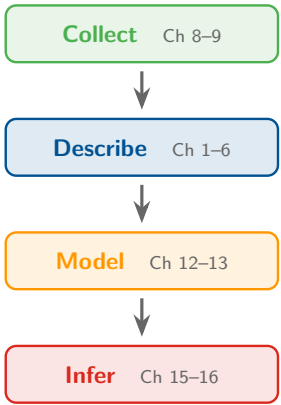
DS 1000B

Data Science Concepts

Last Lecture — Winter 2026

DS 1000B in One Picture



- Gather data with sound methods
- Summarise and visualise.
  Explore the data prior to modelling it.
- Quantify uncertainty with probability
- Infer what the data suggests about the world beyond the data

One Example, Four Stages

A researcher asks: how much sleep do university students get?

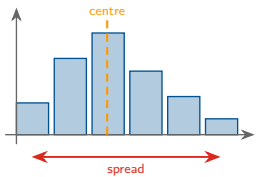
Collect Survey 200 students. SRS from the registrar's list.

Describe $\bar{x} = 6.8$ hours, $\sigma = 1.4$ hours

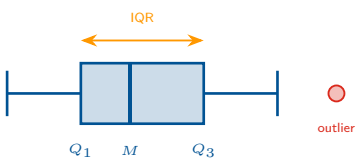
Model Sampling distribution of $\bar{x}$: approximately Normal with $\text{SE} = 1.4/\sqrt{200} \approx 0.10$.

Infer 95% CI: $6.8 \pm (1.96)(0.10) = 6.8 \pm 0.20$, so between 6.6 and 7.0 hours.

Chapter 1 Picturing Distributions with Graphs

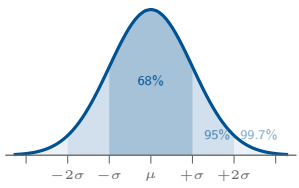


- individuals, variables, variable types
- histograms, bar charts, pie charts, stemplots
- describing distributions
- timeseries plots

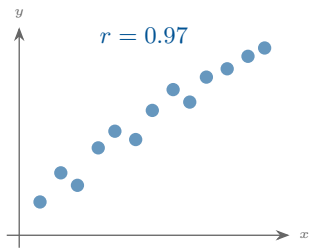


- mean, median, mode
- variance, standard deviation, IQR
- boxplots and modified boxplots
- summary measures

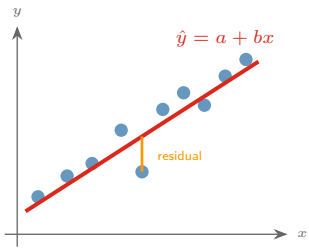
Chapter 3 The Normal Distribution



- density curves
- Normal distribution $N(\mu, \sigma)$
- empirical rule
- calculate and interpret a z-score
- transforming to the standard Normal
- percentiles, probability calculations



- scatterplots
- correlation coefficient



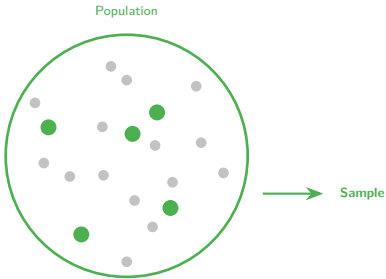
- slope and y-intercept interpretation
- predictions
- coefficient of determination R^2
- residuals
- residual plots for assessing model fit
- influence and leverage

Chapter 6 Contingency Tables

	Dog	Cat	Total
Male	1	4	**5**
Female	3	1	**4**
Total	**4**	**5**	**9**

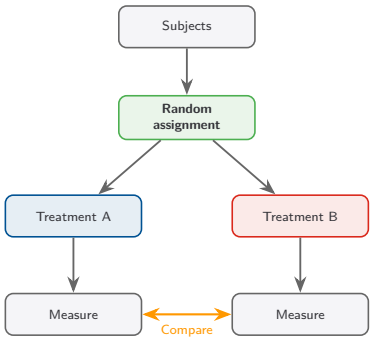
- contingency tables
- marginal and conditional distributions
- independence of two categorical variables
- relative risk, odds ratio
- Simpson's paradox

