$$\mathrm{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \mathrm{E}(S^2) = \sigma^2, \quad \mathrm{Var}(S^2) = \frac{\sigma^4}{n}\Big(\mathrm{E}(\frac{X-\mu}{\sigma})^4 - \frac{n-3}{n-1}\Big)$$

$$I_\mu \approx \bar{x} \pm z(\tfrac{\alpha}{2}) \cdot \frac{s}{\sqrt{n}}.$$

$$I_\mu = \bar{x} \pm t_{n-1}(\tfrac{\alpha}{2}) \cdot s_{\bar{x}}, \quad s_{s^2} = \sqrt{\frac{2}{n-1}} s^2.$$

	Negative decision: do not reject H_0	Positive decision: reject H_0 in favour of H_1
If H_0 is true	True negative outcome	False positive outcome, type I error
If H_1 is true	False negative outcome, type II error	True positive outcome

Planning the sample H_0: $p = p_0$ against H_1: $p = p_1$, $p_1 > p_0$!

$$\sqrt{n} \approx \frac{z(\alpha)\sqrt{p_0(1-p_0)} + z(\beta)\sqrt{p_1(1-p_1)}}{|p_1 - p_0|}$$

$$\frac{z(\alpha)\sqrt{p_0(1-p_0)} + \sqrt{n}(p_0 - p_1)}{\sqrt{p_1(1-p_1)}} \approx -z(\beta)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}, \; I_{\sigma^2} = \Big(\frac{(n-1)s^2}{x_{n-1}(\frac{\alpha}{2})}; \frac{(n-1)s^2}{x_{n-1}(1-\frac{\alpha}{2})}\Big).$$

4.2 Large-sample test for the proportion $\Phi(z(\alpha)) = 1 - \alpha$

normal approximation **Binomial test** $\mathcal{F}(\mu,\sigma) = \mathrm{Bin}(1,p)$ $z_0 = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \cdot Z_0 \overset{H_0}{\approx} \mathrm{N}(0,1)$.

for the maximum likelihood estimator Let $c = x_1 + \ldots + x_n$ **Power of the test** H_0: $p = p_0$ against H_1: $p > p_0$.

$$\hat{\Theta} \approx \mathrm{N}(\theta, \frac{\sigma_\theta}{\sqrt{n}})$$

$C \sim \mathrm{Bin}(n,p)$. define b_α and c_α by

$\mathrm{P}(C \le b_\alpha | H_0) = \alpha, \quad \mathrm{P}(C \ge c_\alpha | H_0) = \alpha.$

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}.$$

$$\mathrm{Pw}(p) = \mathrm{P}(Z_0 \ge z(\alpha)|H_1) = \mathrm{P}\Big(Z \ge \frac{z(\alpha)\sqrt{p_0(1-p_0)} + \sqrt{n}(p_0-p)}{\sqrt{p(1-p)}}|H_1\Big)$$

$$\approx 1 - \Phi\Big(\frac{z(\alpha)\sqrt{p_0(1-p_0)} + \sqrt{n}(p_0-p)}{\sqrt{p(1-p)}}\Big), \quad p > p_0.$$

Large-sample test for mean

H_1	Rejection region
one-sided $p > p_0$	$\mathcal{R} = \{c \ge c_\alpha\}$
one-sided $p < p_0$	$\mathcal{R} = \{c \le b_\alpha\}$
two-sided $p \ne p_0$	$\mathcal{R} = \{c \le b_{\alpha/2}\} \cup \{c \ge c_{\alpha/2}\}$

$$t_0 = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}. \quad T_0 \overset{H_0}{\approx} \mathrm{N}(0,1).$$

works well even if the population distribution $\mathcal{F}(\mu,\sigma)$ is not necessarily normal.

Summarising data

6.1 Empirical probability distribution

empirical variance.

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \le x\}}.$$

$$\hat{\sigma}^2 = \tfrac{n-1}{n}s^2 = \tfrac{1}{n}\sum(x_i - \bar{x})^2 = \overline{x^2} - (\bar{x})^2$$

Empirical survival function If L has the exponential distribution $\mathrm{Exp}(\lambda)$, then the hazard rate $h(x) \equiv \lambda$ is constant over the ages.

$$\hat{S}(x) = 1 - \hat{F}(x), \quad f(x) = F'(x)$$

hazard $h(x)$ $h(x) = \frac{f(x)}{S(x)}$. $h(x) = \frac{f(x)}{S(x)} = -\frac{S'(x)}{S(x)} = -\frac{d}{dx}\ln S(x)$.

$$\frac{\mathrm{P}(x < L \le x + h | L > x)}{} = \frac{\mathrm{P}(x < L \le x + h)}{} = \frac{F(x+\delta) - F(x)}{} \to h(x), \quad \delta \to 0.$$

hazard function can be viewed as the negative slope of the log survival function:

4.5 Chi-squared test of goodness of fit

$(C_1, \ldots, C_J) \sim \mathrm{Mn}(n; p_1, \ldots, p_J)$, $\mathrm{P}(C_1 = c_1, \ldots, C_J = c_J) = \frac{n!}{k_1! \cdots k_J!} p_1^{c_1} \cdots p_J^{c_J}$

$\Omega = \{(p_1, \ldots, p_J) : p_1 + \ldots + p_J = 1, p_1 \ge 0, \ldots, p_J \ge 0\}$

$\dim(\Omega) = J - 1$. $H_0 : (p_1, \ldots, p_J) \in \Omega_0 \dim(\Omega_0) = r, \quad 0 \le r < J - 1.$

$\Omega_0 = \{(p_1, \ldots, p_J) \in \Omega : (p_1, \ldots, p_J) = (p_1(\lambda), \ldots, p_J(\lambda)), \quad \lambda \in \Lambda\},$

$e_j = n \cdot p_j(\hat{\lambda})$, $-2\ln w \approx \mathrm{x}^2$ $\mathrm{x}^2 = \sum_{j=1}^{J} \frac{(c_j - e_j)^2}{e_j}$, $\mathrm{X}^2 \overset{H_0}{\approx} \chi^2_{J-1-r}$,

chi-squared test is approximate in that all *expected* counts are recommended to be at least 5.

$\mathrm{df} = \dim(\Omega) - \dim(\Omega_0) = (J-1) - r = J - 1 - r.$

6.3 Density estimation

histogram $c_j = \sum_{i=1}^{n} 1_{\{x_i \in \mathrm{cell}_j\}}$

kernel density estimate $f_h(x) = \frac{c_j}{nh}$, for $x \in \mathrm{cell}_j$,

$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} \phi(\frac{x - x_i}{h})$, where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. $\int f_h(x)dx = \frac{h}{nh}\sum_j c_j = 1.$

6.4 Skewness and kurtosis

In terms of its normalized version $Z = \frac{X-\mu}{\sigma}$,

Parametric distribution	Skewness	Kurtosis
$\mathrm{N}(\mu,\sigma)$ normal distribution	$\mu_3 = 0$ symmetric	$\mu_4 = 3$
$\mathrm{Gam}(\alpha,\lambda)$ gamma distribution	$\mu_3 = \frac{2}{\sqrt{\alpha}}$ skewed to the right	$\mu_4 = 3 + \frac{6}{\alpha}$ leptokurtic
$\mathrm{Beta}(a,a)$ beta distribution	$\mu_3 = 0$ symmetric	$\mu_4 = 3 - \frac{1}{2a+3}$ platykurtic
t-distribution with $\mathrm{df} = k$ and $k \ge 5$	$\mu_3 = 0$ symmetric	$\mu_4 = 3 + \frac{6}{k-4}$ leptokurtic

$$\mu_1 = \mathrm{E}(Z) = 0, \quad \mu_2 = \mathrm{E}(Z^2) = 1$$

$$\mu_3 = \mathrm{E}(Z^3), \quad \mu_4 = \mathrm{E}(Z^4).$$

$$m_3 = \frac{1}{s^3 n}\sum_{i=1}^{n}(x_i - \bar{x})^3, \quad m_4 = \frac{1}{s^4 n}\sum_{i=1}^{n}(x_i - \bar{x})^4.$$

6.5 Inference on the population median

$$\hat{m} = x_{(k)} \quad \text{or} \quad \hat{m} = \frac{x_{(k)} + x_{(k+1)}}{2},$$

Confidence interval for the median

Suppose that the population distribution $\mathcal{F}(\mu,\sigma)$ is continuous.

