## Introduction

### Population

The entire group we are concerned with.

### Sample

A representative subset of the population.

### Quantitative Data

Numerical data. Could be nominal (discrete or continuous), or ordinal (ordered).

### Qualitative Data

Categorical data, e.g. colour, model.

## Measures of Centrality

### Mean

Given a set of $n$ observations $x_1, x_2, \dots, x_n$, the **arithmetic mean** or **average** is defined as

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

The sample mean is denoted $\overline{x}$. The population mean is denoted $\mu$.

### Median

A drawback to the mean is that it can be misleading when the data is skewed. The **median** is the middle value of a set of $n$ observations when arranged from largest to smallest.
If $n$ is odd:

$$\text{median} = x^{\left(\frac{n+1}{2}\right)}$$

or the $(n+1)/2$th value of the sorted list. If $n$ is even, the median is the :

$$\text{median} = \frac{x^{\left(\frac{n}{2}\right)} + x^{\left(\frac{n}{2}+1\right)}}{2}$$

### Mode

Given discrete data, the mode is defined as the most common value in a set of observations.

## Measures of Dispersion

Dispersion refers to how much variation there is in a set of observations.

### Range

The range is the difference between the maximum and minimum observation.

### Variance

The variance is the average of the squared deviations from the mean.

- Given the observations $x_1, x_2, \dots, x_N$, from a population of size $N$ with mean $\mu$, the **population variance** is defined as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2.$$

- Given the observations $x_1, x_2, \dots, x_n$, from a sample of size $n$ with mean $\overline{x}$, the **sample variance** is defined as

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

The **population variance** is given by

### Standard Deviation

The standard deviation is the square root of the variance. The **population standard deviation** is defined as $\sigma = \sqrt{\sigma^2}$. The **sample standard deviation** is defined as $s = \sqrt{s^2}$.

**Theorem 3.3.1** (Chebyshev's Theorem)**.** *Given a set of n observations, at least*

$$1 - \frac{1}{k^2}$$

*of them are within k standard deviations of the mean, where $k \geq 1$.*

**Theorem 3.3.2** (Empirical Rule)**.** *If a histogram of the data is approximately unimodal and symmetric, then,*

- *68% of the data falls within **one** standard deviation of the mean*
- *95% of the data falls within **two** standard deviations of the mean*
- *99% of the data falls within **three** standard deviations of the mean*

Often the standard deviation cannot be computed directly, but can be approximated using the Empirical rule. Here we assume that

$$\text{range} \approx 4s$$

so that

$$s = \frac{\text{range}}{4}.$$

### Skew

The **skew** describes the asymmetry of the distribution. For a finite population of size $N$, the **population skew** is defined as

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^3$$

For a sample of size $n$, the **sample skew** is defined as

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^3}$$

- When the skew is **positive**, the data is **right-skewed** and the "tail" of the distribution is **longer on the right**
- When the skew is **negative**, the data is **left-skewed** and the "tail" of the distribution is **longer on the left**

## Measures of Rank

It is often useful to know the rank or *relative standing* of a value in a set of observations. This is natural for ordinal data whose ordering has implicit meaning, but it can also be useful for nominal data as a means of measuring dispersion.

### Z-Score

The Z-score is a unitless quantity and can be used to make comparisons of relative rank between members of a population.

$$Z = \frac{x - \mu}{\sigma} \quad \text{or} \quad \frac{x - \overline{x}}{s}$$

### Quantiles

In addition to Z-scores, quantiles can be used to make comparisons of relative ranking between populations, as well as construct intervals bounding a given proportion of the observations. For a set of $n$ observations, $x_q$ is the $q$-th quantile, if $q$% of the observations are less than $x_q$.

### Inter-Quartile Range

The inter-quartile range (IQR) is the difference between the 75th and 25th quantiles, or the range covered by the middle 50% of data. It is a robust measure of the dispersion of the data, as it is not affected by extreme values unlike the range or variance.

## Boxplots

### Five Number Summary

The five number summary is set of measurements that indicates the

- minimum value
- 25% quartile
- median
- 75% quartile
- maximum value

### Outliers

Outliers are extreme observations that fall outside some interval defined either by quantiles (above 95% or below 5% quantiles) or in terms of the Empirical rule (outside two standard deviations from the mean). They should be investigated to determine if they are errors or naturally occurring extreme values.

### Covariance and Correlation Coefficients

Covariance is the measure of the linear correlation between variables. For variables $x$ and $y$,

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

Note that when $x = y$, the formula simplifies to the sample variance of $x$. The covariance has the following characteristics:

- $s_{xy} > 0$: As $x$ increases, $y$ also increases.
- $s_{xy} < 0$: As $x$ increases, $y$ decreases.
- $s_{xy} \approx 0$: No relationship between $x$ and $y$.

