969$

These are 'sample' or 'empirical' versions of moments of a random variable

Empirical means 'derived from the data' *property from data set*

Order statistics

Arrange the sample x_1, \dots, x_n in order of increasing magnitude and define:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

Then $x_{(k)}$ is the *kth order statistic*.

Special cases:

- $x_{(1)}$ is the *sample minimum*
- $x_{(n)}$ is the *sample maximum*

For the example data,

$$x_{(1)} = 0.52, \quad x_{(2)} = 0.73, \quad \dots, \quad x_{(10)} = 6.07$$

↑ without "i" means arbitrary data

What is $x_{(3.25)}$?

Let it be 0.25 of the way from $x_{(3)}$ to $x_{(4)}$,

$$\begin{aligned}x_{(3.25)} &= x_{(3)} + 0.25 \cdot (x_{(4)} - x_{(3)}) \\&= 0.96 + 0.25 \cdot (0.97 - 0.96) \\&= 0.9625\end{aligned}$$

In other words, define it via linear interpolation.

Exercise: verify that $x_{(7.75)} = 3.6480$ $x_{(7.75)} = x_{(8)} - 0.25(x_{(8)} - x_{(7)})$

Why do this? It allows us to define...

Sample quantiles

General definition ('Type 7' quantiles):

→ for sample, use ^ notation

$$\hat{\pi}_p = x_{(k)}, \text{ where } k = 1 + (n - 1)p$$

pth quantile

Special cases:

$$\text{Sample median} = \hat{\pi}_{0.5} = x_{(5.5)} = \frac{1.26 + 1.53}{2} = 1.395$$

$$\text{Sample 1st quartile} = \hat{\pi}_{0.25} = x_{(3.25)} = 0.9625$$

$$\text{Sample 3rd quartile} = \hat{\pi}_{0.75} = x_{(7.75)} = 3.6480$$

Also:

$$\text{Interquartile range} = \hat{\pi}_{0.75} - \hat{\pi}_{0.25} = 2.685$$

$\hat{\pi}_{0.25}$ and $\hat{\pi}_{0.75}$ contain about 50% of the sample between them

Some descriptive statistics in R

```
> x <- round(VEGFC, digit = 2)
```

✓ > x

```
[1] 0.97 0.52 0.73 0.96 1.26 4.06 2.41 1.53 6.07 5.08
```

✓ > sort(x) # order statistics

```
[1] 0.52 0.73 0.96 0.97 1.26 1.53 2.41 4.06 5.08 6.07
```

✓ > summary(x) # sample mean & sample quantiles

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
------	---------	--------	------	---------	------

0.5200	0.9625	1.3950	2.3590	3.6475	6.0700
--------	--------	--------	--------	--------	--------

- ✓ > var(x) # sample variance
[1] 3.98761

- ✓ > sd(x) # sample standard deviation
[1] 1.9969

- ✓ > IQR(x) # interquartile range
[1] 2.685

Frequency statistics

Can also define empirical versions of pdf, pmf, cdf

Will see in the next section...

Outline

Subject information

Review of probability

Descriptive statistics

Basic data visualisations

Box plot

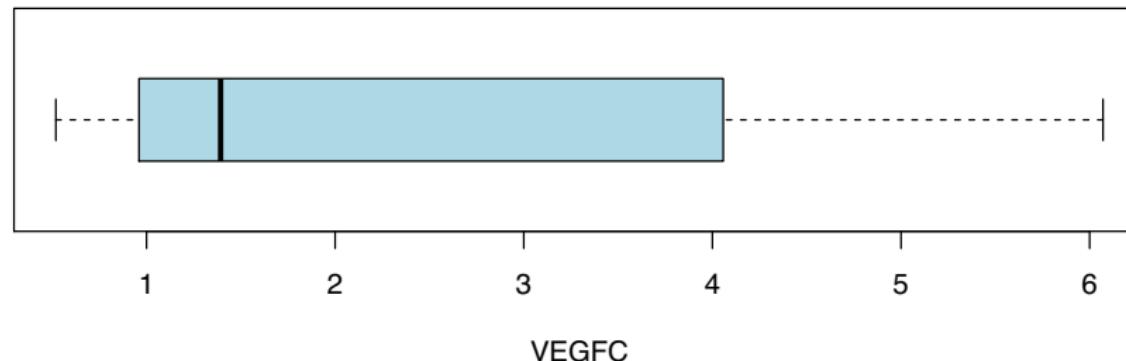
Graphical summary of data from a single variable

