Population

The entire group we are concerned with.

Sample

A representative subset of a population.

Quantitative Data

Numerical data. Nominal (discrete or continuous), or ordinal (ordered).

Qualitative Data

Categorical data, e.g. colour, model.

Measures of Centrality

Mean (arithmetic mean/average)

Given n observations x_1, x_2, ..., x_n

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

Sample: $\overline{x}$. Population: μ.

Median (middle value)

The mean can be misleading when the data is skewed. When arranged from smallest to largest:

$$\text{median} = \begin{cases} x^{\left(\frac{n+1}{2}\right)} & n \text{ odd} \\ \frac{x^{\left(\frac{n}{2}\right)}+x^{\left(\frac{n}{2}+1\right)}}{2} & n \text{ even} \end{cases}$$

Mode (most common value)

Measures of Dispersion

Variation in a set of observations.

Range

Difference between max and min value.

Variance

Average squared deviations from mean. **Sample variance:**

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

Population variance:

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

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

Standard Deviation

Population SD: $\sigma = \sqrt{\sigma^2}$.
Sample standard deviation: $s = \sqrt{s^2}$.

Chebyshev's Theorem

Given n observations, at least

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

of them are within $k\sigma$ of μ, where $k \geq 1$.

Empirical Rule

For unimodal and symmetric data

- 68% of data falls within σ of μ
- 95% of data falls within 2σ of μ
- 99% of data falls within 3σ of μ

Approx. range $\approx 4s$ s.t. $s = \frac{\text{range}}{4}$.

Skew

Asymmetry of the distribution.

- **Positive (right-skewed)** — "tail" of distribution is **longer on the right**
- **Negative (left-skewed)** — "tail" of distribution is **longer on the left**

Measures of Rank

Z-Score

Compare 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**: $x_{75} - x_{25}$.

Covariance

Measure of the linear correlation between variables.

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

When $x = y$, $s_{xy} = s^2$.

- $s_{xy} > 0$: 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

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

No linear relationship if $r_{xy} = 0$, but not necessarily **no relationship**.

Events and Probability

Multiplication Rule

For independent events A and B

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

For dependent events A and B

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

Addition Rule

For independent A and B
$\Pr\left(A \cup B\right) = \Pr\left(A\right) + \Pr\left(B\right) - \Pr\left(AB\right)$.

De Morgan's Laws

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

Bayes' Theorem

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

Probability Mass Function

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

Probability Density Function

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

Cumulative Distribution Function

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

Central Limit Theorem

For a sample of size $n \geq 30$,

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

Standard Error

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

Sample Proportion

For a sample of size n if x members have a particular characteristic:

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

If x follows a binomial distribution, then $\mathrm{E}\left(x\right) = np$ and $\mathrm{Var}\left(x\right) = np\left(1-p\right)$.

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

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

For $np > 5$ and $n\left(1-p\right) > 5$.

Large Sample Estimation

Moments

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

for PDF $f\left(x\right)$, and $\mu_1 = \mathrm{E}\left(X\right)$ and $\mathrm{Var}\left(X\right) = \mu_2 - \mu_1^2$. For samples:

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

where $\overline{x} = m_1$ and $s^2 = m_2 - m_1^2$.

Maximum Likelihood Estimation

$$\boldsymbol{\mathcal{L}}\left(\theta \mid \boldsymbol{x}\right) = \prod_{i=1}^{n} f\left(x_i\right),\ \hat{\theta} = \arg\max_{\theta} \boldsymbol{\mathcal{L}}.$$

$$\ell\left(\theta \mid \boldsymbol{x}\right) = \sum_{i=1}^{n} \log\left(f\left(x_i\right)\right),\ \hat{\theta} = \arg\max_{\theta} \ell.$$

Bias

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

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

Given $\hat{\theta}_1$ and $\hat{\theta}_2$, choose $\hat{\theta}_1$ over $\hat{\theta}_2$ if

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

Mean square error

The estimators of $\theta = \mathrm{E}\left(X\right)$ are selected to minimise **mean square error**:

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

Confidence Intervals

For a **confidence level** of $\left(1-\alpha\right)$%.