Although the covariance is a useful tool to measure relationships, it is only generalisable in terms of its sign. Thus, if we want to compare across data sets, we need to use the correlation coefficient.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

The **correlation coefficient** is a measure of the strength of the relationship between the variables. It is a scale-free and unitless measure bounded between $-1$ and 1 and has the same characteristics as the covariance.
Note that a correlation coefficient of 0 indicates **no linear relationship** between the variables, and not necessarily indicative of **no relationship**.

## Regression and Least Squares

A regression or least squares line of best fit provides both a graphical and numerical summary of the relationship between the variables. A linear relationship between two variables $x$ and $y$ is defined as $y = a + bx$. The least squares best fit determines the coefficients $a$ and $b$ that minimise the sum of the squares of the residuals (errors) between $y$ and the line $\hat{y} = a + bx$. Mathematically, $\min_{a,\,b} \sum_{i=1}^{n} (y_i - \hat{y}(x_i))^2$. The coefficients can be summarised by the formula

$$\begin{aligned} b &= r\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2} \\ a &= \overline{y} - b\overline{x}. \end{aligned}$$

## Events and Probability

### Event

Set of outcomes from an experiment.

### Sample Space

Set of all possible outcomes $\Omega$.

### Intersection

Outcomes occur in both $A$ and $B$

$$A \cap B \qquad \text{or} \qquad AB$$

### Disjoint

No common outcomes, $AB = \varnothing$

$$\Pr(AB) = \Pr(A \mid B) = 0$$

### Union

Set of outcomes in either $A$ or $B$

$$A \cup B$$

### Complement

Set of all outcomes not in $A$, but in $\Omega$

$$\begin{aligned} A\overline{A} &= \varnothing \\ A \cup \overline{A} &= \Omega \end{aligned}$$

### Subset

$A$ is a (non-strict) subset of $B$ if all elements in $A$ are also in $B$ — $A \subset B$.

$$AB = A \qquad \text{and} \qquad A \cup B = B$$

$$\forall A : A \subset \Omega \wedge \varnothing \subset A$$

$$\begin{aligned} \Pr(A) &\leq \Pr(B) \\ \Pr(B \mid A) &= 1 \\ \Pr(A \mid B) &= \frac{\Pr(A)}{\Pr(B)} \end{aligned}$$

### Identities

$$\begin{aligned} A(BC) &= (AB)C \\ A \cup (B \cup C) &= (A \cup B) \cup C \\ A(B \cup C) &= AB \cup AC \\ A \cup BC &= (A \cup B)(A \cup C) \end{aligned}$$

### Probability

Measure of the likeliness of an event occurring

$$\Pr(A) \qquad \text{or} \qquad \mathrm{P}(A)$$

$$0 \leq \Pr(A) \leq 1$$

where a probability of 0 never happens, and 1 always happens.

$$\begin{aligned} \Pr(\Omega) &= 1 \\ \Pr\left(\overline{A}\right) &= 1 - \Pr(A) \end{aligned}$$

### Multiplication Rule

For independent events $A$ and $B$

$$\Pr(AB) = \Pr(A)\Pr(B).$$

For dependent events $A$ and $B$

$$\Pr(AB) = \Pr(A \mid B)\Pr(B)$$

### Addition Rule

For independent $A$ and $B$

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(AB).$$

If $AB = \varnothing$, then $\Pr(AB) = 0$, so that $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

### De Morgan's Laws

$$\begin{aligned} \overline{A \cup B} &= \overline{A}\,\overline{B} \\ \overline{AB} &= \overline{A} \cup \overline{B}. \end{aligned}$$

$$\begin{aligned} \Pr(A \cup B) &= 1 - \Pr\left(\overline{A}\,\overline{B}\right) \\ \Pr(AB) &= 1 - \Pr\left(\overline{A} \cup \overline{B}\right) \end{aligned}$$

### Bayes' Theorem

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$$

## Random Variables

Measurable variable whose value holds some uncertainty. An event is when a random variable assumes a certain value or range of values.

### Probability Distribution

The probability distribution of a random variable $X$ is a function that links all outcomes $x \in \Omega$ to the probability that they will occur $\Pr(X = x)$.

### Probability Mass Function

$$\Pr(X = x) = p_x$$

### Probability Density Function

$$\Pr(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)\,\mathrm{d}x$$

**Cumulative Distribution Function**

Probability that a random variable is less than or equal to a particular realisation $x$.