Main box: $\hat{\pi}_{0.25}, \hat{\pi}_{0.5}, \hat{\pi}_{0.75}$

'Whiskers': $x_{(1)}, x_{(n)}$

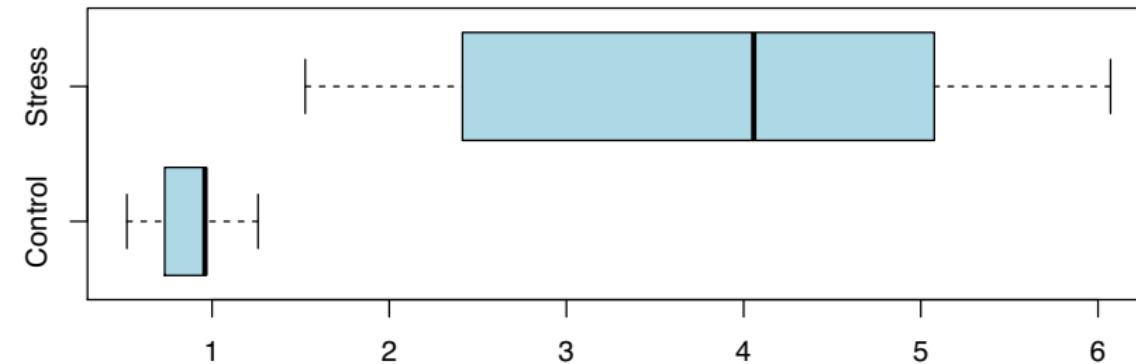
(but R does something more complicated, see tutorial problems)

clear visualization
easily compare for
different size sample



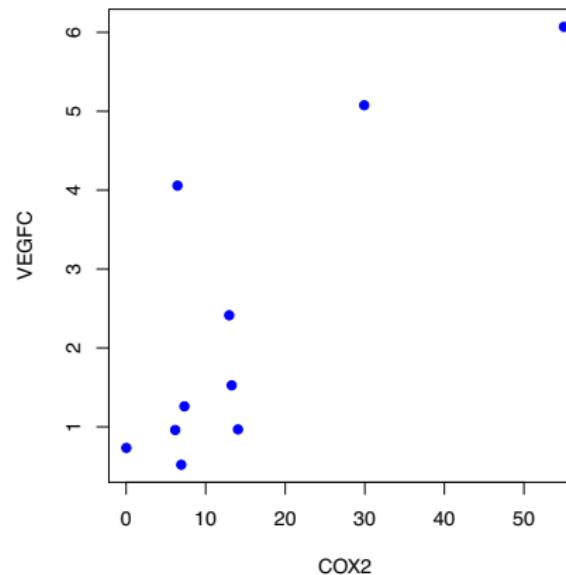
Convenient way of comparing data from different groups

Example: VEGFC (Stress vs Control)



Scatter plot

For comparing data from two variables (usually continuous)