Observe that the number of observations below the true population median m,

$$y = \sum_{i=1}^{n} 1_{\{x_i \le m\}}$$

$$p_k = \mathrm{P}(Y < k) = \sum_{i=0}^{k-1}\binom{n}{i}2^{-n}, \quad k = 1, \ldots, n.$$

the following non-parametric formula for an exact confidence interval for the median.

$I_m = (x_{(k)}, x_{(n-k+1)})$ is a $100 \cdot (1 - 2p_k)\%$ confidence interval for the population median m

Sign test

$$y_0 = \sum_{i=1}^{n} 1_{\{x_i \le m_0\}} \; Y_0 \overset{H_0}{\approx} \mathrm{Bin}(n, 0.5).$$

non-parametric test of H_0: $m = m_0$ against the two-sided alternative H_1: $m \ne m_0$.

$$I_m = (x_{(k)}, x_{(n-k+1)}),$$

6.6 Measures of dispersion

sample range $x_{(n)} - x_{(1)}$.

the interquartile range, $\mathrm{IQR} = x_{0.75} - x_{0.25}$,

the median of absolute values of deviations, MAD,

is defined as the sample median of $\{|x_i - \hat{m}|, i = 1, \ldots, n\}$

general normal distribution $\mathrm{N}(\mu,\sigma)$, **estimates of σ:**

theoretical lower and upper quartiles are $\mu \pm 0.675 \cdot \sigma$ s, $\frac{\mathrm{IQR}}{1.35}$, $\frac{\mathrm{MAD}}{0.675}$.

$\mathrm{IQR} = (\mu + 0.675 \cdot \sigma) - (\mu - 0.675 \cdot \sigma) = 1.35 \cdot \sigma.$

we obtain $\mathrm{MAD} = 0.675 \cdot \sigma.$

$\mathrm{P}(|X - \mu| \le 0.675 \cdot \sigma) = (\Phi(0.675) - 0.5) \cdot 2 = 0.5,$

Box

- upper edge of the box = upper quartile (UQ)
- box center = median
- lower edge of the box = lower quartile (LQ)
- box height = interquartile range (IQR)

Wiskers

- upper whisker end = {largest data point ≤ UQ + 1.5 × IQR}
- lower whisker end = {smallest data point ≥ LQ − 1.5 × IQR}

Outliers

- upper dots = {data points > UQ + 1.5 × IQR}
- lower dots = {data points < LQ − 1.5 × IQR}

7.1 Two independent samples

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m},$$

$$s^2_{\bar{x}-\bar{y}} = s^2_{\bar{x}} + s^2_{\bar{y}} = \frac{s_1^2}{n} + \frac{s_2^2}{m}$$

unbiased estimate of $\mathrm{Var}(\bar{X} - \bar{Y})$.

Two-sample t-test three assumptions

1. two random samples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$ are independent of each other,

2. both underlying population distributions are normal,

$\mathcal{F}(\mu_1,\sigma_1) = \mathrm{N}(\mu_1,\sigma_1)$ and $\mathcal{F}(\mu_2,\sigma_2) = \mathrm{N}(\mu_2,\sigma_2)$

3. these two normal distributions have equal variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

pooled sample variance. unbiased estimate of σ^2:

$$s_{\mathrm{p}}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2}{n + m - 2},$$

In the case of equal variances and independent samples, the variance

$\mathrm{Var}(\bar{X} - \bar{Y}) = \sigma^2 \cdot \frac{n+m}{nm}$, unbiased estimate $s^2_{\bar{x}-\bar{y}} = s_{\mathrm{p}}^2 \cdot \frac{n+m}{nm}$.

normality assumption with equal variances, the sampling distribution of the two-sample t-score

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{s_{\mathrm{p}}} \cdot \sqrt{\frac{nm}{n+m}}$$ t-distribution with $(n + m - 2)$ degrees of freedom.

Exact confidence interval $I_{\mu_1-\mu_2} = \bar{x} - \bar{y} \pm t_{n+m-2}(\frac{\alpha}{2}) \cdot s_{\mathrm{p}} \cdot \sqrt{\frac{n+m}{nm}}$

two sample t-test for testing H_0: $\mu_1 = \mu_2$ $t_0 = \frac{\bar{x} - \bar{y}}{s_{\mathrm{p}}} \cdot \sqrt{\frac{nm}{n+m}}$

$T_0 \overset{H_0}{\sim} t_{n+m-2}.$

Large-sample test for the difference between two means

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_X^2 + S_Y^2}} \approx \mathrm{N}(0,1). \quad H_0 : \mu_1 = \mu_2$$

Approximate $100(1-\alpha)\%$ confidence interval $I_{\mu_1-\mu_2} \approx \bar{x} - \bar{y} \pm z(\frac{\alpha}{2}) \cdot \sqrt{s_x^2 + s_y^2}$.

Rank sum test

$$z_0 = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2 + s_y^2}}$$

when either one of the distributions F_1 and F_2 is not normal, or their variances are not equal

pool the samples and replace the data values by their ranks $1, 2, \ldots, n+m$, starting from the smallest sample value to the largest, and then compute two test statistics r_1 = sum of the ranks of x-observations, and r_2 = sum of y-ranks.

$r_1 + r_2 = 1 + 2 + \ldots + (n+m) = \frac{(n+m)(n+m+1)}{2}.$

H_0: $F_1 = F_2$ against H_1: $F_1 \ne F_2$ $\mathrm{E}(R_1) = \frac{n(n+m+1)}{2}$,

For $n \ge 10$, $m \ge 10$, we may apply the normal approximation

exact null distributions for R_1 and R_2 depend only on the sample sizes n and m,

$\mathrm{E}(R_2) = \frac{m(n+m+1)}{2}$ $\mathrm{Var}(R_1) = \mathrm{Var}(R_2) = \frac{mn(n+m+1)}{12}.$

7.2 Two independent samples: comparing population proportions

$X \sim \mathrm{Bin}(1, p_1)$ and $Y \sim \mathrm{Bin}(1, p_2)$. $\bar{x} = \hat{p}_1$, $\bar{y} = \hat{p}_2$,

$$s_{\hat{p}_1} = \sqrt{\tfrac{\hat{p}_1(1-\hat{p}_1)}{n-1}}, \quad s_{\hat{p}_2} = \sqrt{\tfrac{\hat{p}_2(1-\hat{p}_2)}{m-1}} \cdot H_1 : p_1 \ne p_2$$

Large sample test for two proportions

$$I_{p_1-p_2} \approx \hat{p}_1 - \hat{p}_2 \pm z(\tfrac{\alpha}{2}) \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n-1} + \frac{\hat{p}_2(1-\hat{p}_2)}{m-1}}.$$

Fisher's exact test $H_0 : p_1 = p_2$,

	Sample 1	Sample 2	Total
Number of successes	c_{11}	c_{12}	$c_{11} + c_{12}$
Number of failures	c_{01}	c_{02}	$c_{01} + c_{02}$
Sample sizes	n	m	$N = n + m$

$c_{11} = x_1 + \ldots + x_n, \quad c_{01} = n - c_{11}, \; p = \frac{c_{11} + c_{12}}{N},$

$c_{12} = y_1 + \ldots + y_m, \quad c_{02} = m - c_{12}.$

$C_{11} \overset{H_0}{\sim} \mathrm{Hg}(N, n, p).$

7.3 Paired samples

$\mathrm{E}(X) = \mu_1, \quad \mathrm{E}(Y) = \mu_2, \quad \mathrm{Var}(X) = \sigma_1^2, \quad \mathrm{Var}(Y) = \sigma_2^2,$

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_1\sigma_2} > 0.$$

$\mathrm{Var}(\bar{X} - \bar{Y}) = \frac{1}{n}(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho).$

if the two samples are independent and have equal sizes, then $\rho = 0$ and

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \tfrac{1}{n}(\sigma_1^2 + \sigma_2^2).$$

It follows that with $\rho > 0$, the paired sampling ensures a smaller standard error for the estimate $\bar{x} - \bar{y}$ compared to the independent X and Y.

