DS1000B – Notation and Terminology Glossary

Winter 2026

Contents

Quick Reference: Common Notation

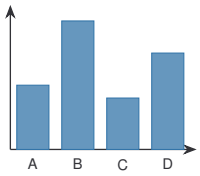
Symbol	Name	Meaning
$\bar{x}$	x-bar	Sample mean
s		Sample standard deviation
$\hat{p}$	p-hat	Sample proportion
n		Sample size
μ	mu	Population mean
σ	sigma	Population standard deviation
p		Population proportion
Q_1, Q_3		First, third quartile
$\hat{y}$	y-hat	Predicted value
e_i		Residual
a, b		Intercept, slope
$corr(x, y), r$		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
$\Phi(z)$		Left tail Normal probability $P(Z \leq z)$
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

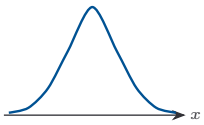
A typical or representative value for the distribution.

Dataset

Structured collection of data on individuals and their variables.

Distribution

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



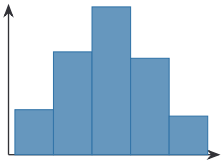
x	Count
1	2
2	7
3	12
4	6
5	3

Exploratory Data Analysis (EDA)

Using graphs and summaries to describe variables and relationships.

Histogram

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



Individuals

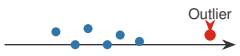
The objects described by a set of data.

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

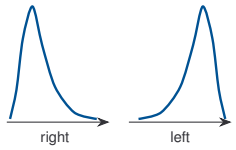
Takes numerical values where arithmetic makes sense.

Shape

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

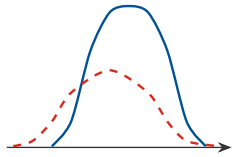
Skewed distribution

A distribution with one tail longer than the other; named by the direction of the longer tail.



Spread

How much data values can differ from one another.



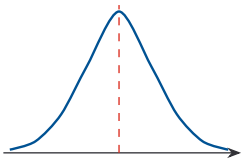
Stemplot

Splits each observation into a stem (all but last digit) and leaf (last digit).

1	235
2	1468
3	03

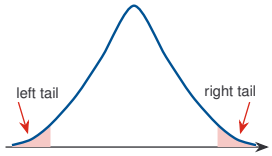
Symmetric distribution

A distribution whose left and right sides are approximately mirror images.



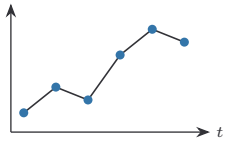
Tail of a distribution

The far left or far right end of a distribution, where observations become increasingly sparse.



Time plot (time series)

Shows how a variable changes over time; time on x-axis.



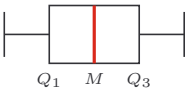
Variable

Any characteristic of an individual that can take different values.

Ch.2 – Describing Distributions with Numbers

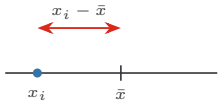
Boxplot

Graphical summary based on the five-number summary; box spans Q_1 to Q_3.



Deviation

Difference between an observation and the mean: $x_i - \bar{x}$.



First quartile · Q_1

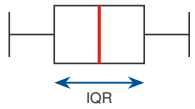
Value below which 25% of data fall.

Five-number summary

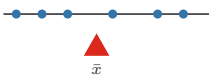
Min, Q_1, Median, Q_3, Max arranged in order.

Interquartile range · IQR $= Q_3 - Q_1$

Range of the middle 50% of the data.



Mean



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

Sum of all values divided by number of observations. ▷ Population mean, Ch.3; Sample mean, Ch.15

Median · Q_2

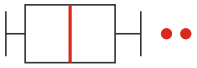
Middle value when data are sorted; 50% of observations fall below.

Mode

Value that appears most frequently; the only centre measure usable for categorical data.

Modified boxplot

Boxplot displaying potential outliers as individual points; whiskers extend only to fences.



Numerical summary

Reduces a distribution to a few key numbers capturing centre and variability.