Empirical cdf

The sample cdf, or empirical cdf, is defined as

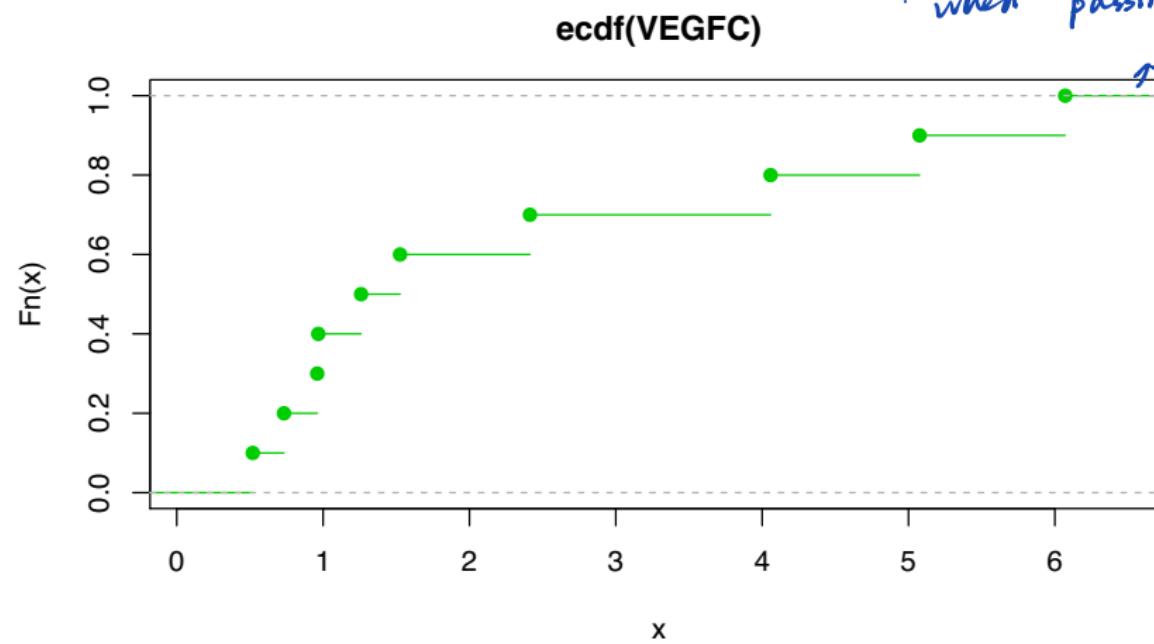
$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x)$$

for sample
what proportion of points are $\leq x$
total number of points
→ how many points $\leq x$

where $I(\cdot)$ is the indicator function ($I(x_i \leq x)$ has value 1 if $x_i \leq x$ and value 0 if $x_i > x$).

For example, for the previous data,

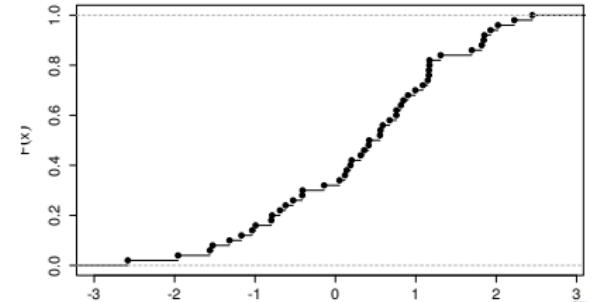
$$\hat{F}(2) = \frac{1}{10} \sum_{i=1}^{10} I(x_i \leq 2) = \frac{6}{10} = 0.6$$



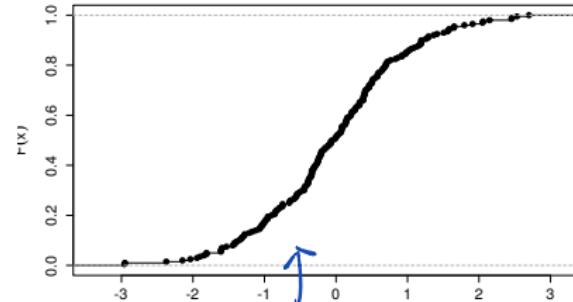
when passing all data points
get value of
 $F(x)$ be 1

It has the form of a discrete cdf. However, it will approximate the cdf of a continuous variable if the sample size is large. The following diagram shows cdfs based on $n = 50$ and $n = 200$ observations sampled from a standard normal distribution, $N(0, 1)$.