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

Hypothesis Testing

Define a falsifiable **null hypothesis** H_0 and reject if the test statistic $T\left(\boldsymbol{x}\right)$ lies in the rejection region $R = \neg CI_{1-\alpha}$. R satisfies: $\Pr\left(T\left(\boldsymbol{x} \in R\right)\right) = \alpha$.

Small Sample Inference

When $n < 30$:

$$T\left(\boldsymbol{x}\right) \sim t_{\nu,\alpha/2}$$

where the degrees of freedom $\nu = n - 1$.

$$\mathrm{E}\left(X\right) = 0$$

$$\mathrm{Var}\left(X\right) = \frac{\nu}{\nu - 2}$$

Population Mean

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

$$T(\boldsymbol{x}) = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

for small samples use $t_{\nu,\alpha/2}$.

Population Proportion

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

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

$n\hat{p} > 5$ and $n(1-\hat{p}) > 5$.

Paired Differences

For two dependent samples,

$$\overline{d} = \overline{x}_1 - \overline{x}_2.$$

$$T(\boldsymbol{x}) = \frac{\overline{d} - d_0}{s_d/\sqrt{n}}$$

Difference of Population Means

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

for $n_1,\ n_2 \geq 30$.

$$T(\boldsymbol{x}) = \frac{(\overline{x}_1 - \overline{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

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

For small independent samples,

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

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

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

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

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

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

Difference of Population Proportions

$$\begin{aligned} CI_{1-\alpha} &= \hat{p}_1 - \hat{p}_2 \\ &\pm Z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}. \end{aligned}$$

$n_i\hat{p}_i > 5$, and $n_i(1-\hat{p}_i) > 5$ for $i = 1, 2$.
When the hypothesised difference is zero

$$p_0 = \frac{x_1 + x_2}{n_1 + n_2} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}.$$

so that $p_0 = p_1 = p_2$.

$$T(\boldsymbol{x}) = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{p_0(1-p_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

When the hypothesised difference is not 0, i.e., $p_1 - p_2 = \Delta_0$:

$$T(\boldsymbol{x}) = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

Significance of Results

The test can only be used to reject the null hypothesis. When the strength against the null hypothesis is weak, we cannot assume that the null hypothesis is true, rather, the test is inconclusive and that there is no statistical significance.

Power

$$\begin{aligned} \text{Power} &= 1 - \beta \\ &= 1 - \Pr\left(|T(\boldsymbol{x})| \leq Z_{\alpha/2} \,|\, \theta = \theta^*\right) \\ &= \Pr\left(|T(\boldsymbol{x})| \geq Z_{\alpha/2} \,|\, \theta = \theta^*\right). \end{aligned}$$

For the true value θ^* of θ (not θ_0).
As n increases, the β decreases, and the power increases. n can therefore be selected to achieve a desired true value of θ and a desired power.

P-Values

Rather than constructing R based on α, measure the strength of evidence against H_0 using

$$\alpha = \Pr(Z \geq T(\boldsymbol{x})).$$

The strength of evidence against H_0 increases as the p-value decreases.

ANOVA

Analyse effects of various factors that have more than two levels.

- **Experimental unit** — object whose response is measured, called the **dependent variable**.
- **Factor** — independent variable controlled/varied in experiment. The **levels** are the values that the factor can take.
- **Treatment** — specific combination of factor levels applied to an experimental unit.
- **Response** — variable measured for each experimental unit.

The F-statistic F_{test} ("F" in ANOVA table) is the ratio of the mean squares for the treatment and error sources of variation, or the ratio of the variation between treatments to the variation within treatments. Reject H_0 if MSXX $\gg$ MSE, (XX accounted for more of the total variance than the error). $F_{\text{test}} \gg F_{\nu_1,\,\nu_2,\,\alpha}$, $\Pr\left(F < F_{\nu_1,\,\nu_2,\,\alpha}\right) = 1 - \alpha$.