Chapter 8 Sampling



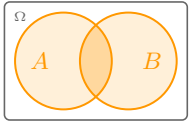
- population vs. sample; parameter vs. statistic
- problems with non-random sampling methods
- SRS, stratified, systematic, cluster, and multistage sampling
- practical issues with good designs

Chapter 9 Designing Experiments



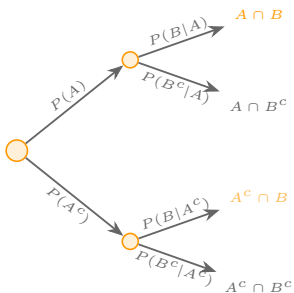
- observational studies vs. experiments
- explanatory and response variables
- factors, levels, treatments, and subjects
- principles of control, blocking, randomisation, and replication
- design types: CRD, block design, matched pairs
- blinding and placebo effects

Chapter 12 Introduction to Probability

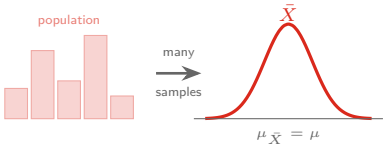


- outcomes, events, sample space, and probability
- algebra of events
- addition and complement rules
- Venn diagrams
- random variables

Chapter 13 Conditional Probability

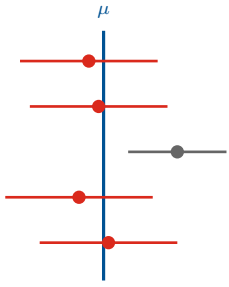


- conditional probability
- multiplication rule
- independent events
- tree diagrams
- Bayes' theorem
- Law of Total Probability



- sampling distribution of a statistic
- the behaviour of the sampling distribution of $\bar{X}$

Chapter 16 Confidence Intervals: The Basics



- CI interpretation
- CI construction
- relationship between CI and sample size, confidence level and population variability
- CI for a transformed parameter

Western Science

DS 1000B Final Exam Self-Assessment & Study Plan

Winter 2026 Final Exam: Sunday, April 19, 2026 10:00 AM - 1:00 PM

The goal is to help you gather an accurate assessment of your strengths and weaknesses across the course material, and to use that information to create a focused study plan for the final exam.

Unit 1 Exploring Data Chapters 1 – 6

Chapter 1 Picturing Distributions with Graphs

I can ...	Not yet	Developing	Proficient
Distinguish categorical variables from quantitative variables	○	○	○
Read and interpret a histogram	○	○	○
Read and interpret a bar chart or pie chart for categorical data	○	○	○
Interpret the features of a time series-plot: trends, seasonality and outliers	○	○	○

Chapter 2 Describing Distributions with Numbers

I can ...	Not yet	Developing	Proficient
Calculate and interpret the mean, median and quartiles of a dataset	○	○	○
Calculate and interpret the IQR and standard deviation	○	○	○
State and apply the properties of standard deviation	○	○	○
Identify outliers from a modified boxplot	○	○	○
Read and construct a boxplot or a modified boxplot	○	○	○
Choose appropriate summary measures for a distribution	○	○	○

DS1000B – Notation and Terminology Glossary

Winter 2026

Contents

Quick Reference: Common Notation

Symbol	Name	Meaning
$\bar{x}$	x-bar	Sample mean
s		Sample standard deviation
$\hat{p}$	p-hat	Sample proportion
n		Sample size
μ	mu	Population mean
σ	sigma	Population standard deviation
p		Population proportion
Q_1, Q_3		First, third quartile
$\hat{y}$	y-hat	Predicted value
ϵ_i		Residual
a, b		Slope, intercept
r		Correlation coefficient
R^2		Coefficient of determination

Symbol	Name	Meaning
$P(A)$		Probability of event A
$P(A \mid B)$		Prob. of A given B
$A \cap B$		Intersection (A and B)
$A \cup B$		Union (A or B)
$\overline{A}$, A^c, A'		Complement (not A)
$N(\mu, \sigma)$		Normal distribution
Z		Standard Normal variable
$\Phi(z)$		Cumul. Normal prob. $P(Z \leq z)$
z^*		Critical value
$\bar{X}_n$		Sample mean (random var.)
SE		Standard error
MOE		Margin of error

Ch.1 – Picturing Distributions with Graphs

Bar graph

Displays distribution of a categorical variable using rectangular bars.