empirical cdf
link to probability
distribution version



approximating by step function



approximating smooth curve
from the distribution in which
it sampling from

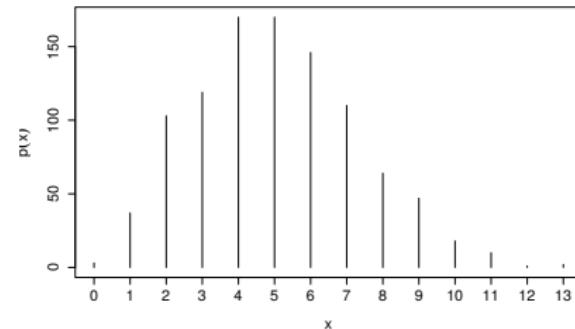
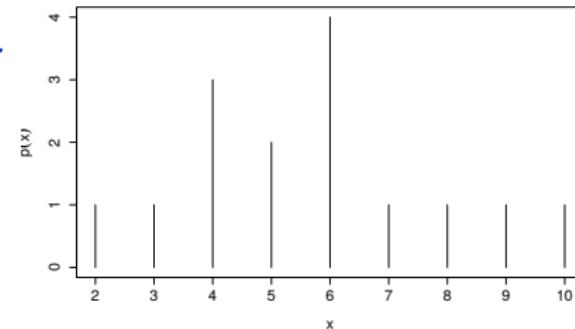
Empirical pmf

If the underlying variable is discrete we use the pmf corresponding to the sample cdf \hat{F}

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^n I(x_i = x)$$

For example, the following shows $\hat{p}(x)$ of size $n = 15$ from $Pn(5)$ (left) and the true pmf $p(x)$ of $Pn(5)$ (right)

if $n \uparrow$
left get closer to right



Histograms and smoothed pdfs

If the underlying variable is continuous we would prefer to obtain an approximation of the pdf. There are several approaches that can be used:

- step function*
1. **Histogram**, \hat{f}_h (h is the bin length). First divide the entire range of values into a series of small intervals (bins) and then count how many values fall into each interval. For interval $[a, b)$, where $b - a = h$, draw a rectangle with height:

$$\hat{f}_h(x) = \frac{1}{hn} \sum_{i=1}^n I(a \leq x_i < b)$$

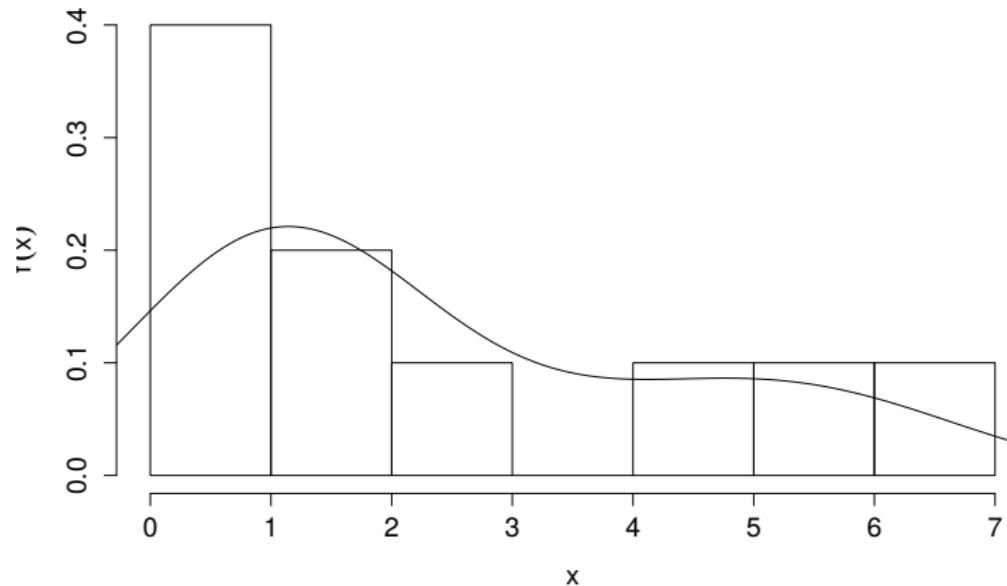
*indicator function,
cause step
count how many points in the interval*

- smooth version
of step function*
2. **Smoothed pdf**, \hat{f}_h (h is the 'bandwidth parameter'),

$$\hat{f}_h(x) = \frac{1}{hn} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right),$$

where $K(\cdot)$ is the **kernel** (a non-negative function that integrates to 1 and with mean zero) and h is a parameter that controls the level of smoothing.