One-Way ANOVA

Let the outcome of an experiment for replication j of treatment i be denoted by y_{ij}.

$$y_{ij} = \mu_i + \epsilon_{ij}$$

for mean treatment effect μ_i and $n = IJ$ experimental units. Use the hypothesis:

$$H_0 : \mu_i = \mu_j \quad \text{for all} \quad i \neq j$$

$$H_A : \text{at least one } \mu_i \text{ is different}$$

Tukey's Honest Significant Difference Test

Tests the hypothesis that the mean of each treatment is equal to the mean of the other treatments.

$$H_0 : \mu_i = \mu_j \quad \text{for all} \quad i \neq j$$

where we can reject individual pairs of treatment means if the p-value is less than the significance level.

Two-way ANOVA with Blocking

When variation in the responses arise due to factors other than the treatment factors. Uses I treatments and $J > I$ blocks with I subjects to isolate block-to-block variability.

$$y_{ij} = \alpha_i + \beta_j + \epsilon_{ij}$$

for α_i and β_j the mean treatment effect and block effect, respectively.
The null hypotheses are:

$$H_0 : \alpha_i = \alpha_j \quad \text{for all} \quad i \neq j$$

$$H_A : \text{at least one } \alpha_i \text{ is different}$$

or

$$H_0 : \beta_i = \beta_j \quad \text{for all} \quad i \neq j$$

$$H_A : \text{at least one } \beta_i \text{ is different}$$

Two-Way ANOVA with Interaction

When both factors, and interactions between factors are of interest.

$$y_{ijk} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where α_i is the mean effect of the first factor, β_j is the mean effect of the second factor and $(\alpha\beta)_{ij}$ is the mean effect of the interaction between the two factors.
Factor A has I levels and factor B has J levels, and each of these factors is replicated K times. The null hypotheses are:

$$H_0 : \alpha_i = \alpha_j \quad \text{for all} \quad i \neq j$$

$$H_A : \text{at least one } \alpha_i \text{ is different}$$

or

$$H_0 : \beta_i = \beta_j \quad \text{for all} \quad i \neq j$$

$$H_A : \text{at least one } \beta_i \text{ is different}$$

or

$$H_0 : \text{No interaction between A and B}$$

$$H_A : \text{A and B interact.}$$

Linear Regression

Relationship between a response variable y and a predictor x:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i is the residual,

$$\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}\left(0,\, \sigma^2\right).$$

$$\therefore y_i \sim \mathrm{N}\left(\beta_0 + \beta_1 x_i,\, \sigma^2\right).$$

Coefficients

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \sim \mathrm{N}\left(\beta_0,\, s^2_{\hat{\beta}_0}\right)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \sim \mathrm{N}\left(\beta_1,\, s^2_{\hat{\beta}_1}\right)$$

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

$s^2 = \frac{\mathrm{SSE}}{n-2}$ from ANOVA.

$$s^2_{\hat{\beta}_0} = \frac{s^2 \overline{x}^2}{n \sum_{i=1}^n \left(x_i - \overline{x}\right)^2}$$

$$s^2_{\hat{\beta}_1} = \frac{s^2}{\sum_{i=1}^n \left(x_i - \overline{x}\right)^2}$$

Hypothesis Testing

Confidence intervals:

$$\hat{\beta}_0 \pm t_{\alpha/2, n-2} s_{\hat{\beta}_0}$$

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} s_{\hat{\beta}_1}$$

where

$$\Pr\left(t < t_{\alpha/2, n-2}\right) = 1 - \alpha/2.$$

The null hypothesis for β_1 tests whether there is indeed a linear relationship between the predictor and response variables:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

with test statistic

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \cancelto{0}{\beta_1}}{s_{\hat{\beta}_1}} \sim t_{n-2}$$

Assumptions

1. The parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased, i.e., the expected value of $\epsilon_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ is zero.
2. The residuals ϵ_i are independent, i.e., $\mathrm{Corr}\left(\epsilon_i,\, \epsilon_j\right) = 0$ for $i \neq j$.
3. The residuals follow a Gaussian distribution, i.e., $\epsilon_i \sim \mathrm{N}\left(0,\, \sigma^2\right)$.

Diagnostics to test assumptions:

1. Using a histogram of the residuals, we can check whether the residuals are unimodal and thus normally distributed.
2. Using a *q-q* plot, we can check whether the residuals lie on a straight line.
3. Using a plot of the residuals, we can check whether the residuals are independent, i.e., there are no patterns in the residuals and their variance is roughly equal or constant (they lie randomly around 0). This is known as homoscedasticity.