Categorical variable
Places individuals into groups or categories.

Centre
A typical or representative value for the distribution.

Dataset
Structured collection of data on individuals and their variables.

Distribution

The values a variable can take and how often it takes them.

Exploratory Data Analysis (EDA)
Using graphs and summaries to describe variables and relationships.

Histogram

Visualisation for quantitative data: x-axis divided into bins, bar height shows count.

Individuals
The objects described by a set of data.

Observation
A single row in a dataset; all variable values for one individual.

Outlier
A value that lies far from the rest of the data. ▷ Outlier, Ch.4

Pie chart

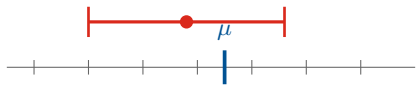
Circle divided into slices proportional to category frequencies.

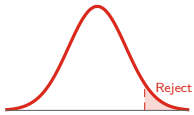
Quantitative variable
Takes numerical values where arithmetic makes sense.

Shape
Overall pattern of a distribution: symmetric, left skewed or right skewed.

Where It Leads

Inference



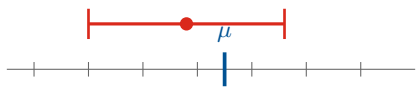


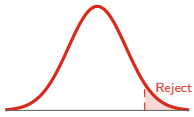
Hypothesis testing

Is the observed effect real,

or just noise?

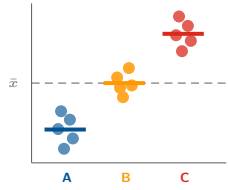
Inference





Hypothesis testing

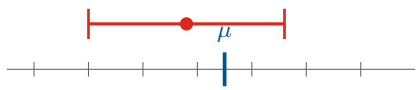
Is the observed effect real, or just noise?

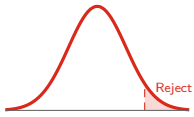


ANOVA

Is there a difference in centre between the groups?

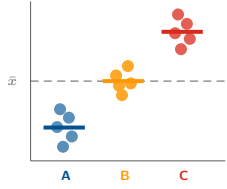
Inference





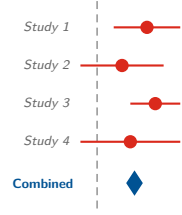
Hypothesis testing

Is the observed effect real, or just noise?



ANOVA

Is there a difference in centre between the groups?

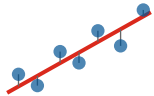


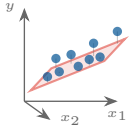
Meta-analysis

Several studies have conflicting conclusions.

How can we combine their evidence, and what does it say?

Regression



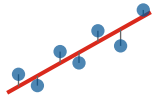


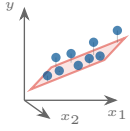
Multiple regression

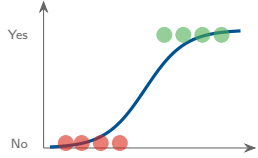
What if we have more than one independent variable?

SS 3859 Regression
SS 4860 Advanced Regression

Regression







Multiple regression

What if we have more than one independent variable?

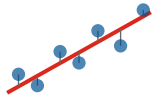
SS 3859 Regression
SS 4860 Advanced Regression

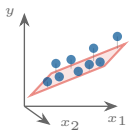
Logistic regression

How can we predict other types of dependent variables?