Example: VEGFC

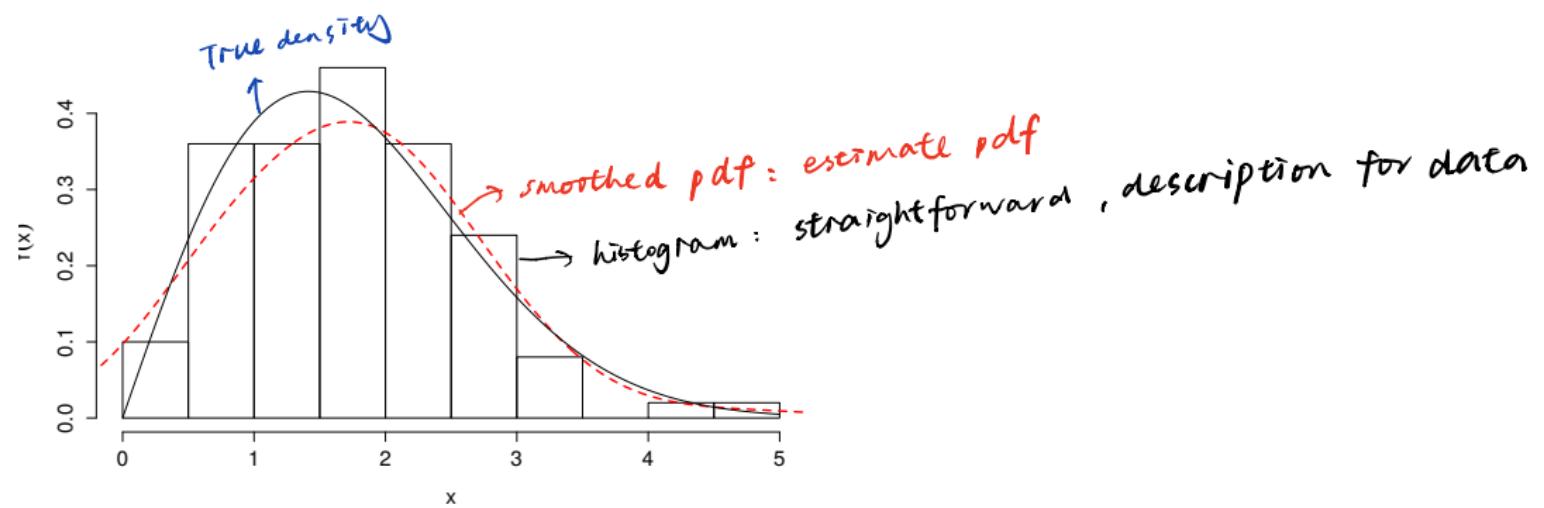


Simulated data

Consider $n = 100$ observations from the Weibull distribution with pdf

$$f(x) = \frac{1}{2}xe^{-(x/2)^2}, x > 0$$

True density (solid black curve), smoothed pdf (red dashed curve)



Quantile-quantile (QQ) plots

- For comparing the similarity of two probability distributions
- We plot their quantiles against each other (as a scatter plot)
- Typically, we compare data against a theoretical distribution
- The points in the plot are $(\hat{\pi}_p, \pi_p)$
- **Note:** some people plot these the other way around: $(\pi_p, \hat{\pi}_p)$
sample quantile *actual quantile (theoretical from distribution)*.

- One axis shows the data, written here as sample quantiles:

$$\hat{\pi}_p = x_{(k)}, \quad \text{where } p = \frac{k}{n+1} \quad (\text{'Type 6' quantiles})$$

for $k = 1, \dots, n$

- Other axis shows corresponding quantiles for a theoretical distribution:

 $\pi_p = F^{-1}(p) = F^{-1}\left(\frac{k}{n+1}\right)$

- The points in the plot therefore are,

$$\left\{ x_{(k)}, F^{-1}\left(\frac{k}{n+1}\right) \right\}.$$

Example: VEGFC

Is the sample from an exponential distribution with cdf
 $F(x) = 1 - e^{-\lambda x}$?