ANOVA on Linear Regression

How much variation in y is explained by the model. The null hypothesis is that *the model explains more variation in y than the sample mean $\overline{y}$*, with the alternative explaining *less*. For 1 independent variable, $F \equiv t^2$, and the null hypothesis is simply $H_0 : \beta_1 = 0$.

Coefficient of Determination R^2

Proportion of the total variation in y that is explained by the model.

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

R^2 is subjective: $R^2 \approx 1$ is good but may indicate "over-fitting" in the model, small values may also be acceptable.

Estimation and Prediction

Given the estimate $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$:

$$s_{\hat{y}_i} = \sqrt{s^2 \left(\frac{1}{n} + \frac{\left(x_i - \overline{x}\right)^2}{\sum_{i=1}^n \left(x_i - \overline{x}\right)^2}\right)}$$

the confidence interval for the true value of y_i is

$$\hat{y}_i \pm t_{n-2, 1-\alpha/2} s_{\hat{y}_i}$$

based on the sampling distribution

$$\frac{\hat{y}_i - \mathrm{E}\left(y_i\right)}{s_{\hat{y}_i}} \sim t_{n-2}.$$

A predicted value y^* for an unobserved x^* gives $y^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$, so that:

$$s_{y^*} = \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{\left(x^* - \overline{x}\right)^2}{\sum_{i=1}^n \left(x_i - \overline{x}\right)^2}\right)}$$

which gives the prediction interval,

$$y^* \pm t_{n-2, 1-\alpha/2} s_{y^*}.$$

We are most confident when x is near $\overline{x}$ (both bands are narrowest near $\overline{x}$), with the prediction interval being wider than the confidence interval. Caution should be used when extrapolating outside the data domain.

ANCOVA

Allows for the inclusion of continuous covariates that should be accounted for, but could not be controlled for. The ANOVA model is extended to

$$y_{ij} = \alpha_i + \epsilon_{ij} \to \alpha_i + \beta\left(x_{ij} - \overline{x}\right) + \epsilon_{ij}.$$

where the ANCOVA model accounts for covariate values x_{ij}.

These terms increase the power of the test in detecting treatment effects. All constraints of linear regression regarding the independence, homogeneity, and normality of residuals also apply to the covariate effects in the model.

This model also assumes that there is homogeneity in regression slopes, or β is approximately equal for all levels of treatment. This can be confirmed visually, or by fitting an interaction term between the treatment and covariates and testing if the interaction term is **not** significant ($p > 0.05$), indicating that it is the same for all treatments.

Categorical Data Analysis

Data are counts of members in each category, and we are interested in the relationships between categories.

2×2 Contingency Tables

2 factors with 2 categories, where cells are the counts of observations in each category. Counts are random variables and have sampling distributions, but because row/column sums are fixed, they are not independent.

Hypergeometric Distribution

Probability of drawing k successes from a population of size N without replacement, where the population contains K successes, and the draw size is n:

$$\Pr\left(X = k\right) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Using the example table below, the probability of drawing k_{11} successes from the population of size N without replacement, where the population contains a total of K_1 successes, and we draw a sample of size n_1 is given by

$$\Pr\left(X = k_{11}\right) = \frac{\binom{K_1}{k_{11}}\binom{N-K_1}{n_1-k_{11}}}{\binom{N}{n_1}}.$$

We can construct a null hypothesis that tests whether the two factors have an equal probability of being associated with a category, then we can use the hypergeometric distribution to calculate the probability of observing the data given the null hypothesis, and reject the null hypothesis if the probability (two-sided multiplied by 2) is less than 0.05.

Chi-Squared Distribution

$$X = \sum_{i=1}^n Z_i^2 \sim \chi^2_\nu$$

with $Z_i \sim \mathrm{N}\left(0,\, 1\right)$ and $\nu = n - 1$.

Test of Homogeneity

Tests whether the distribution of items across categories is the same for different factors.