Potential outlier (boxplot)

Value more than $1.5 \times$ IQR beyond Q_1 or Q_3.

Quartiles · Q_1, Q_2, Q_3

Divide a sorted dataset into four equal parts.

Range · Range $=$ Max $-$ Min

Difference between the largest and smallest values; simplest measure of spread.

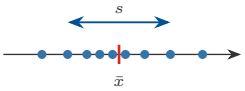
Resistant measure

Statistic not sensitive to a few extreme observations (outliers).

Second quartile · Q_2

The median; 50% of data fall below.

Standard deviation



$$s = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

Square root of variance; in same units as data. ▷ Population standard deviation, Ch.3; Sample standard deviation, Ch.15

Third quartile · Q_3

Value below which 75% of data fall.

Variance

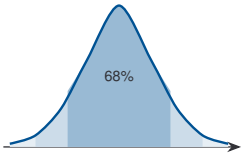
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Average squared deviation from the mean.

Ch.3 – The Normal Distribution

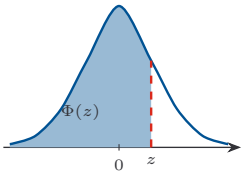
Empirical rule / 68–95–99.7 Rule

In $N(\mu, \sigma)$: 68% within $\mu \pm \sigma$, 95% within $\mu \pm 2\sigma$, 99.7% within $\mu \pm 3\sigma$.



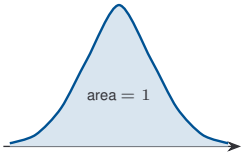
Cumulative Normal probability · $\Phi(z) = P(Z \leq z)$

Probability that $Z \sim N(0,1)$ is at most z.



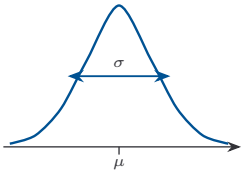
Density curve

A function allowing to compute probabilities for a continuous random variable. The density curve is nonnegative and area under the curve is 1.



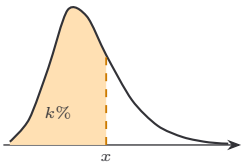
Normal distribution · $X \sim N(\mu, \sigma)$

Symmetric bell-shaped density curve determined by mean μ and SD σ.



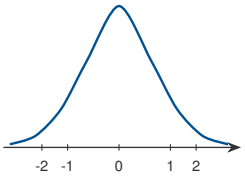
Percentile

The kth percentile is the value x such that $k\%$ of observations fall at or below x.



Standard Normal distribution · $Z \sim N(0, 1)$

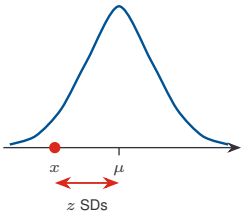
Normal distribution with mean 0 and standard deviation 1.



z-score

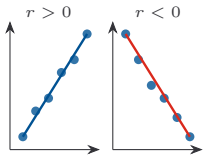
$$z = \frac{x - \mu}{\sigma}$$

Number of standard deviations an observation x lies from the mean μ.



Ch.4 – Scatterplots and Correlation

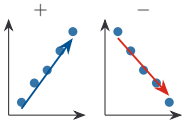
Correlation coefficient



$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

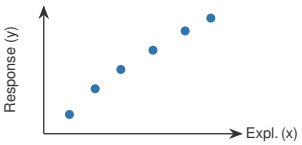
Measures strength and direction of linear association; $-1 \le r \le 1$.

Direction (scatterplot)



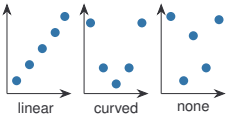
Positive: y increases as x increases; negative: y decreases.

Explanatory variable · x



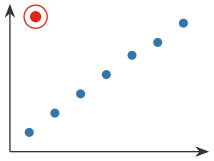
Variable used to explain or predict the response; plotted on the x-axis. ▷ Explanatory variable, Ch.9

Form (scatterplot)



Overall shape of the relationship.