$(d_1, \ldots, d_n)$, $d_i = x_i - y_i$. $\bar{d} = \bar{x} - \bar{y}$,

$\mathrm{E}(\bar{D}) = \mu_1 - \mu_2, \quad \mathrm{E}(\bar{D}) = \frac{1}{n}(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho).$

Signed rank test H_0: $m = 0$ against H_1: $m \ne 0$.

assumption: the population distribution of $D = X - Y$ is symmetric about its median m.

$r_i = \mathrm{rank}(|d_i|), \quad i = 1, \ldots, n.$

signed rank test statistic w: either as the sum of positive ranks

$$w = \sum_{i=1}^{n} r_i \cdot 1_{\{d_i > 0\}}$$ or the sum of the negative ranks

For $n \ge 20$, one can use the normal approximation $\mathrm{N}(\mu,\sigma)$

$$w = \sum_{i=1}^{n} r_i \cdot 1_{\{d_i < 0\}}. \quad \mu = \frac{n(n+1)}{4}, \quad \sigma^2 = \frac{n(n+1)(2n+1)}{24}.$$

The signed rank test uses more data information than the sign test but requires symmetric distribution of differences.

7.4 Paired samples: comparing population proportions

two dependent Bernoulli random variables $X \sim \mathrm{Bin}(1, p_1)$ and $Y \sim \mathrm{Bin}(1, p_2)$.

$(0,0), (0,1), (1,0), (1,1)$ with probabilities denoted as $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$.

$(C_{00}, C_{01}, C_{10}, C_{11}) \sim \mathrm{Mn}(n, \pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}).p_1 - p_2 = \pi_{10} - \pi_{01},$

$\hat{p}_1 - \hat{p}_2 = \hat{\pi}_{10} - \hat{\pi}_{01}, \quad \hat{\pi}_{10} = \frac{c_{10}}{n}, \quad \hat{\pi}_{01} = \frac{c_{01}}{n}.$

$$s^2_{\hat{p}_1-\hat{p}_2} = \frac{\hat{\pi}_{10} + \hat{\pi}_{01} - (\hat{\pi}_{10} - \hat{\pi}_{01})^2}{n-1},$$

$$I_{p_1-p_2} \approx \hat{p}_1 - \hat{p}_2 \pm z(\tfrac{\alpha}{2}) \cdot s_{\hat{p}_1-\hat{p}_2}.$$

8.1 One-way layout ANOVA

$H_0 : \mu_1 = \ldots = \mu_I$, against $H_1 : \mu_i \ne \mu_j$ for some pair of indices (i,j).

independently collected a random sample $(y_{i1}, \ldots, y_{in})$ of size n. **Normal theory model**

$Y_{ik} = \mu + \alpha_i + \sigma Z_{ik}, \quad Z_{ik} \sim \mathrm{N}(0,1), \quad i = 1, \ldots, I, \quad k = 1, \ldots, n,$

$N = I \cdot n$ $\hat{\mu} = \bar{y}_{..}, \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..},$

$$\bar{y}_{i.} = \frac{1}{n}\sum_{k=1}^{n} y_{ik},$$

$$\bar{y}_{..} = \frac{\bar{y}_{1.} + \ldots + \bar{y}_{I.}}{I} = \frac{1}{N}\sum_{i=1}^{I}\sum_{k=1}^{n} y_{ik} \cdot \quad y_{ik} = \hat{\mu} + \hat{\alpha}_i + \hat{e}_{ik}, \quad \hat{e}_{ik} = y_{ik} - \bar{y}_{i.},$$

$\mathrm{ss_T} = \mathrm{ss_A} + \mathrm{ss_E},$

$\mathrm{ss_T} = \sum_i \sum_k (y_{ik} - \bar{y}_{..})^2$ is the total sum of squares for the pooled sample,

$\mathrm{ss_A} = n \sum_i \hat{\alpha}_i^2$ is the factor A sum of squares $\mathrm{df_T} = N - 1, \quad \mathrm{df_A} = I - 1, \quad \mathrm{df_E} = I \cdot (n-1).$

$\mathrm{ss_E} = \sum_i \sum_k \hat{e}_{ik}^2$ is the error sum of squares.

$$\mathrm{ms_A} = \frac{\mathrm{ss_A}}{\mathrm{df_A}}, \quad \mathrm{ms_E} = \frac{\mathrm{ss_E}}{\mathrm{df_E}}.$$

One-way F-test

Under the null hypothesis $H_0 : \alpha_1 = \ldots = \alpha_I = 0$, we expect

$$\mathrm{E}(\mathrm{MS_A}) = \mathrm{E}(\mathrm{MS_E}) = \sigma^2.$$

if H_0 is violated, then $\mathrm{E}(MS_A) > \mathrm{E}(MS_E)$, as given some $\alpha_i \ne 0$, we have

$$\sigma^2 + \frac{n}{I-1}\sum \alpha_i^2 > \sigma^2. \quad f = \frac{\mathrm{ms_A}}{\mathrm{ms_E}}$$

$F \overset{H_0}{\sim} F_{k_1,k_2}, \qquad k_1 = I - 1, \; k_2 = I(n-1),$

$$\frac{X_1/k_1}{X_2/k_2} \sim F_{k_1,k_2},$$

$$s_{\mathrm{p}}^2 = \mathrm{ms_E} = \frac{\sum_{i=1}^{I}\sum_{k=1}^{n}(y_{ik} - \bar{y}_{i.})^2}{I(n-1)}. \quad \mathrm{E}(\mathrm{MS_A}) = \sigma^2 + \frac{n}{I-1}\sum_i \alpha_i^2, \qquad \mathrm{E}(\mathrm{MS_E}) = \sigma^2.$$

$X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$ are two independent random variables having chi-squared distributions

One-sample t-test $\mathcal{F}(\mu,\sigma) = \mathrm{N}(\mu,\sigma)$,

H_1	Rejection region
one-sided $p > p_0$	$\mathcal{R} = \{z \ge z(\alpha)\}$
one-sided $p < p_0$	$\mathcal{R} = \{z \le -z(\alpha)\}$
two-sided $p \ne p_0$	$\mathcal{R} = \{z \le -z(\frac{\alpha}{2})\} \cup \{z \ge z(\frac{\alpha}{2})\}$

t-test of H_0: $\mu = \mu_0$

Likelihood-ratio test $T_0 \overset{H_0}{\sim} t_{n-1}$.

rejects H_0 for small values of the likelihood-ratio.

$H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. $\frac{L(\theta_0)}{L(\theta_1)}$

optimal test in the case of two simple

Nested hypotheses

$H_0 : \theta \in \Omega_0, \quad H : \theta \in \Omega, w = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})},$

rejects H_0 for smaller values of w.

equivalently for larger values of

$-\ln w = \ln L(\theta) - \ln L(\theta_0)$ sufficiently large sample sizes n.

$-2\ln W \overset{H_0}{\approx} \chi^2_{\mathrm{df}}$, where $\mathrm{df} = \dim(\Omega) - \dim(\Omega_0)$.

Quantiles and QQ-plots

inverse of the cumulative distribution function $F(x)$ is called the quantile function

$Q(p) = F^{-1}(p), \quad 0 < p < 1.$

$x_{(k)}$ is called the empirical $(\frac{k-0.5}{n})$-quantile

$$\hat{F}(x_{(k)}) = \frac{k}{n}, \qquad \hat{F}(x_{(k)} - \epsilon) = \frac{k-1}{n}.$$

8.2 Simultaneous confidence interval

Bonferroni method

$\alpha = \mathrm{P}(X_c \geq 1|H_0), X_c \overset{H_0}{\sim} \mathrm{Bin}(c, \alpha_c),$

where X_c is the number of positive results among these c tests.