Sample quantiles (data):

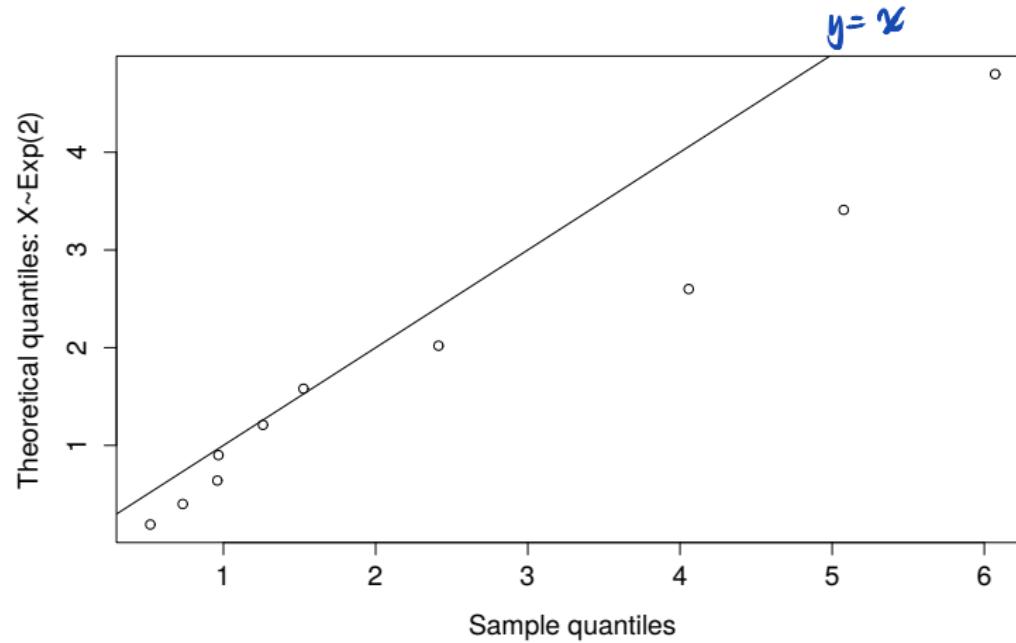
$$x_{(1)} = 0.52, x_{(2)} = 0.73, \dots, x_{(10)} = 6.07$$

Theoretical quantiles:

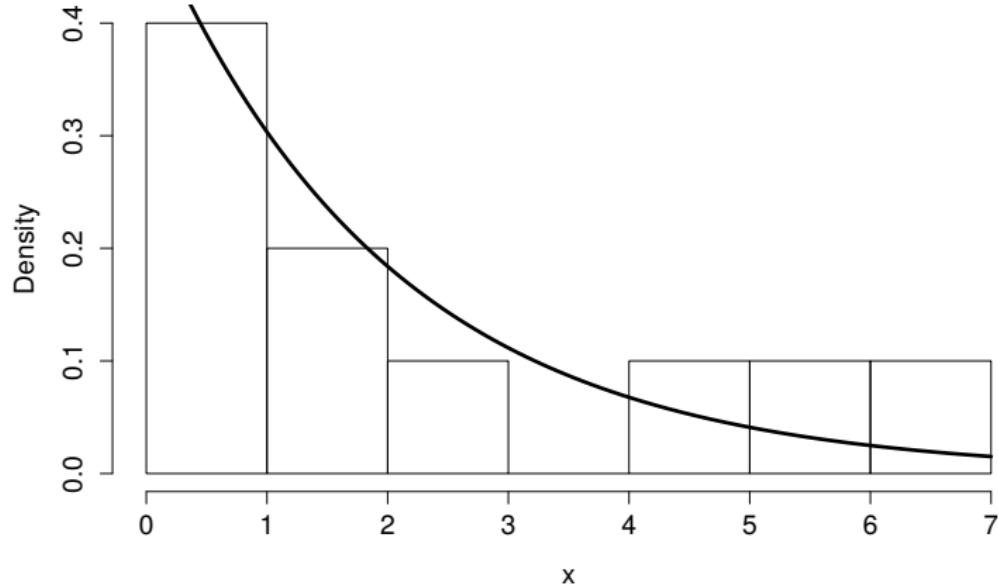
$$F^{-1}(p) = -\ln(1 - p)/\lambda \quad (\text{e.g. set } \lambda = 0.5)$$