Outlier (Ch.4)

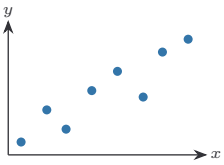


A point that falls far from the overall pattern of the scatterplot. ▷ Outlier, Ch.1

Response variable · y

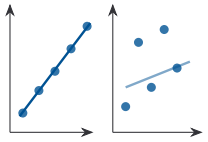
The outcome of interest; plotted on the y-axis. ▷ Response variable, Ch.9

Scatterplot



Displays the relationship between two quantitative variables as plotted points (x_i, y_i).

Strength (scatterplot)



How closely data points follow the overall pattern (strong, moderate, weak).

Ch.5 – Regression

Coefficient of determination · R^2

Proportion of variation in y explained by x via the regression line.

Ecological fallacy

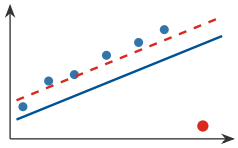
Incorrectly inferring individual behaviour from aggregate (group-level) data.

Extrapolation

Using a regression line to predict y for values of x outside the observed range; unreliable because the linear pattern may not continue.

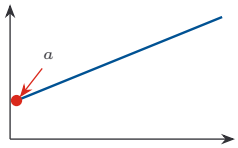
Influential point

Observation whose removal would substantially change the regression line.



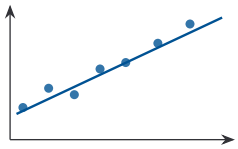
Intercept · $a = \bar{y} - b\bar{x}$

Predicted value of y when $x = 0$.



Least squares regression line · $\hat{y} = a + bx$

Line minimising the sum of squared vertical distances between observed and predicted values.

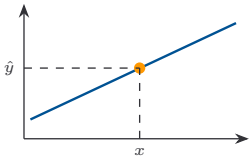


Lurking variable

Unobserved variable influencing both explanatory and response variables.

Predicted value · $\hat{y}$

Value of y estimated from the regression equation ($\hat{y}$-hat).

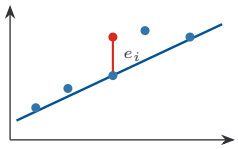


Regression line

Straight line describing how y changes as x changes; used for prediction.

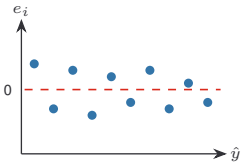
Residual · $e_i = y_i - \hat{y}_i$

Difference between an observed value and its predicted value.



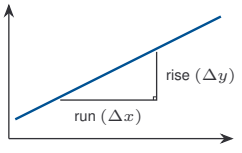
Residual plot

Plot of residuals versus predicted values (or x) to assess fit.



Slope · $b = r \cdot \frac{s_y}{s_x}$

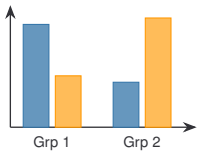
Predicted change in y for a one-unit increase in x.



Ch.6 – Contingency Tables

Conditional distribution

Distribution of a variable within a specific group defined by another variable.



Confounding variable

Measured variable influencing both explanatory and response variables. ▷ Confounding variable, Ch.9

Independence (categorical)

Two categorical variables are independent if conditional distributions are the same across all levels of the other variable.

Marginal distribution

Distribution of one variable alone, ignoring the other; found in the row or column totals of a two-way table.

	A	B	Tot
G1	35	18	**53**
G2	23	41	**64**
Tot	**58**	**59**	**117**

Odds · $\frac{P}{1-P}$

Ratio of the probability of an outcome to the probability of its complement.

Odds ratio

$$\text{OR} = \frac{\text{Odds}_{\text{exposed}}}{\text{Odds}_{\text{unexposed}}}$$

Compares the odds in two groups; measures strength of association.

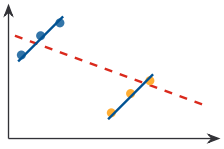
Relative risk

$$\text{RR} = \frac{P(\text{outcome} \mid \text{exposed})}{P(\text{outcome} \mid \text{unexposed})}$$

Ratio of risk in the exposed group to risk in the baseline (unexposed) group.