$\mathrm{P}(X \geq 1|H_0) = 1 - (1-\alpha_c)^c.$

$\alpha = 1 - (1-\alpha_c)^c \approx c\alpha_c.$

each individual test should be performed at the level α/c.

current setting of $c = \binom{I}{2}$ pairwise differences $(\mu_i - \mu_j)$,

$100(1-\alpha)\%$ simultaneous confidence interval:

$$B_{\mu_i-\mu_j} = \bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm t_{I(n-1)}\left(\frac{\alpha}{I(I-1)}\right) \cdot s_{\mathrm{p}}\sqrt{\frac{2}{n}},$$

e confidence interval for a single difference

$$I_{\mu_i-\mu_j} = \bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm t_{I(n-1)}\left(\frac{\alpha}{2}\right) \cdot s_{\mathrm{p}}\sqrt{\frac{2}{n}}, \quad 1 \leq i, j \leq I$$

$$\alpha \to \frac{\alpha}{c} = \frac{2\alpha}{I(I-1)}.$$

pairwise differences $\mu_i - \mu_j$ are not independent as required by Bonferroni i

Bonferroni method gives slightly wider intervals compared to the Tukey

Tukey method

normality assumption with equal variances,

$X_i = \bar{Y}_{i\cdot} - \mu_i \sim \mathrm{N}(0, \frac{\sigma}{\sqrt{n}}), \quad i = 1, \ldots, I,$

are independent normal variables.

$\frac{\max\{X_1, \ldots, X_I\} - \min\{X_1, \ldots, X_I\}}{S_{\mathrm{p}}/\sqrt{n}} \sim \mathrm{SR}_{k_1,k_2}, \quad k_1 = I,\ k_2 = I(n-1).$

studentised range distribution SR_{k_1,k_2}

number of independent samples k_1

degrees of freedom k_2 used in the variance estimate s_{p}^2.

Tukey's $100(1-\alpha)\%$ simultaneous confidence interval

$$T_{\mu_i-\mu_j} = \bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm q_{I,I(n-1)}(\alpha) \cdot \frac{s_{\mathrm{p}}}{\sqrt{n}}$$

Tukey takes into account the dependences

8.3 Kruskal-Wallis test

non-parametric I independent random samples

H_0 : the underlying I population distributions are equal.

pooled sample of size $N = I \cdot n$. $\sum_i \sum_k r_{ik} = 1 + 2 + \ldots + N = \frac{N(N+1)}{2}$

overall mean rank is $\bar{r}_{\cdot\cdot} = \frac{N(N+1)}{2N} = \frac{N+1}{2}$.

Let r_{ik} be the pooled ranks measures the discrepancy between the sample means of the ranks

$$w = \frac{12n}{N(N+1)} \sum_{i=1}^{I} (\bar{r}_{i\cdot} - \tfrac{N+1}{2})^2 \quad \bar{r}_{i\cdot} = \frac{r_{i1} + \ldots + r_{in}}{n}, \quad i = 1, \ldots, I.$$

A large value of w would indicate a deviation from the null distribution.

For $I = 3$, $n \geq 5$ or $I \geq 4$, $n \geq 4$, one can use the approximate null distribution

$W \overset{H_0}{\approx} \chi^2_{I-1}$.

Categorical data analysis

$(C_1, \ldots, C_k) \sim \mathrm{Mn}(n, \pi_1, \ldots, \pi_k).$

	b_1	b_2	...	b_J	total
a_1	π_{11}	π_{12}	...	π_{1J}	$\pi_{1\cdot}$
a_2	π_{21}	π_{22}	...	π_{2J}	$\pi_{2\cdot}$
...	...	...	...	...	...
a_I	π_{I1}	π_{I2}	...	π_{IJ}	$\pi_{I\cdot}$
total	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	...	$\pi_{\cdot J}$	1

$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ for all pairs (i, j).

$\pi_{i|j} = \mathrm{P}(A = a_i | B = b_j) = \frac{\pi_{ij}}{\pi_{\cdot j}}$

	b_1	b_2	...	b_J
a_1	$\pi_{1\|1}$	$\pi_{1\|2}$	...	$\pi_{1\|J}$
a_2	$\pi_{2\|1}$	$\pi_{2\|2}$	...	$\pi_{2\|J}$
...	...	...	...	...
a_I	$\pi_{I\|1}$	$\pi_{I\|2}$	...	$\pi_{I\|J}$
total	1	1	...	1

$H_0 : \pi_{i|j} = \pi_{i\cdot}$ for all pairs (i, j).

null hypothesis of homogeneity states the equality of J population distributions

9.1 Chi-squared test of homogeneity

	sample 1	sample 2	...	sample J	total counts
category a_1	c_{11}	c_{12}	...	c_{1J}	c_1
category a_2	c_{21}	c_{22}	...	c_{2J}	c_2
...	...	...	...	...	...
category a_I	c_{I1}	c_{I2}	...	c_{IJ}	c_I
sample sizes	n_1	n_2	...	n_J	n

J multinomial distributions

$(C_{1j}, \ldots, C_{Ij}) \sim \mathrm{Mn}(n_j; \pi_{1|j}, \ldots, \pi_{I|j}), \quad j = 1, \ldots, J.$

degrees of freedom for J independent samples from I-dimensional multinomial

s $J(I-1)$. $H_0 : \pi_{i|j} = \pi_i$ for all (i, j),

maximum likelihood estimates of π_i a

$\hat{\pi}_i = c_i/n, \quad i = 1, \ldots, I.$ estimates consumes $(I-1)$ degrees of freedom,

expected cell counts $e_{ij} = n_j \cdot \hat{\pi}_i = c_i n_j / n$

$$\mathrm{x}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(c_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(c_{ij} - c_i n_j/n)^2}{c_i n_j / n}.$$

reject H_0 for larger values of x^2 $\mathrm{X}^2 \approx \chi^2_{\mathrm{df}}$

df $= J(I-1) - (I-1) = (I-1)(J-1)$.

9.2 Chi-squared test of independence

contrast to the previous setting of J independent samples,

s, the total counts $(n_1, n_2, \ldots, n_J)$ are random outcomes

$(C_{11}, \ldots, C_{IJ}) \sim \mathrm{Mn}(n_{\cdot\cdot}; \pi_{11}, \ldots, \pi_{IJ})$,

characterised by $IJ - 1$ degrees of freedom.

	b_1	b_2	...	b_J	total
a_1	c_{11}	c_{12}	...	c_{1J}	c_1
a_2	c_{21}	c_{22}	...	c_{2J}	c_2
...	...	...	...	...	...
a_I	c_{I1}	c_{I2}	...	c_{IJ}	c_I
total	n_1	n_2	...	n_J	n

$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ for all pairs (i, j).

maximum likelihood estimates of $\pi_{i\cdot}$ and $\pi_{\cdot j}$

$\hat{\pi}_{i\cdot} = \frac{c_i}{n}, \quad \hat{\pi}_{\cdot j} = \frac{n_j}{n}.$

null hypothesis of independence the expected cell counts

$e_{ij} = n\hat{\pi}_{ij} = n\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j} = \frac{c_i n_j}{n}$ $\mathrm{df} = (IJ-1) - (I-1+J-1) = (I-1)(J-1)$

The chi-squared tests of homogeneity and independence have the same test rejection rule.

9.3 Matched-pairs designs

9.4 Odds ratios

$\mathrm{odds}(A) = \frac{\mathrm{P}(A)}{\mathrm{P}(A^c)} = \frac{\mathrm{P}(A)}{1-\mathrm{P}(A)}$ $\mathrm{P}(A) = \frac{\mathrm{odds}(A)}{1+\mathrm{odds}(A)}$, $\mathrm{odds}(A|B) = \frac{\mathrm{P}(A|B)}{\mathrm{P}(A^c|B)} = \frac{\mathrm{P}(AB)}{\mathrm{P}(A^cB)}$

odds ratio for a pair of events (A, B) $\Delta_{AB} = \frac{\mathrm{odds}(A|B)}{\mathrm{odds}(A|B^c)}$,

$\Delta_{AB} = \frac{\mathrm{P}(AB)\mathrm{P}(A^cB^c)}{\mathrm{P}(A^cB)\mathrm{P}(AB^c)}$, $\Delta_{AB} = \Delta_{BA}$, $\Delta_{AB^c} = \frac{1}{\Delta_{AB}}$.

if $\Delta_{AB} = 1$, then events A and B are independent,

if $\Delta_{AB} > 1$, then $\mathrm{P}(A|B) > \mathrm{P}(A|B^c)$ so that B increases the probability of A,

if $\Delta_{AB} < 1$, then $\mathrm{P}(A|B) < \mathrm{P}(A|B^c)$ so that B decreases the probability of A.

10.1 Simple linear regression model

$Y = \beta_0 + \beta_1 x + \sigma Z, \qquad Z \sim \mathrm{N}(0,1),$

$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n,$ where $C = (2\pi)^{-n/2}$.

