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 $x_1, x_2, \ldots, x_n$, 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 μ.

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 $x_1, x_2, \ldots, x_N$, from a population of size N with mean μ, the **population variance** is defined as

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

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

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

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

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

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.

Covariance and Correlation Coefficients

Covariance is the measure of the linear correlation between variables.

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

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.

Correlation Coefficient

$$-1 \leq r_{xy} = \frac{s_{xy}}{s_x s_y} \leq 1$$

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 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

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

Events and Probability

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

$$\overline{A \cup B} = \overline{A}\ \overline{B}$$
$$\overline{AB} = \overline{A} \cup \overline{B}.$$

$$\Pr(A \cup B) = 1 - \Pr\left(\overline{A}\ \overline{B}\right)$$
$$\Pr(AB) = 1 - \Pr\left(\overline{A} \cup \overline{B}\right)$$

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)

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

Expectation (Mean)

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

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

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

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

Central Limit Theorem

For a sample of size n from any random probability distribution with expected value μ and variance σ^2,

$$\frac{\sqrt{n}(\overline{x} - \mu)}{\sigma} \xrightarrow{p} \mathrm{N}(0,\ 1)$$

meaning that increasing the sample size will lead to a more normal distribution. In this case, a sample size of $n = 30$ is sufficient to approximate a normal distribution.

Standard Error

$$\mathrm{SE}(\overline{x}) = \frac{\sigma^2}{n}$$

Sample Proportion

For a sample of size n let x be the number of members with a particular characteristic. The sample estimate of the population proportion p is

$$\hat{p} = \frac{x}{n}.$$

By assuming that the samples statistic x follows a binomial distribution with probability p and size n, then $\mathrm{E}(x) = np$ and $\mathrm{Var}(x) = np(1-p)$. Therefore the expectation is

$$\mathrm{E}(\hat{p}) = p$$

and the standard error is

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

For the above to apply, we must assume that the sample proportion and size are sufficiently large. In general, if $np > 5$ and $n(1-p) > 5$, then we can assume that the sampling distribution of $\hat{p}$ is approximately normal.

Assessing Normality

- Histograms: if the data is approximately normal, then the histogram will be approximately symmetric and unimodal.
- Boxplots: boxplots can be useful for showing outliers and skewness. Extreme clusters of an excessive number of outliers can be evidence of non-normality.
- Normal probability plots (q-q plots): these plots are constructed by plotting the sorted data values against their Z-scores. If the data is approximately normal, then the points will lie approximately on a straight line.

Large Sample Estimation

Point Estimation

Method of Moments

The moments of a probability distribution are defined

$$\mu_n = \mathrm{E}(X^n) = \int_{-\infty}^{\infty} x^n f(x)\,\mathrm{d}x$$

where $f(x)$ is the probability density function of the distribution. Here $\mu_1 = \mathrm{E}(X)$ and $\mathrm{Var}(X) = \mu_2 - \mu_1^2$. Sample moments are defined similarly

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

where $\overline{x} = m_1$.

Method of Maximum Likelihood Estimation

Definition 9.1 (Likelihood function)**.**

$$\mathcal{L}\left(\theta \,|\, \boldsymbol{x}\right) = \prod_{i=1}^{n} f\left(x_i\right)$$

Definition 9.2 (Maximum likelihood estimator)**.**

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}\left(\theta \,|\, \boldsymbol{x}\right).$$

Definition 9.3 (Log-likelihood function)**.** The **log-likelihood function** is defined as

$$\ell\left(\theta \,|\, \boldsymbol{x}\right) = \sum_{i=1}^{n} \log\left(f\left(x_i\right)\right)$$

Due to the monotonicity of the log function, the maximum likelihood estimator is the same as the maximum log-likelihood estimator.

Definition 9.4 (Maximum log-likelihood estimator)**.** The **maximum log-likelihood estimator** is defined as

$$\hat{\theta} = \arg\max_{\theta} \ell\left(\theta \,|\, \boldsymbol{x}\right).$$

Properties of Estimators

Definition 9.5 (Bias)**.** The **bias** of an estimator is defined as the difference between the expected value of the estimator $\mathrm{E}\left(\hat{\theta}\right)$ and the true value of the parameter θ_0.

$$\mathrm{Bias}\left(\hat{\theta}\right) = \mathrm{E}\left(\hat{\theta}\right) - \theta$$

An estimator $\hat{\theta}$ is **unbiased** if

$$\mathrm{E}\left(\hat{\theta}\right) = \theta$$

so that the bias is zero.