Simpson's paradox

Trend in aggregate data reverses or disappears when data are split into subgroups.



Two-way table

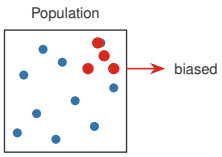
Displays counts for each combination of two categorical variables (also: contingency table).

	A	B
G1	35	18
G2	23	41

Ch.8 – Sampling

Bias

A sampling method is biased if it consistently over- or under-represents part of the population.



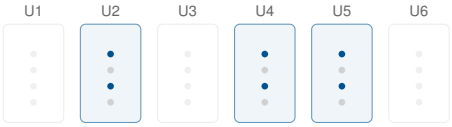
Census
Collects data from every individual in the population.

Cluster sampling
Divides population into clusters, randomly selects clusters, and measures all individuals in selected clusters.

Convenience sample
Sample selected by choosing the easiest members of the population to reach.

Multistage sampling

Sampling plan combining multiple sampling methods in successive stages.



Nonresponse
Occurs when a selected individual cannot be contacted or refuses to participate.

Nonresponse bias
Systematic error arising when non-respondents differ from respondents in relevant ways.

Parameter
Numerical summary describing a population characteristic; typically unknown.

Population
The entire group of individuals we want information about.

Response bias
Systematic pattern of incorrect responses in a sample survey.

Sample
The part of the population we actually observe.

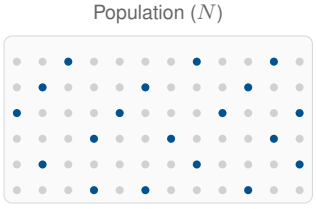
Sampling design
Description of exactly how to choose a sample from the population.

Sampling frame
The list or source used to choose the sample.

Selection bias
Sampling method systematically over- or under-represents certain groups.

Simple random sample · SRS of size n

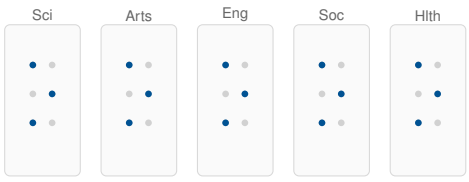
Every set of n individuals has an equal chance of being selected.



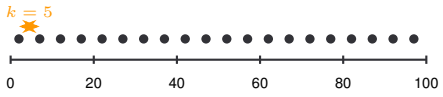
Statistic
Numerical summary computed from sample data; known but varies across samples.

Stratified random sampling

Divides population into non-overlapping strata, then takes an SRS from each.



Systematic sampling



Selects every kth element after a random start.

Under-coverage
Some groups in the population have no chance of being selected.

Voluntary response bias
Bias from self-selection in a voluntary response sample.

Voluntary response sample
Consists of people who choose themselves by responding to a general appeal.

Wording effects
Influence of question wording on answers given in a survey.

Ch.9 – Designing Experiments

Blinding
Keeps subjects, researchers, or both unaware of which treatment a subject receives.

Block
A group of similar subjects formed before random assignment to treatments.

Blocking
Groups similar subjects together before randomly assigning them to treatments.

Completely randomized design · CRD
All subjects are randomly assigned to treatments without prior grouping.

Control group
Group that does not receive the treatment; used for comparison.

Double-blind
Neither subjects nor researchers measuring outcomes know which treatment is applied.

Experiment
Researchers deliberately impose a treatment on subjects and measure the response.

Factor
An explanatory variable that is deliberately manipulated in an experiment.

Level
A specific value of a factor used in an experiment.

Matched pairs design
Block design where each block has exactly two units; one assigned to each treatment.

Observational study
Researchers observe and record data without manipulating any variable.

Placebo
A fake treatment given to the control group so subjects cannot tell which group they are in.

Randomization
Randomly assigning subjects to treatment and control groups so groups are comparable before treatment.

Randomised block design
Subjects grouped into blocks of similar units; then randomly assigned to treatments within each block.

Replication
Using enough subjects in each group so results are not easily explained by chance alone.