Using the table below, calculate the expected counts for each cell in the table as

$$E_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}.$$

The null hypothesis is that the

distribution of items across categories is the same for all factors,

$H_0 : \pi_{ij} = \pi_i \forall i, j$

H_1 : at least one π_{ij} is different from π_i.

$$\pi_i \approx \hat{\pi}_i = \frac{n_{i.}}{n_{..}}.$$

Assume $\pi_{ij} \sim$ Poisson then the Z-score Z_{ij} is given by

$$Z_{ij} = \frac{\pi_{ij} - E_{ij}}{\sqrt{E_{ij}}} \sim \mathrm{N}\left(0,\, 1\right).$$

The χ^2 statistic is given by

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(\pi_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_\nu$$

for $\nu = (I-1)(J-1)$ degrees of freedom. We can therefore reject the null hypothesis if

$$X^2 > \chi^2_{\nu,\, 1-\alpha}.$$

This test assumes that row/column sums ($\pi_{i.}$ and $\pi_{.j}$) are fixed.

Test of Independence

Tests whether the distribution of items across categories is independent of the factors:

H_0 : The categories and factors are independent

H_A : at least one category is not independent of a factor.

We can reject the null hypothesis if

$$X^2 > \chi^2_{\nu,\, 1-\alpha}.$$

This test assumes only the total n is fixed.

Uniform Distribution

Single trial X in a set of equally likely elements.

Bernoulli (Binary) Distribution

Boolean-valued outcome X, i.e., success (1) or failure (0).

Binomial Distribution

Number of successes X for n independent trials.

Geometric Distribution

Number of trials N up to and including the first success.

Poisson Distribution

Number of events N which occur over a fixed interval of time: $\lambda = \eta t$.

Uniform Distribution

Outcome X within some interval, where the probability of an outcome in one interval is the same as all other intervals of the same length.

Exponential Distribution

Time T between events with rate η.

Memoryless Property

$$\Pr\left(T > s + t \,|\, T > t\right) = \Pr\left(T > s\right)$$

Equality	**Difference**	R
$\theta = \theta_0$	$\theta_1 - \theta_2 = 0$	$\lvert T(\boldsymbol{x}) \rvert > Z_{\alpha/2}$
$\theta \leq \theta_0$	$\theta_1 - \theta_2 \leq 0$	$T(\boldsymbol{x}) > Z_\alpha$
$\theta \geq \theta_0$	$\theta_1 - \theta_2 \geq 0$	$T(\boldsymbol{x}) < -Z_\alpha$

Table 1: Rejection regions for hypothesis testing.

$$\Pr\left(Z \geq Z_{\alpha/2}\right) = \frac{\alpha}{2},\ \Pr\left(Z \geq Z_\alpha\right) = \Pr\left(Z \leq -Z_\alpha\right) = \alpha$$

Decision	H_0 **true**	H_0 **false**
Reject	α (Type I error)	$1 - \beta$ (Power)
Fail	$1 - \alpha$	β (Type II error)

Table 2: Probability of rows given columns.

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	$np(1-p)$
$N \sim \mathrm{Geometric}(p)$	$n \geq 1$	$(1-p)^{n-1} p$	$1 - (1-p)^n$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$N \sim \mathrm{Poisson}(\lambda)$	$n \geq 0$	$\frac{\lambda^n e^{-\lambda}}{n!}$	$e^{-\lambda} \sum_{u=0}^{n} \frac{\lambda^u}{u!}$	λ	λ

Table 3: Discrete probability distributions.

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

Table 4: Continuous probability distributions.

	Factor 1	**Factor 2**	**Total Successes** (K)
Category 1	k_{11}	k_{12}	K_1
Category 2	k_{21}	k_{22}	K_2
Total Draws (n)	n_1	n_2	N

Table 5: Example 2×2 contingency table.

	Factor 1	$\cdots$	**Factor** J	**Total Successes** ($n_{i.}$)
Category 1	π_{11}	$\cdots$	π_{1J}	$n_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Category I	π_{I1}	$\cdots$	π_{IJ}	$n_{..}$
Total Draws $n_{.j}$	$n_{.1}$	$\cdots$	$n_{.J}$	$n_{..}$

Table 6: Example $I \times J$ contingency table.