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\Big\} = C\sigma^{-n} \exp\Big\{-\frac{1}{2\sigma^2}\sum e_i^2\Big\},$$

$l(\theta) = \ln C - (n/2)\ln\sigma^2 - \frac{\sum_{i=1}^n e_i^2}{2\sigma^2}$ $\bar{x} = \frac{x_1 + \ldots + x_n}{n}, \quad \bar{y} = \frac{y_1 + \ldots + y_n}{n},$

$e_i = y_i - \beta_0 - \beta_1 x_i, \quad \overline{x^2} = \frac{x_1^2 + \ldots + x_n^2}{n}, \quad \overline{y^2} = \frac{y_1^2 + \ldots + y_n^2}{n},$

$n^{-1}\sum_{i=1}^{n} e_i^2 = \beta_0^2 + 2\beta_0\beta_1\bar{x} - 2\beta_0\bar{y} - 2\beta_1\overline{xy} + \beta_1^2\overline{x^2} + \overline{y^2},$

$\overline{xy} = \frac{x_1y_1 + \ldots + x_ny_n}{n}$ $\frac{\partial l}{\partial \beta_0} = -\frac{n}{\sigma^2}(\beta_0 + \beta_1\bar{x} - \bar{y}),$

$\frac{\partial l}{\partial \beta_1} = -\frac{n}{\sigma^2}(\beta_0\bar{x} - \overline{xy} + \beta_1\overline{x^2}),$

$\hat{e}_i = y_i - b_0 - b_1x_i.$ $\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} e_i^2,$

$\mathrm{ss_E} = \hat{e}_1^2 + \ldots + \hat{e}_n^2,$

$b_0 + b_1\bar{x} = \bar{y}, \quad b_0\bar{x} + b_1\overline{x^2} = \overline{xy}, \quad \hat{\sigma}^2 = \frac{\mathrm{ss_E}}{n},$

$b_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}, \qquad b_0 = \bar{y} - b_1\bar{x}.$

$\mathrm{ss_E} = \min_{\beta_0,\beta_1}\Big\{\sum_{i=1}(y_i - \beta_0 - \beta_1 x_i)^2\Big\},$

The maximum likelihood estimate of $\hat{\sigma}^2 = \frac{\mathrm{ss_E}}{n}$ is a

biased but asymptotically unbiased estimate of σ^2.

unbiased estimate of σ^2 $s^2 = \frac{\mathrm{ss_E}}{n-2}$.

10.2 Confidence intervals and hypothesis testing

$B_0 \sim \mathrm{N}(\beta_0, \sigma_0), \quad \sigma_0^2 = \frac{\sigma^2\sum x_i^2}{n(n-1)s_x^2}, \quad s_{b_0}^2 = \frac{s^2\sum x_i^2}{n(n-1)s_x^2}, \quad \frac{B_0 - \beta_0}{S_{B_0}} \sim t_{n-2},$

$B_1 \sim \mathrm{N}(\beta_1, \sigma_1), \quad \sigma_1^2 = \frac{\sigma^2}{(n-1)s_x^2}, \quad s_{b_1}^2 = \frac{s^2}{(n-1)s_x^2}, \quad \frac{B_1 - \beta_1}{S_{B_1}} \sim t_{n-2}.$

$\mathrm{Cov}(B_0, B_1) = -\frac{\sigma^2\bar{x}}{(n-1)s_x^2}$ which is negative, if $\bar{x} > 0$, and positive, if $\bar{x} < 0$.

H_0: $\beta_i = \beta^*$.

For β_0 and β_1, we get two exact $100(1-\alpha)\%$

% confidence intervals $I_{\beta_i} = b_i \pm t_{n-2}\left(\frac{\alpha}{2}\right) \cdot s_{b_i}$

$t = \frac{b_i - \beta^*}{s_{b_i}}$,

$T \sim t_{n-2}$.

Residuals

$\hat{y}_i = b_0 + b_1x_i,$

$\hat{e}_i = y_i - \hat{y}_i. \quad \hat{e}_1 + \ldots + \hat{e}_n = 0,$

$x_1\hat{e}_1 + \ldots + x_n\hat{e}_n = 0,$ $\mathrm{Cov}(\hat{E}_i, \hat{E}_j) = -\sigma^2\frac{c_{ij}}{n-1},$

$\hat{y}_1\hat{e}_1 + \ldots + \hat{y}_n\hat{e}_n = 0.$

$\mathrm{Var}(\hat{E}_i) = \sigma^2\left(1 - \frac{c_{ii}}{n-1}\right), \quad c_{ij} = \frac{\sum_{k=1}^n (x_i - x_k)(x_j - x_k)}{ns_x^2}.$

Sample correlation coefficient

$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$

$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$ sample correlation coefficient

$y = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x}), \quad r = \frac{s_{xy}}{s_x s_y}. \quad b_1 = \frac{rs_y}{s_x},$

$\left(\frac{y - \bar{y}}{s_y}\right) = r\left(\frac{x - \bar{x}}{s_x}\right).$

Intervals for individual observations

$y_p = \beta_0 + \beta_1 x_p + \sigma z_p, \quad z_p$ is generated by the $\mathrm{N}(0,1)$

$\mu_p = \beta_0 + \beta_1 x_p$

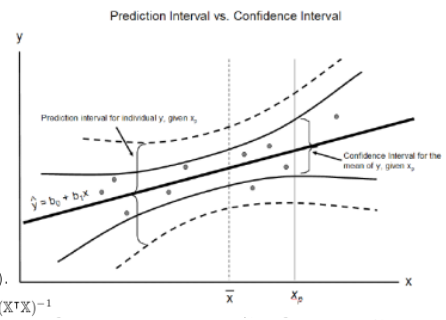
$\hat{\mu}_p = b_0 + b_1 x_p.$ prediction interval of the response value y_p

$\mathrm{Var}(B_0 + B_1 x_p) = \mathrm{Var}(B_0) + x_p^2\mathrm{Var}(B_1) + 2x_p\mathrm{Cov}(B_0, B_1) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n-1} \cdot (\frac{x_p - \bar{x}}{s_x})^2.$

$$I_{\mu_p} = b_0 + b_1 x_p \pm t_{n-2}\left(\frac{\alpha}{2}\right) \cdot s\sqrt{\frac{1}{n} + \frac{1}{n-1}\left(\frac{x_p - \bar{x}}{s_x}\right)^2}.$$

$\mathrm{Var}(Y_p - \hat{\mu}_p) = \mathrm{Var}(\mu_p + \sigma Z_p - \hat{\mu}_p) =$

$= \sigma^2 + \mathrm{Var}(\hat{\mu}_p) = \sigma^2(1 + \frac{1}{n} + \frac{1}{n-1} \cdot (\frac{x_p - \bar{x}}{s_x})^2).$



confidence interval goes to zero, while the width of the 95% prediction interval converges to 1.96σ.

10.4 Coefficients of multiple determination and model utility tests

$R^2 = 1 - \frac{\mathrm{ss_E}}{\mathrm{ss_T}}, \qquad \mathrm{ss_T} = (n-1)s_y^2. \quad R_a^2 = 1 - \frac{n-1}{n-p} \cdot \frac{\mathrm{ss_E}}{\mathrm{ss_T}}.$

$R_a^2 = 1 - \frac{s^2}{s_y^2},$

10.3 Multiple linear regression

$y_1 = \beta_0 + \beta_1 x_{1,1} + \ldots + \beta_{p-1}x_{1,p-1} + e_1,$

$\ldots$

$y_n = \beta_0 + \beta_1 x_{n,1} + \ldots + \beta_{p-1}x_{n,p-1} + e_n,$

$\mathbf{y} = (y_1, \ldots, y_n)^\intercal, \quad \boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\intercal, \quad \boldsymbol{e} = (e_1, \ldots, e_n)^\intercal,$

$$\mathbb{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,p-1} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n,1} & \ldots & x_{n,p-1} \end{pmatrix}.$$

$\boldsymbol{b} = (\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal\mathbf{y},$

:cted responses $\hat{\mathbf{y}} = \mathbb{X}\boldsymbol{b}$:

$\hat{\boldsymbol{y}} = \mathbb{P}\boldsymbol{y}$, where $\mathbb{P} = \mathbb{X}(\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal$

$p \times p$ matrix with elements $\mathrm{Cov}(B_i, B_j)$.