Response variable (Ch.9)
The outcome measured in an experiment. ▷ Response variable, Ch.4

Single-blind
Either subjects or researchers (but not both) do not know which treatment is applied.

Treatment
A specific combination of factor levels applied to a subject.

Treatment group
Group that receives the intervention being studied.

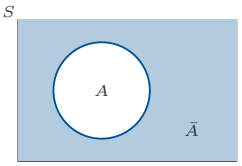
Ch.12 – Introduction to Probability

Addition rule (general) · $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Probability of A or B for any two events.

Addition rule (mutually exclusive) · $P(A \cup B) = P(A) + P(B)$
Holds when A and B cannot both occur.

Complement · $\overline{A}$ or A^c or A'

The event containing all outcomes not in A.

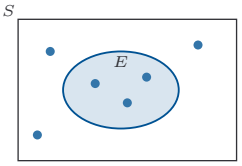


Complement rule · $P(\overline{A}) = 1 - P(A)$
Probability of the complement equals one minus the probability of the event.

Continuous probability model
Probability model with continuous sample space; probabilities described by a density curve.

Event

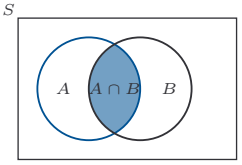
A set of outcomes from a sample space.



Finite probability model
Probability model with a finite number of outcomes; each outcome assigned a probability.

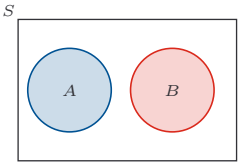
Intersection · $A \cap B$

Outcomes in both A and B; "A and B."



Mutually exclusive (disjoint)

Two events that cannot both occur; they share no outcomes.



Outcome

A single possible result of a random phenomenon.

Probability · $P(A)$

Number between 0 and 1 measuring how likely event A is to occur.

Probability distribution

Describes which values a random variable can take and how likely each is.

Random phenomenon

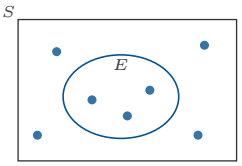
Any process whose result is uncertain before it occurs.

Random variable

Assigns a numerical value to each outcome of a random phenomenon.

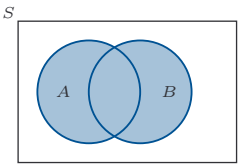
Sample space · S

The set of all possible outcomes of a random phenomenon.



Union · $A \cup B$

All outcomes in A, in B, or in both; "A or B."



Venn diagram

Represents events as regions inside a rectangle (the sample space); overlapping regions show intersections.

Ch.13 – Conditional Probability

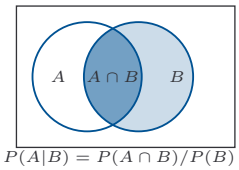
Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

Reverses the direction of a conditional probability.

Conditional probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



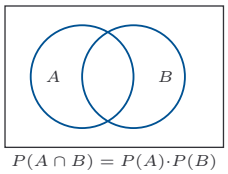
Probability of A given that B has occurred.

"Given" bar · $\mid$

Read as "given"; denotes conditioning: $P(A \mid B)$ = "probability of A given B."

Independence (events) · $P(A \cap B) = P(A) \cdot P(B)$

Events A and B are independent if knowing one does not change the probability of the other.



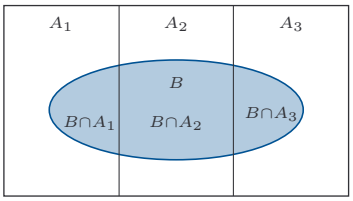
Joint probability · $P(A \cap B)$

Probability that both events A and B occur.

Law of Total Probability

$$P(B) = \sum_i P(B \mid A_i)\, P(A_i)$$

Expresses $P(B)$ using a partition of the sample space.



Marginal probability · $P(A)$ or $P(B)$

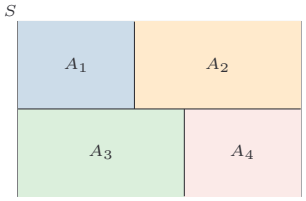
Probability of a single event, ignoring other events.