We can also compare the variance of two estimators, to assess which one is more preferable. If the variance of the estimator is small, then the estimator is more precise. Given two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, we would choose $\hat{\theta}_1$ over $\hat{\theta}_2$ if

$$\mathrm{Var}\left(\hat{\theta}_1\right) < \mathrm{Var}\left(\hat{\theta}_2\right)$$

Definition 9.6 (Mean square error)**.** Given data x_i with variance σ^2, the estimators of $\theta = \mathrm{E}\left\{X\right\}$ are selected such that they minimise the **mean square error**:

$$\begin{aligned}\mathrm{MSE}\left(\hat{\theta}\right) &= \mathrm{E}\left(\left(\hat{\theta} - \theta\right)^2\right) \\ &= \mathrm{Bias}\left(\hat{\theta}\right)^2 + \mathrm{Var}\left(\hat{\theta}\right).\end{aligned}$$

This quantity is used to determine the **bias-variance trade-off** of an estimator. The **root mean square error** is defined as

$$\mathrm{RMSE}\left(\hat{\theta}\right) = \sqrt{\mathrm{MSE}\left(\hat{\theta}\right)}.$$

Confidence Intervals

This interval ranges from the lower confidence limit (UCL) to the upper confidence limit (LCL)

$$L < \theta < U.$$

This interval has a **confidence coefficient** of $1-\alpha$, or a **confidence level** of $(1-\alpha)\,\%$. The confidence interval is defined as

$$CI_{1-\alpha} = \hat{\theta} \pm Z_{\alpha/2}\,\mathrm{SE}\left(\hat{\theta}\right)$$

Confidence Interval for the Mean

$$CI_{1-\alpha} = \overline{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $\mathrm{SE}\left(\overline{x}\right) = \frac{\sigma}{\sqrt{n}}$.

Confidence Interval for the Proportion

Given the sample size n and sample proportion $\hat{p}$,

$$\hat{p} \sim \mathrm{N}\left(p,\ \frac{p\left(1-p\right)}{n}\right).$$

The confidence interval for the population proportion is

$$CI_{1-\alpha} = \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

where the standard error is given by

$$\mathrm{SE}\left(\hat{p}\right) = \sqrt{\frac{p\left(1-p\right)}{n}}.$$

with the approximation $p = \hat{p}$. Note that $n\hat{p} > 5$ and $n\left(1-\hat{p}\right) > 5$ are required for the approximation to be valid.

Confidence Interval for the Difference of Two Means

$$CI_{1-\alpha} = \overline{x}_1 - \overline{x}_2 \pm Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

If the two populations follow a normal distribution, then the sampling distribution is exactly normal. If the two populations are not normal, then the sampling distribution is approximately normal, for $n_1 > 30$ and $n_2 > 30$.

Confidence Interval for the Difference of Two Proportions

$CI_{1-\alpha} = \hat{p}_1 - \hat{p}_2 \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ Note $n_i\hat{p}_i > 5$, and $n_i\left(1-\hat{p}_i\right) > 5$ for $i = 1, 2$ to use this.

Hypothesis Testing

Hypothesis Testing for the Population Mean

Given the sample statistic $\overline{x}$,

$$\overline{x} \sim \mathrm{N}\left(\mu,\ \frac{\sigma^2}{n}\right)$$

the test statistic is defined

$$T\left(\boldsymbol{x}\right) = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}\left(0,\ 1\right).$$

Hypothesis Testing for the Population Proportion

Given the sample statistic $\hat{p}$,

$$\hat{p} \sim \mathrm{N}\left(p,\ \frac{p\left(1-p\right)}{n}\right)$$

for $n\hat{p} > 5$ and $n\left(1-\hat{p}\right) > 5$, the test statistic is defined

$$T\left(\boldsymbol{x}\right) = \frac{\sqrt{n}\left(\hat{p} - p_0\right)}{\sqrt{p_0\left(1-p_0\right)}}.$$

Hypothesis Testing with Differences

Hypothesis Testing for the Difference in Population Means

The point estimator of $\mu_1 - \mu_2$ is given by

$$\overline{x}_1 - \overline{x}_2$$

and the standard error is given by

$$\mathrm{SE}_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

The test statistic is defined

$$T\left(\boldsymbol{x}\right) = \frac{\left(\overline{x}_1 - \overline{x}_2\right) - \Delta_0}{\mathrm{SE}_{\overline{x}_1 - \overline{x}_2}}.$$

where $\Delta_0 = \mu_1 - \mu_2$ is the hypothesized difference between the two population means.

Hypothesis Testing for the Difference in Population Proportions

The point estimator of the difference in proportions where $p_1 = p_2$ is given by

$$\hat{p}_1 - \hat{p}_2$$

and the standard error is defined

$$\mathrm{SE}_{\hat{p}_1 - \hat{p}_2} = \sqrt{p_0\left(1-p_0\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where

$$p_0 = \frac{x_1 + x_2}{n_1 + n_2}$$

$$p_0 = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}.$$

so that $p_0 = p_1 = p_2$. The resulting test statistic is defined:

$$T\left(\boldsymbol{x}\right) = \frac{\left(\hat{p}_1 - \hat{p}_2\right)}{\mathrm{SE}_{\hat{p}_1 - \hat{p}_2}}.$$

Significance of Results

When interpreting the results from a test statistic, the test can only be used to reject the null hypothesis.