$\mathrm{E}\{(\boldsymbol{B} - \boldsymbol{\beta})(\boldsymbol{B} - \boldsymbol{\beta})^\intercal\} = \sigma^2(\mathbb{X}^\intercal\mathbb{X})^{-1}.$

$\hat{\boldsymbol{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbb{I} - \mathbb{P})\mathbf{y}$

a zero mean vector and a covariance matrix $\sigma^2(\mathbb{I} - \mathbb{P})$.

Denote by $d_0^2, \ldots, d_{n-1}^2$ the diagonal elements of the matrix $(\mathbb{X}^\intercal\mathbb{X})^{-1}$

standard error of b_i $s_{b_j} = sd_j,$

$\frac{B_j - \beta_j}{S_{B_j}} \sim t_{n-p}, \quad j = 0, 1, \ldots, p-1.$

Parametric models

$\mu = \mathrm{E}(X), \quad \sigma^2 = \mathrm{Var}(X). \quad \mathrm{Var}(X) = \mathrm{E}((X-\mu)^2).$

$\mathrm{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \mathrm{E}(S^2) = \sigma^2, \quad \mathrm{Var}(S^2) = \frac{\sigma^4}{n}\left(\mathrm{E}(\frac{X-\mu}{\sigma})^4 - \frac{n-3}{n-1}\right).$

$\mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)dx \quad$ or $\quad \mathrm{E}(X) = \sum_{i=1}^{\infty} x_i p_i,$

z-score, $\mathrm{Cov}(X_1, X_2) = \mathrm{E}((X_1-\mu_1)(X_2-\mu_2)).$

$Z = \frac{X-\mu}{\sigma}$, $\mathrm{E}(X) = \mu$, variance $\mathrm{Var}(X) = \sigma^2$,

$X \sim \mathrm{N}(\mu, \sigma) \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$

$\Phi(x) = \mathrm{P}(Z \le x), \quad -\infty < x < \infty.$

$\mathrm{P}(X \le x) = \mathrm{P}(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}) = \mathrm{P}(Z \le \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma}).$

$\bar{X} = \frac{X_1 + \ldots + X_n}{n} \quad \bar{X} \approx \mathrm{N}(\mu, \frac{\sigma}{\sqrt{n}}),$

Mixtures of normal distribution

$\sigma^2 = \sum_{j=1}^{k} w_j(\mu_j - \mu)^2 + \sum_{j=1}^{k} w_j \sigma_j^2.$

law of total variance,

variation between the strata $\sum_{j=1}^{k} w_j(\mu_j - \mu)^2$,

variation within the strata $\sum_{j=1}^{k} w_j \sigma_i^2$.

Poisson,

$X \sim \mathrm{Pois}(\mu)$ is a discrete distribution

$$\mathrm{P}(X = x) = \frac{\mu^x}{x!} e^{-\mu}, \quad x = 0, 1, \ldots,$$

$\mathrm{E}(X) = \mu, \quad \mathrm{Var}(X) = \mu.$

Poisson distribution is obtained as an approximation for the $\mathrm{Bin}(n, p)$ distribution i

$n \to \infty,\ p \to 0,$ and $np \to \mu.$

describe the number of rear events (like accidents) observed during a given time interval.

Geometric distribution

sequence of independent Bernoulli trials with probability p of success.

the distribution of the number X of trials needed to get one success,

the distribution of the number $Y = X - 1$ of failures before the first success.

$X \sim \mathrm{Geom}(p) \quad \mathrm{P}(X = x) = (1-p)^{x-1}p, \quad x = 1, 2, \ldots, \quad \mu = \frac{1}{n}, \quad \sigma^2 = \frac{1-p}{n^2}.$

Beta distribution

Beta distribution $\mathrm{Beta}(a, b)$ is determined by two parameters $a > 0$, $b > 0$ which are called pseudo-counts. It is defined by the probability density function

$$g(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}, \quad 0 < p < 1,$$

its mean and variance are given by

$$\mu = \frac{a}{a+b}, \quad \sigma^2 = \frac{\mu(1-\mu)}{a+b+1}.$$

2.1 Point estimation

mean square error $\mathrm{E}((\widehat{\Theta} - \theta)^2) = \mathrm{Var}(\widehat{\Theta}) + (\mathrm{E}(\widehat{\Theta}) - \theta)^2$

the bias size $\mathrm{E}(\widehat{\Theta}) - \theta$, measuring the lack of accuracy (systematic error),

$\mathrm{Var}(\widehat{\Theta})$, measuring the lack of precision (random error).

mean square error vanishes $\mathrm{E}((\widehat{\Theta} - \theta)^2) \to 0$ as $n \to \infty$, the point estimate $\hat{\theta}$ is called *consistent*.

$$I_\mu \approx \bar{x} \pm z(\tfrac{\alpha}{2}) \cdot \frac{s}{\sqrt{n}}.$$

2.6 Exact confidence intervals

$$I_\mu = \bar{x} \pm t_{n-1}(\tfrac{\alpha}{2}) \cdot s_{\bar{x}},$$

$I_{\sigma^2} = \left(\frac{(n-1)s^2}{x_{n-1}(\frac{\alpha}{2})}; \frac{(n-1)s^2}{x_{n-1}(1-\frac{\alpha}{2})}\right).$

$\mathcal{F}(\mu, \sigma) = \mathrm{N}(\mu, \sigma), \ \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

t-distribution with $n-1$ degrees of freedom

$f(x) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}\left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}, \quad -\infty < x < \infty$

gamma function $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx,$

$\Gamma(k) = (k-1)!, \quad k = 1, 2, \ldots$

if $Z, Z_1, \ldots, Z_k$ are independent random variables with $\mathrm{N}(0,1)$-distribution,

$$\frac{Z}{\sqrt{(Z_1^2 + \ldots + Z_k^2)/k}} \sim t_k.$$

1.5 Bernoulli, binomial, and multinomial distributions

Bernoulli, $\mathrm{P}(X = 1) = p, \quad \mathrm{P}(X = 0) = 1 - p.$

binomial, $X \sim \mathrm{Bin}(1, p) \quad \mu = p, \quad \sigma^2 = p(1-p).$

$\mathrm{Bin}(n, p).$ $\mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, \ldots, n,$

$\mu = np, \quad \sigma^2 = np(1-p).$ $\mathrm{Bin}(n, p) \approx \mathrm{N}(np, \sqrt{np(1-p)}).$

rule of thumb good enough if both $np \ge 5$ and $n(1-p) \ge 5$,

$\mathrm{P}(X \le x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right).$

Continuity correction

suggests replacing x by $x + \frac{1}{2}$ on the right hand side

$\mathrm{P}(X \le x) = \mathrm{P}(X < x + 1), \quad \mathrm{P}(X \le x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right),$

$\mathrm{P}(X \le x) \approx \Phi\left(\frac{x + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right),$

$\mathrm{P}(X < x) \approx \Phi\left(\frac{x - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$

2.3 Simple random sample

- *sampling with replacement* produces what we call a random sample consisting of independent and identicall distributed observations,
- sampling without replacement produces a so called *simple random sample* having identically distributed but dependent observations.