$$1/(10 + 1) = 0.09, 2/(10 + 1) = 0.18, \dots, 10/(10 + 1) = 0.91$$

$$F^{-1}(0.09) = 0.19, F^{-1}(0.18) = 0.40, \dots, F^{-1}(0.91) = 4.80$$



The right tail of the sample does not quite match the theoretical model (tail of the sample distribution is heavier).



The right tail of the sample does not quite match the theoretical model (tail of the sample distribution is heavier).

Normal QQ plots

standardize

If $X \sim N(\mu, \sigma^2)$, then $X = \mu + \sigma Z$, where $Z \sim N(0, 1)$. Therefore, if the normal model is correct

$$x_{(k)} \approx \mu + \sigma \Phi^{-1} \left(\frac{k}{n+1} \right) \rightarrow \text{quantile of } Z.$$

where $\Phi(z) = P(Z \leq z)$ is the standard normal cdf.

So, if we plot the points

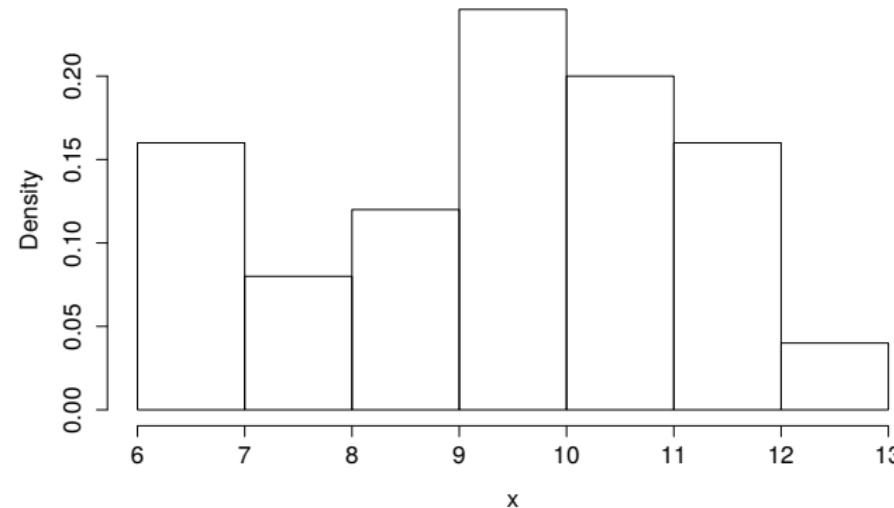
$$\left(x_{(k)}, \Phi^{-1} \left(\frac{k}{n+1} \right) \right), \quad k = 1, \dots, n$$

the result should be a straight line with intercept μ and slope σ .
The values $\Phi^{-1}(k/(n+1))$ are called **normal scores**.