Small Sample Inference

Student's t-distribution:

$$T(\boldsymbol{x}) \sim t_\nu$$

where the degrees of freedom ν is equal to $n-1$.

$$\mathrm{E}(X) = 0$$
$$\mathrm{Var}(X) = \frac{\nu}{\nu-2}$$

Inferencing

$$T(\boldsymbol{x}) = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \sim t_{\nu,\alpha/2}.$$

Hypothesis Testing for the Population Mean

$$T(\boldsymbol{x}) = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \sim t_{\nu,\alpha/2}.$$

Hypothesis Testing for the Difference in Population Means

For independent samples,

$$T(\boldsymbol{x}) = \frac{\overline{x}_1-\overline{x}_2}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{\nu,\alpha/2}.$$

If $s_1^2 \neq s_2^2$ use

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{\nu}.$$

where $\nu = n_1+n_2-2$ for the two-sample t-test. The population variances between two samples vary *greatly*, if $\frac{\max(s_1^2,\, s_2^2)}{\min(s_1^2,\, s_2^2)} > 3.$. If so use

$$T(\boldsymbol{x}) = \frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

$$\nu = \left\lfloor \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1}+\frac{\left(s_2^2/n_2\right)^2}{n_2-1}} \right\rfloor$$

ANOVA

An experimental unit is the object whose outcome or response is measured and is of interest in the process of the experiment. The outcome or response measured is called the dependent variable. A factor is an independent variable controlled and varied by the experimenter. A treatment is a specific combination of factor levels. A response is the variable measured by the experimenter, typically continuous numeric.

$$\frac{S_1^2}{s_2^2} \sim F_{\nu_1,\nu_2} \quad \nu_i = n_i - 1$$

$$s_j^2 = \frac{1}{\nu_j}\sum_{i=1}^{n_j}(x_i-\bar{x})^2$$

H_0 : All are same, H_A : At least one is different. Reject if $F > F_{critical} = F_{\nu_1,\nu_2,\alpha}$. $\Pr(F < F_{critical}) = 1-\alpha$.

Tukey's HSD

$$q_{I,I(J-1)} \max_{i_1,i_2} \frac{|(\bar{y}_{i_1}-\mu_{i_1})-(\bar{y}_{i_2}-\mu_{i_2})|}{s_p/\sqrt{J}}$$
$$H_0: \mu_i = \mu_j, \quad H_A: \mu_i \neq \mu_j$$

Reject if $|\bar{y}_{i_1}-\bar{y}_{i_2}| > q_{I,I(J-1),\alpha}\frac{s_p}{\sqrt{J}}$

Blocking

Two-factor Experiments

ANCOVA

Linear Regression

Relationship is linear if $y_i = \beta_0+\beta_1 x_i+\epsilon_i$.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_1-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2}$$
$$s^2 = \frac{\sum_{i=1}^n (y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2}{n-2}$$

Caveat that errors/residuals $e_i \overset{iid}{\sim} N(0,\sigma^2)$.

Testing Inference

Coefficient of Determination

$R^2 = \frac{SSR}{SST} = 1-\frac{SSE}{SST}$. Close to 1 is good but could overfit, small can still be good, weak measure. We can use $F = R^2$ for test.

Regression Diagnostics

Estimation and Prediction

Doing ANCOVA

Categorical Data

2x2 Tables

χ^2 Homogeneity

χ^2 Independence

Decision	H_0	$\neg H_0$
Reject H_0	α	$1-\beta$
Fail to reject	$1-\alpha$	β

H_0	Rejection Region R
$\theta = \theta_0$	$\lvert T(\boldsymbol{x})\rvert > Z_{\alpha/2}$
$\theta \leq \theta_0$	$T(\boldsymbol{x}) > Z_\alpha$
$\theta \geq \theta_0$	$T(\boldsymbol{x}) < -Z_\alpha$

	Discrete	Continuous
$\mathrm{E}(X)$	$\sum_\Omega x p_x$	$\int_\Omega x f(x)\,\mathrm{d}x$
$\mathrm{E}(g(X))$	$\sum_\Omega g(x) p_x$	$\int_\Omega g(x) f(x)\,\mathrm{d}x$
$\mathrm{Var}(X)$	$\sum_\Omega (x-\mu)^2 p_x$	$\int_\Omega (x-\mu)^2 f(x)\,\mathrm{d}x$

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	$np(1-p)$
$N \sim \mathrm{Poisson}(\lambda)$	$n \geq 0$	$\frac{\lambda^n e^{-\lambda}}{n!}$	$e^{-\lambda}\sum_{u=0}^n \frac{\lambda^u}{u!}$	λ	λ

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