SS 3860 Generalized Linear Models
SS 3869 Applied Linear Models
BIOSTATS 3110 Multivariable Methods

Regression

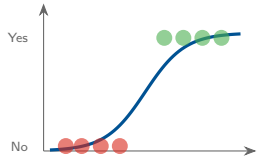




Multiple regression

What if we have more than one independent variable?

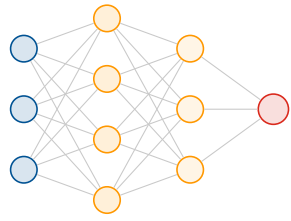
SS 3859 Regression
SS 4860 Advanced Regression



Logistic regression

How can we predict other types of dependent variables?

SS 3860 Generalized Linear Models
SS 3869 Applied Linear Models
BIOSTATS 3110 Multivariable Methods

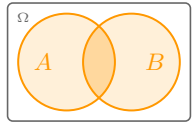


Neural networks

How does AI translate languages or generate images?

DS 3000 Intro to ML
CS 4452 Deep Learning
AM 4264 Intro to Neural Networks

Probability



$\mathbb{E}[X]$

Expectation

$\mathrm{Var}(X)$

Variance

$\mathbb{E}[\mathbb{E}[X \mid Y]]$

Law of Total Expectation

$M_Z(t) = e^{\frac{t^2}{2}}$

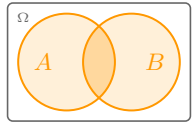
Moment generating functions

Probability theory

Why is the Central Limit Theorem true? How to make decisions under uncertainty? What are good betting strategies?

SS 2857 Probability I
SS 3657 Intermediate Probability

Probability



$\mathbb{E}[X]$

Expectation

$\mathrm{Var}(X)$

Variance

$\mathbb{E}[\mathbb{E}[X \mid Y]]$

Law of Total Expectation

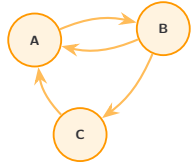
$M_Z(t) = e^{\frac{t^2}{2}}$

Moment generating functions

Probability theory

Why is the Central Limit Theorem true? How to make decisions under uncertainty? What are good betting strategies?

SS 2857 Probability I
SS 3657 Intermediate Probability

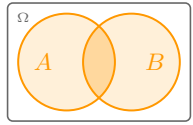


Markov chains

How does Google rank billions of webpages?

SS 4654 Markov Chains

Probability



$\mathbb{E}[X]$ Expectation

$\mathrm{Var}(X)$ Variance

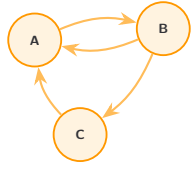
$\mathbb{E}[\mathbb{E}[X \mid Y]]$ Law of Total Expectation

$M_Z(t) = e^{\frac{t^2}{2}}$ Moment generating functions

Probability theory

Why is the Central Limit Theorem true? How to make decisions under uncertainty? What are good betting strategies?

SS 2857 Probability I
SS 3657 Intermediate Probability

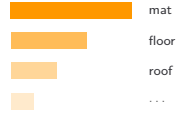


Markov chains

How does Google rank billions of webpages?

SS 4654 Markov Chains

The cat sat on the . . .



Large Language Models

Conditional probability:

$P(\text{next word} \mid \text{context})$

AISE 4010 Deep Learning (Transformers)
CS 3346 Intro to AI

Before You Go

Show Up

- Attend your classes
- Ask questions in lecture, in office hours, and beyond

Build Experience

- Pursue internships and research opportunities
- Develop skills beyond your coursework

Adapt

- AI
- Struggle is part of learning

Thank You

It has been a genuine pleasure teaching you this course.

Thank you for your curiosity, your effort, and your questions. They made the lectures a lot more enjoyable.

Good luck on the final and in everything that comes next. Don't hesitate to reach out.

One Last Ask

Your feedback shapes how this course is taught in the future.

Please take a few minutes to complete the **SQCT course evaluations** they are anonymous and I go through them all carefully.

go.blueja.io/Uu165VRnRUC1AZaZnmGHhA

https://go.blueja.io/Uu165VRnRUC1AZaZnmGHhA