Multiplication rule · $P(A \cap B) = P(B) \cdot P(A \mid B)$

Expresses joint probability via a conditional and a marginal probability.

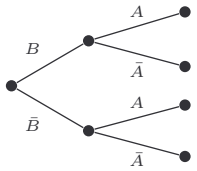
Partition

Mutually exclusive and exhaustive collection of events $A_1, A_2, \ldots, A_k$ that together cover all of S.



Tree diagram

Displays all possible outcomes of a sequence of events, with each branch labelled by a conditional probability.

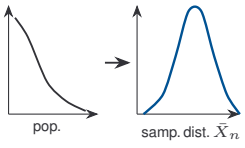


Ch.15 – Sampling Distributions

Central Limit Theorem (CLT)

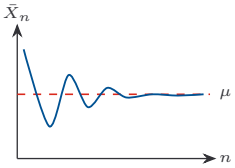
$$\bar{X}_n \approx N\left(\mu,\, \frac{\sigma}{\sqrt{n}}\right)$$

For large n, the sampling distribution of $\bar{X}_n$ is approximately Normal regardless of the population's shape.



Law of Large Numbers (LLN)

As n increases, the sample mean $\bar{X}_n$ converges to the population mean μ with high probability.



Parameter vs. Statistic

A *parameter* is a fixed number describing the population (e.g. μ, σ, p); a *statistic* is a number computed from a sample (e.g. $\bar{x}$, s, $\hat{p}$) and varies from sample to sample.

Population proportion · p

True proportion in the entire population; a parameter.

Sample mean · $\bar{x}$ or $\bar{X}_n$

Mean computed from a sample. Lowercase $\bar{x}$ denotes a specific sample; uppercase $\bar{X}_n$ denotes the random variable representing the sample mean.

Sample proportion · $\hat{p}$

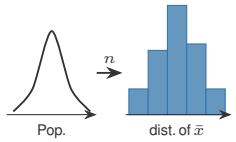
Proportion computed from a sample; estimates p.

Sample standard deviation · s

Standard deviation computed from a sample; estimates σ (see Population standard deviation, Ch.3).

Sampling distribution

Distribution of values taken by a statistic over all possible samples of the same size.

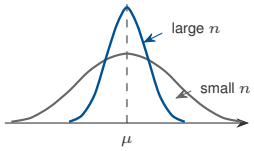


Sampling variability

The natural tendency for a statistic to take different values in different samples, even when drawn from the same population. The sampling distribution quantifies this variability.

Standard error of the mean

$$\mathrm{SE}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$



Standard deviation of the sampling distribution of the sample mean.

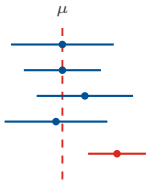
Unbiased estimator · $E(\bar{X}_n) = \mu$

A statistic whose sampling distribution is centred on the true parameter value; $\bar{X}_n$ is an unbiased estimator of μ.

Ch.16 – Confidence Intervals: The Basics

Confidence interval for μ

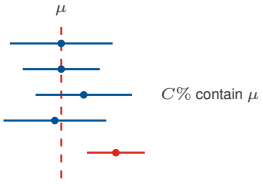
$$\bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$$



Interval estimate for the population mean when σ is known.

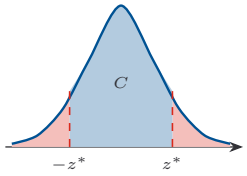
Confidence level · C

In repeated sampling, $C\%$ of all resulting intervals will contain the true parameter.



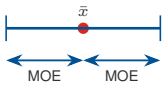
Critical value · z^*

Value such that the standard Normal curve has area C between $-z^*$ and z^*.



Margin of error

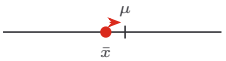
$$\text{MOE} = z^* \cdot \frac{\sigma}{\sqrt{n}}$$



Amount added to and subtracted from the point estimate to form the interval.