$\mathrm{E}(S^2) = \frac{n}{n-1}\mathrm{E}\left(\frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2\right) = \frac{n}{n-1}(\mathrm{E}(X^2) - \mathrm{E}(\bar{X}^2))$

$= \frac{n}{n-1}(\sigma^2 + \mu^2 - \frac{\sigma^2}{n}(1 - \frac{n-1}{N-1}) - \mu^2) = \frac{n}{n-1}(\sigma^2 - \frac{\sigma^2}{n}(1 - \frac{n-1}{N-1})) = \sigma^2 \frac{N}{N-1},$

$s_{\bar{x}}^2 = \frac{s^2}{n}\frac{N-1}{N}(1 - \frac{n-1}{N-1}) = \frac{s^2}{n}(1 - \frac{n}{N}).$ unbiased estimate of $\mathrm{Var}(\bar{X})$:

sampling without replacement, the formula for the estimated standard error of the sample mean $\bar{x}$

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}. \quad I_\mu \approx \bar{x} \pm z(\alpha/2)s_{\bar{x}},$$

Gamma, $\mathrm{Gam}(\alpha, \lambda) \quad \mu = \frac{\alpha}{\lambda}, \quad \sigma^2 = \frac{\alpha}{\lambda^2}.$

shape parameter $\alpha > 0$, inverse scale (or the rate) parameter $\lambda > 0$.

restricted to positive values,

if $\alpha = 1$, $\mathrm{Gam}(1, \lambda) = \mathrm{Exp}(\lambda)$. exponential,

if $X_i \sim \mathrm{Exp}(\lambda)$, $i = 1, \ldots, k$ are independent,

$X_1 + \ldots + X_k \sim \mathrm{Gam}(k, \lambda), \quad k = 1, 2, \ldots$

large values of the shape parameter

$\mathrm{Gam}(\alpha, \lambda) \approx \mathrm{N}(\frac{\alpha}{\lambda}, \frac{\sqrt{\alpha}}{\lambda}), \qquad \alpha \gg 1$ **chi-squared**

chi-squared distribution with k degrees of freedom is

the gamma distribution with $\alpha = \frac{k}{2}, \lambda = \frac{1}{2}$.

if $Z_1, \ldots, Z_k$ are independent random variables with $\mathrm{N}(0,1)$-distribution $Z_1^2 + \ldots + Z_k^2 \sim \chi_k^2.$

if $(X_1, \ldots, X_n)$ are independent and random variables each having the $\mathrm{N}(\mu, \sigma)$

$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2.$

multinomial distribution $(X_1, \ldots, X_r) \sim \mathrm{Mn}(n; p_1, \ldots, p_r)$

$\mathrm{P}(X_1 = x_1, \ldots, X_r = x_r) = \binom{n}{x_1, \ldots, x_r} p_1^{x_1} \ldots p_r^{x_r}, x_i = 0, \ldots, n, \quad i = 1, \ldots, r,$

$p_1 + \ldots + p_r = 1.$ marginal distribution of X_i is binomial $X_i \sim \mathrm{Bin}(n, p_i)$

$\mathrm{Cov}(X_i, X_j) = -np_ip_j, \quad i \ne j.$

Hypergeometric distribution $X \sim \mathrm{Hg}(N, n, p)$

number x of black balls among n balls drawn without replacement from a box with N balls, of which

$B = Np$ balls are black and

$W = N(1-p)$ balls are white.

X is the number of successes in n Bernoulli trials which depend on each other.

$\mathrm{P}(X = x) = \frac{\binom{B}{x}\binom{W}{n-x}}{\binom{N}{n}}, \quad \max(0, n - W) \le x \le \min(n, B).$

$\mu = np, \quad \sigma^2 = np(1-p)\frac{N-n}{N-1}.$

Compared to the variance of the $\mathrm{Bin}(n, p)$ distribution, the last formula contains the factor $\frac{N-n}{N-1} = 1 - \frac{n-1}{N-1}$, *finite population correction* factor.

$\mathrm{Hg}(N, n, p) \approx \mathrm{N}(\mu, \sigma), \qquad \mu = np, \quad \sigma = \sqrt{np(1-p)}\sqrt{1 - \frac{n-1}{N-1}},$

applied provided $np \ge 5$ and $n(1-p) \ge 5$.

2.4 Dichotomous data

The estimated standard error for the sample proportion $\hat{p}$ is $s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$.

Simple random sample $s_{\hat{x}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\sqrt{1 - \frac{n}{N}}.$

An approximate 100(1–α)% two-sided confidence interval for p is given by $I_p \approx \hat{p} \pm z(\frac{\alpha}{2}) \cdot s_{\hat{p}}$

valid both for sampling with replacement and for sampling without

2.5 Stratified random sample

$\mu = w_1\mu_1 + \ldots + w_k\mu_k,$

$\sigma^2 = \overline{\sigma^2} + \sum_{i=1}^{k} w_j(\mu_j - \mu)^2.$

$\overline{\sigma^2} = w_1\sigma_1^2 + \ldots + w_k\sigma_k^2$

$\mathrm{Var}(\bar{X}_s) = w_1^2\mathrm{Var}(\bar{X}_1) + \ldots + w_k^2\mathrm{Var}(\bar{X}_k) = \frac{w_1^2\sigma_1^2}{n_1} + \ldots + \frac{w_k^2\sigma_k^2}{n_k}.$

$s_{\bar{x}_s}^2 = w_1^2 s_{\bar{x}_1}^2 + \ldots + w_k^2 s_{\bar{x}_k}^2 = \frac{w_1^2 s_1^2}{n_1} + \ldots + \frac{w_k^2 s_k^2}{n_k},$

A confidence interval for the mean based on a stratified sample: $I_\mu \approx \bar{x}_s \pm z(\frac{\alpha}{2}) \cdot s_{\bar{x}_s}$

$\bar{\sigma} = w_1\sigma_1 + \ldots + w_k\sigma_k.$

The stratified sample mean $\bar{X}_{so}$ with the optimal allocation $n_j = n\frac{w_j\sigma_j}{\bar{\sigma}}$ has the smallest variance $\mathrm{Var}(\bar{X}_{so}) = \frac{\bar{\sigma}^2}{n}$ among all allocations of n observations.

The stratified sample mean $\bar{X}_{sp}$ for the proportional allocation $n_j = nw_j$ has the variance $\mathrm{Var}(\bar{X}_{sp}) = \frac{\overline{\sigma^2}}{n}$.

$\mathrm{Var}(\bar{X}_{so}) \le \mathrm{Var}(\bar{X}_{sp}) \le \mathrm{Var}(\bar{X}),$

$\frac{(\bar{\sigma})^2}{n} \le \frac{\overline{\sigma^2}}{n} \le \frac{\sigma^2}{n}.$

The statistical tools introduced in this course so far are based on the so called *frequentist approach*. In the parametric case, the frequentist treats the data x as randomly generated by a distribution $f(x|\theta)$ involving the unknown true population parameter value θ, which may be estimated using the method of maximum likelihood. This section presents basic concepts of the *Bayesian approach* relying on the following model for the observed data x:

$$\text{Apriori distribution} \xrightarrow{g(\theta)} \text{generates a value } \theta \xrightarrow{f(x|\theta)} \text{data } x.$$

The model assumes that before the data is collected the parameter of interest θ is randomly generated by a prior distribution $g(\theta)$. The computational power of the Bayesian approach stems from the possibility to treat θ as a realisation of a random variable Θ.

The prior distribution $g(\theta)$ brings into the statistical model our knowledge (or lack of knowledge) on θ before the data x is generated using a conditional distribution $f(x|\theta)$, which in this section is called the likelihood function. After the data x is generated by such a two-step procedure involving the pair $g(\theta)$ and $f(x|\theta)$, we may update our knowledge on θ and compute a posterior distribution $h(\theta|x)$ using the Bayes formula

$$h(\theta|x) = \frac{f(x|\theta)g(\theta)}{\phi(x)}.$$

The denominator, depending on whether the distribution is continuous or discrete,

$$\phi(x) = \int f(x|\theta)g(\theta)d\theta \quad \text{or} \quad \phi(x) = \sum_{\theta} f(x|\theta)g(\theta)$$

gives the marginal distribution of the random data X. For a fixed realization x, treating the denominator $\phi(x)$ as a constant which does not explicitly involve θ, the Bayes formula can be summarized as

$$\boxed{\text{posterior} \propto \text{likelihood} \times \text{prior}}$$

where the sign $\propto$ means proportional.

If we have no prior knowledge on θ, the prior distribution is often modelled by the uniform distribution. In this case of uninformative prior, with $g(\theta)$ being a constant over a certain interval, we have $h(\theta|x) \propto f(x|\theta)$, implying that the posterior knowledge comes solely from the likelihood function.