$F(x)$ is a valid CDF if:

1. $F$ is monotonically increasing and continuous
2. $\lim_{x\to-\infty} F(x) = 0$
3. $\lim_{x\to\infty} F(x) = 1$

$$\frac{\mathrm{d}F(x)}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}\int_{-\infty}^{x} f(u)\,\mathrm{d}u = f(x)$$

**Complementary CDF (Survival Function)**

$$\Pr(X > x) = 1 - \Pr(X \leq x) = 1 - F(x)$$

**$p$-Quantiles**

$$F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u = p$$

**Special Quantiles**

$$\begin{aligned} \text{Lower quartile } q_1\text{:} \quad & p = \frac{1}{4} \\ \text{Median } m\text{:} \quad & p = \frac{1}{2} \\ \text{Upper quartile } q_2\text{:} \quad & p = \frac{3}{4} \\ \text{Interquartile range IQR:} \quad & q_2 - q_1 \end{aligned}$$

**Quantile Function**

$$x = F^{-1}(p) = Q(p)$$

**Expectation (Mean)**

Expected value given an infinite number of observations. For $a < c < b$:

$$\begin{aligned} \mathrm{E}(X) = & -\int_a^c F(x)\,\mathrm{d}x \\ & + \int_c^b (1 - F(x))\,\mathrm{d}x + c \end{aligned}$$

**Variance**

Measure of spread of the distribution (average squared distance of each value from the mean).

$$\mathrm{Var}(X) = \sigma^2 = \mathrm{E}(X^2) - \mathrm{E}(X)^2$$

**Standard Deviation**

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

| **Distribution** | **Restrictions** | **PMF** | **CDF** | $\mathrm{E}(X)$ | $\mathrm{Var}(X)$ |
|---|---|---|---|---|---|
| $X \sim \mathrm{Uniform}(a,\ b)$ | $x \in \{a, \dots, b\}$ | $\frac{1}{b-a+1}$ | $\frac{x-a+1}{b-a+1}$ | $\frac{a+b}{2}$ | $\frac{(b-a+1)^2-1}{12}$ |
| $X \sim \mathrm{Bernoulli}(p)$ | $p \in [0,1], x \in \{0,1\}$ | $p^x(1-p)^{1-x}$ | $1-p$ | $p$ | $p(1-p)$ |
| $X \sim \mathrm{Binomial}(n,\ p)$ | $x \in \{0, \dots, n\}$ | $\binom{n}{x}p^x(1-p)^{n-x}$ | $\sum_{u=0}^{x}\binom{n}{u}p^u(1-p)^{n-u}$ | $np$ | $np(1-p)$ |
| $N \sim \mathrm{Poisson}(\lambda)$ | $n \geq 0$ | $\frac{\lambda^n e^{-\lambda}}{n!}$ | $e^{-\lambda}\sum_{u=0}^{n}\frac{\lambda^u}{u!}$ | $\lambda$ | $\lambda$ |

Table 1: Discrete probability distributions.

| **Distribution** | **Restrictions** | **PDF** | **CDF** | $\mathrm{E}(X)$ | $\mathrm{Var}(X)$ |
|---|---|---|---|---|---|
| $X \sim \mathrm{Uniform}(a,\ b)$ | $a < x < b$ | $\frac{1}{b-a}$ | $\frac{x-a}{b-a}$ | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ |
| $T \sim \mathrm{Exp}(\eta)$ | $t > 0$ | $\eta e^{-\eta t}$ | $1 - e^{-\eta t}$ | $1/\eta$ | $1/\eta$ |
| $X \sim \mathrm{N}(\mu,\ \sigma^2)$ | $x \in \{0, \dots, n\}$ | $\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ | $\frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$ | $\mu$ | $\sigma^2$ |

Table 2: Continuous probability distributions.

| | **Discrete** | **Continuous** |
|---|---|---|
| Valid probabilities | $0 \leq p_x \leq 1$ | $f(x) \geq 0$ |
| Cumulative probability | $\sum_{u \leq x} p_u$ | $\int_{-\infty}^{x} f(u)\,\mathrm{d}u$ |
| $\mathrm{E}(X)$ | $\sum_\Omega x p_x$ | $\int_\Omega x f(x)\,\mathrm{d}x$ |
| $\mathrm{E}(g(X))$ | $\sum_\Omega g(x)\,p_x$ | $\int_\Omega g(x)\,f(x)\,\mathrm{d}x$ |
| $\mathrm{Var}(X)$ | $\sum_\Omega (x-\mu)^2\,p_x$ | $\int_\Omega (x-\mu)^2 f(x)\,\mathrm{d}x$ |

Table 3: Probability rules for univariate $X$.