Point estimate · $\bar{x}$ or $\hat{p}$

Single number computed from data serving as the estimate for the unknown parameter.



Common Points of Confusion

1. Why $n-1$ in the sample standard deviation?

The sample variance divides by $n-1$, not n. This is called *Bessel's correction*. Dividing by n would systematically underestimate the true population variance σ^2 because the deviations are measured from $\bar{x}$ (which was computed from the same sample) rather than from the true mean μ. Dividing by $n-1$ compensates, making s^2 an *unbiased* estimator of σ^2. The denominator $n-1$ is called the *degrees of freedom*: once you know $\bar{x}$ and $n-1$ of the deviations, the last deviation is determined.

2. Greek letters for parameters, Latin letters for statistics

A notational convention runs through the entire course:

Quantity	Parameter (population)	Statistic (sample)
Mean	μ	$\bar{x}$
Standard deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r

Parameters are fixed but unknown; statistics are computed from data and vary from sample to sample. When we write $\bar{X}_n$ (capital, subscripted) we are treating the sample mean as a *random variable* whose behaviour across all possible samples is described by its sampling distribution.

3. Sample standard deviation s vs. population standard deviation σ

Both measure spread, but they play different roles at different points in the course:

1. **Describing data (Ch. 2).** You compute s from the sample to summarise how spread out the observations are. Here s is a purely descriptive number.
2. **Modelling with the Normal distribution (Ch. 3).** The Normal model $N(\mu, \sigma)$ uses σ as a fixed population parameter that controls the width of the bell curve. We implicitly make the assumption that we know σ in this chapter.
3. **Building regression (Ch. 4–5).** The slope formula $b = r \cdot s_y/s_x$ uses sample standard deviations. Regression (Ch. 5) is entirely descriptive: we are fitting a line through observed data points, with no population model, no inference, and no assumptions about a bell-curve shape. Because we are computing from data, we use s, not σ.
4. **Sampling distributions and confidence intervals (Ch. 15–16).** The standard error $\mathrm{SE} = \sigma/\sqrt{n}$ and the confidence interval $\bar{x} \pm z^* \cdot \sigma/\sqrt{n}$ both use σ, which we treat as known. A follow-up course relaxes this assumption by replacing σ with s and the z^* critical value with a t^* critical value from the Student's t-distribution.

Rule of thumb: if you are computing a number from your data, use s. If you are working inside a probability model or a theoretical formula, use σ.

4. Standard deviation vs. Standard error

The two measure different things:

- **Standard deviation (s)** describes the spread of *individual observations* in a dataset.
- **Standard error (SE $= \sigma/\sqrt{n}$)** describes the spread of a *statistic* (e.g. $\bar{X}_n$) across repeated samples.

Because the SE divides by $\sqrt{n}$, it shrinks as the sample size grows, even when the population spread σ stays the same.

5. Mutually exclusive $\neq$ Independent

Students often mix these up:

- **Mutually exclusive:** $P(A \cap B) = 0$. If A happens, B cannot.

- **Independent:** $P(A \cap B) = P(A) \cdot P(B)$. Knowing A tells you nothing about B.

If two events with non-zero probabilities are mutually exclusive, knowing that A occurred tells you B did *not* occur, so they are *not* independent. The two properties are incompatible (unless one event has probability zero).

6. Confounding variable vs. Lurking variable

Both affect the relationship between explanatory and response variables. The difference:

- A **confounding variable** is *measured* and known to be associated with both variables.
- A **lurking variable** is *not measured* (and may not even be identified) but still influences both.

In an observational study, confounders and lurking variables make causal conclusions unreliable. Randomised experiments handle both by distributing all variables (measured or not) equally across treatment groups.

Alphabetical Index

Ch. 1 2 | Ch. 2 3 | Ch. 3 5 | Ch. 4 6 | Ch. 5 6 | Ch. 6 8 | Ch. 8 8 | Ch. 9 10 | Ch. 12 11 | Ch. 13 12 | Ch. 15 13 | Ch. 16 14