Multinomial-Dirichlet model

The multinomial-Dirichlet model is a multivariate version of the binomial-beta model. For both the binomial-beta and multinomial-Dirichlet models, the updating rule has the form

$$\boxed{\text{the posterior pseudo-counts} = \text{the prior pseudo-counts plus the sample counts}}$$

Dirichlet distribution

The Dirichlet distribution $\text{Dir}(\alpha_1, \dots, \alpha_r)$ is a multivariate extension of the beta distribution. It is a probability distribution over the vectors $(p_1, \dots, p_r)$ with non-negative components such that

$$p_1 + \dots + p_r = 1.$$

The positive parameters $\alpha_1, \dots, \alpha_r$ of the Dirichlet distribution are often called the pseudo-counts. The probability density function of $\text{Dir}(\alpha_1, \dots, \alpha_r)$ is given by

$$g(p_1, \dots, p_r) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_r)} p_1^{\alpha_1 - 1} \dots p_r^{\alpha_r - 1}, \quad \alpha_0 = \alpha_1 + \dots + \alpha_r.$$

The marginal distributions of the random vector $(X_1, \dots, X_r) \sim \text{Dir}(\alpha_1, \dots, \alpha_r)$ are the beta distributions

$$X_j \sim \text{Beta}(\alpha_j, \alpha_0 - \alpha_j), \quad j = 1, \dots, r.$$

Different components of the vector have negative covariances

$$\text{Cov}(X_i, X_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \text{ for } i \neq j.$$

The figure below illustrates four examples of $\text{Dir}(\alpha_1, \alpha_2, \alpha_3)$ distribution. Each triangle contains $n = 300$ points generated using different sets of parameters $(\alpha_1, \alpha_2, \alpha_3)$:

$$\text{upper left } (0.3, 0.3, 0.1), \text{ upper right } (13, 16, 15), \text{ lower left } (1, 1, 1), \text{ lower right } (3, 0.1, 1).$$

A dot in a triangle gives a realisation (x_1, x_2, x_3) of the vector $(X_1, X_2, X_3) \sim \text{Dir}(\alpha_1, \alpha_2, \alpha_3)$ as the distances to the bottom edge of the triangle (x_1), to the right edge of the triangle (x_2), and to the left edge of the triangle (x_3).

5.3 Credibility interval

Given data x coming from a parametric model with the likelihood function $f(x|\theta)$, a $100(1-\alpha)\%$ confidence interval for the parameter θ,

$$I_\theta = (a_1(x), a_2(x)),$$

is viewed as a realisation of a random interval $(a_1(X), a_2(X))$ such that

$$\text{P}(a_1(X) < \theta < a_2(X)) = 1 - \alpha.$$

This frequentist interpretation of the confidence level $100(1-\alpha)\%$ is rather cumbersome as it requires mentioning other samples and potential confidence intervals which as a group cover the true unknown value of θ with probability $1 - \alpha$.

In the framework of Bayesian inference we can refer to θ as a realisation of a random variable Θ with a certain posterior distribution $h(\theta|x)$. This allows us to define a $100(1-\alpha)\%$ credibility interval (or credible interval)

$$J_\theta = (b_1(x), b_2(x))$$

by the relation based on the posterior distribution

$$\text{P}(b_1(x) < \Theta < b_2(x)|x) = 1 - \alpha.$$

The interpretation of the credibility interval is more intuitive as it does not refer to some potential, never observed data values.

5.1 Conjugate priors

Suppose the data x is generated by a parametric model having the likelihood function $f(x|\theta)$. Consider a parametric family of the prior distributions $\mathcal{G}$.

$$\boxed{\begin{array}{l}\mathcal{G} \text{ is called a family of conjugate priors for the likelihood function } f(x|\theta) \\ \text{if for any prior } g(\theta) \in \mathcal{G}, \text{ the corresponding posterior distribution } h(\theta|x) \in \mathcal{G}\end{array}}$$

The next table presents five Bayesian models involving conjugate priors. The details of the first three models come next. Notice that the posterior variance is always smaller than the prior variance. This list also illustrates that the contribution of the prior distribution to the posterior distribution becomes smaller as the sample size n increases.

Parametric model for the data	Unknown θ	Prior	Posterior distribution
$X_1, \dots, X_n \sim \text{N}(\mu, \sigma)$	$\theta = \mu$	$\text{N}(\mu_0, \sigma_0)$	$\text{N}(\gamma_n\mu_0 + (1-\gamma_n)\bar{x}; \sigma_0\sqrt{\gamma_n})$
$X \sim \text{Bin}(n, p)$	$\theta = p$	$\text{Beta}(a, b)$	$\text{Beta}(a + x, b + n - x)$
$(X_1, \dots, X_r) \sim \text{Mn}(n; p_1, \dots, p_r)$	$\theta = (p_1, \dots, p_r)$	$\text{Dir}(\alpha_1, \dots, \alpha_r)$	$\text{Dir}(\alpha_1 + x_1, \dots, \alpha_r + x_r)$
$X_1, \dots, X_n \sim \text{Geom}(p)$	$\theta = p$	$\text{Beta}(a, b)$	$\text{Beta}(a + n, b + n\bar{x} - n)$
$X_1, \dots, X_n \sim \text{Pois}(\mu)$	$\theta = \mu$	$\text{Gam}(\alpha_0, \lambda_0)$	$\text{Gam}(\alpha_0 + n\bar{x}, \lambda_0 + n)$
$X_1, \dots, X_n \sim \text{Gam}(\alpha, \lambda)$	$\theta = \lambda$	$\text{Gam}(\alpha_0, \lambda_0)$	$\text{Gam}(\alpha_0 + \alpha n, \lambda_0 + n\bar{x})$

Normal-normal model

Suppose a random sample $(x_1, \dots, x_n)$ is drawn from the normal distribution $\text{N}(\mu, \sigma)$ with a known standard deviation σ and the unknown mean $\theta = \mu$. Taking the normal prior $\Theta \sim \text{N}(\mu_0, \sigma_0)$ with known (μ_0, σ_0) results in the normal posterior $\text{N}(\mu_1, \sigma_1)$ with

$$\mu_1 = \gamma_n\mu_0 + (1 - \gamma_n)\bar{x}, \quad \sigma_1^2 = \sigma_0^2\gamma_n,$$

where

$$\gamma_n = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2} = \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \sigma_0^2}$$

is the shrinkage factor which becomes smaller for the larger sample sizes n. As a result for the large samples, the posterior mean μ_1 gets close to the maximum likelihood estimate and the input $\gamma_n\mu_0$ involving the prior mean becomes negligible.

Binomial-beta model

Next, we introduce the beta distribution which serves as a convenient family of conjugate priors for Bayesian inference for p, in the case when the data x is generated by the $\text{Bin}(n, p)$.

5.2 Bayesian estimation

In the language of decision theory, finding a point estimate a for the unknown population parameter θ is an action of assigning the value a to the unknown parameter θ. In the frequentist setting, the optimal a is found by maximising the likelihood function. In the Bayesian setting, the optimal choice of a is determined by the so-called loss function $l(\theta, a)$. The so-called Bayes action minimises the posterior risk

$$R(a|x) = \text{E}\big(l(\Theta, a)|x\big),$$

computed using the posterior distribution

$$R(a|x) = \int l(\theta, a)h(\theta|x)d\theta \quad \text{or} \quad R(a|x) = \sum_{\theta} l(\theta, a)h(\theta|x).$$

We consider two loss functions leading to two different Bayesian estimators. These two loss functions called the zero-one loss and the squared error loss

$$\boxed{\text{Zero-one loss function: } l(\theta, a) = 1_{\{\theta \neq a\}}} \quad \boxed{\text{Squared error loss: } l(\theta, a) = (\theta - a)^2}$$

are schematically depicted on the figure below.

Zero-one loss function and maximum a posteriori probability

With the zero-one loss function, the posterior risk is equal to the probability of misclassification

$$R(a|x) = \sum_{\theta \neq a} h(\theta|x) = 1 - h(a|x).$$

In this case, to minimise the risk we have to maximise the posterior probability $h(a|x)$. We define of θ that maximises $h(\theta|x)$. Observe that with the uninformative prior, $\hat{\theta}_{\text{map}} = \hat{\theta}_{\text{mle}}$.

define $\hat{\theta}_{\text{map}}$ as the value

Squared error loss function and posterior mean estimate

Using the squared error loss function we find that the posterior risk is a sum of two components

$$R(a|x) = \text{E}((\Theta - a)^2|x) = \text{Var}(\Theta|x) + (\text{E}(\Theta|x) - a)^2.$$

Since the first component is independent of a, we minimise the posterior risk by putting

$$\hat{\theta}_{\text{pme}} = \text{E}(\Theta|x),$$

resulting in the posterior mean value as the Bayesian point estimate of θ.