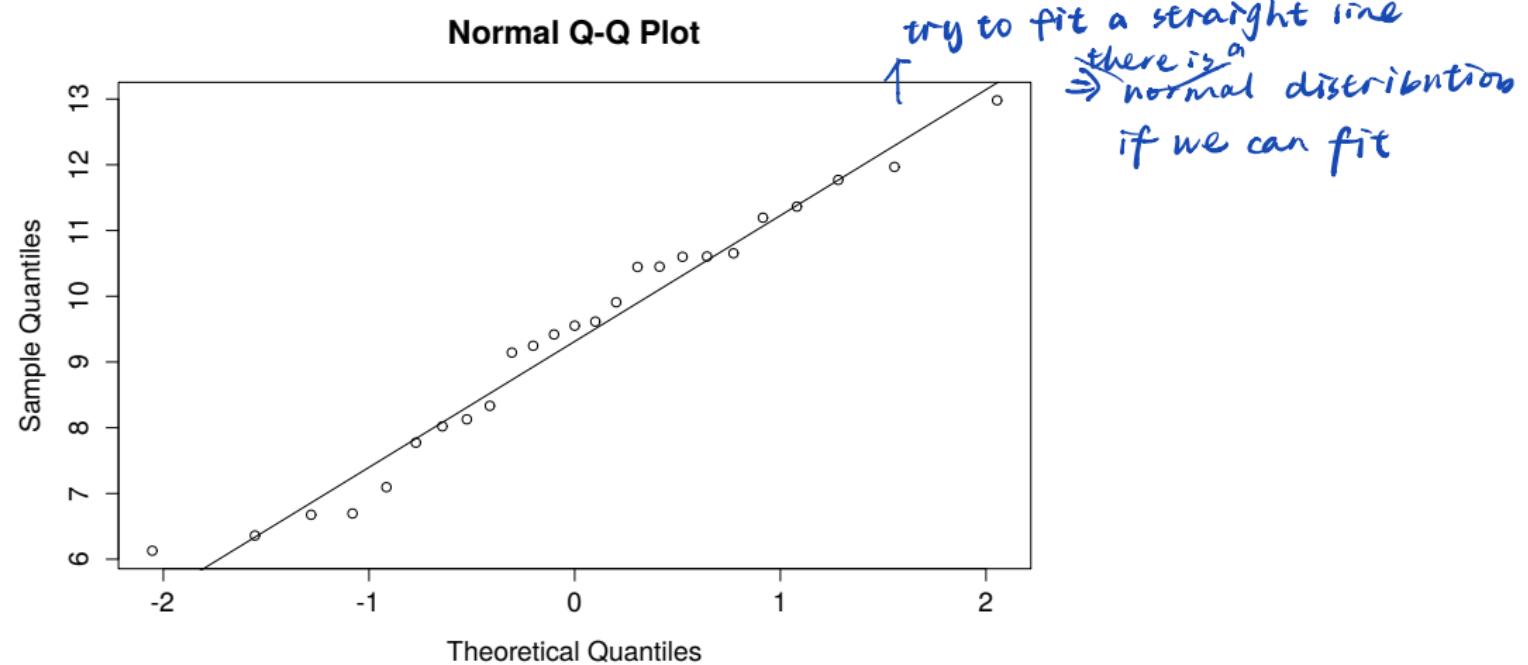
Example: simulated data

Consider 25 observations from $X \sim N(10, 2)$.

The histogram is (not very helpful)



But the QQ plot is much clearer:



when use GeoGebra

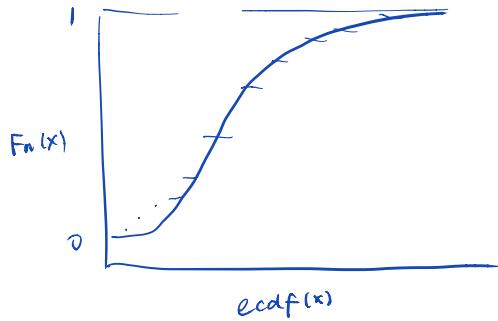
even if we know it samples from $N(\mu, \sigma^2)$, it sometimes still have variation
(may not fit into a straight line)

tail deviation may cause from small sample size

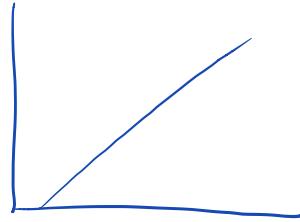
deviation is more easily in the tail → some right theoretical distribution may have some tail deviation

How a QQ plot relates to the empirical CDF (Type b Quantile)

?



⇒ use Quantile QQ plot



hard to compare curves