Central limit theory

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

$$M_{Z_n(t)} = E\left[\exp\left(t \cdot \frac{S_n - n\mu}{\sigma\sqrt{n}}\right)\right]$$

$$= E\left[\prod_{i=1}^{n} \exp\left(\frac{X_i - \mu}{\sigma\sqrt{n}} \cdot t\right)\right]$$

$X_1 \cdots X_n$ are iid

$$\Rightarrow = \prod_{i=1}^{n} E\left[\exp\left(\frac{X_i - \mu}{\sigma\sqrt{n}} \cdot t\right)\right]$$

$$= \left\{E\left(\exp\left(t \cdot \frac{(X-\mu)}{\sigma\sqrt{n}}\right)\right)\right\}^n$$

$$= \left\{E\left[\right) + \frac{t}{\sigma\sqrt{n}}(X-\mu) + \frac{t^2}{2\sigma^2 n}(X-\mu)^2 + \text{higher order}\right\}^n$$

($\frac{t}{\sigma\sqrt{n}}(X-\mu)$ → 0)

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty} \left(1 + \frac{t^2}{2n} + R(t)\right)^n$$

$$= e^{t^2/2}$$

MGF of $N(0,1)$

Statistic for mean and var

$$\hat{\mu} = \frac{1}{n}\sum X_i$$

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

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

$$\text{Var } S^2 = \frac{E(X-\mu)^4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}$$

quatiles vs order statistic

$$\hat{F}(x) = \begin{cases} 0 & \text{if } x < X_{(1)} \\ k/n & \text{if } X_{(k)} \le x \le X_{(k+1)} \\ 1 & \text{if } x \ge x_{(n)} \end{cases}$$

$$\hat{C}_q = x_{(k)} \quad \text{if } \frac{k-1}{n} < q < \frac{k}{n}$$

$$\hat{C}_q = \frac{1}{2}\left(X_{(q\cdot n)} + X_{(q\cdot n+1)}\right)$$

$$f_{X_{(1)},\cdots,X_{(n)}}(x_{(1)},\cdots x_{(n)}) = n!\, f_x(x_1)\cdots f_{n-k}(x_n)$$

$$f_{X_{(k)}}(u) = k\binom{n}{k} F_x(u)^{k-1}(1-F_x(u))^{n-k} f_x(u)$$

$$= n\binom{n-1}{k-1} \cdots$$

$$\hat{C}_q \overset{d}{\approx} N\left(C_q, \frac{q(1-q)}{n f(q)^2}\right)$$

Statistic dist

$X_1 \cdots X_n$, iid, $X \sim N(0,1)$

① $U = \sum_{i=1}^{n} X_n^2 \sim \chi_n^2$ $\quad M_U(t) = (1-2t)^{-\frac{n}{2}}$

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum X_i^2 - n\bar{X}^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 - \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi_{n-1}^2$$

② $U = \sum_{i=1}^{n} X_n$, $U \sim N(0, \frac{n}{n})$

Statistic dist

$X_1 \cdots X_n$, iid, $X \sim N(\mu, \sigma^2)$

① for true σ^2 $\quad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

② don't know σ^2, approxi σ^2 by s^2

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \overset{d}{\sim} t_{n-1}; \quad \frac{Z}{\sqrt{U/\nu}} \overset{d}{=} t_{\nu-1}, \quad Z=N(0,1),\ U\sim\chi_\nu^2$$

for prediction $\quad \frac{(\bar{X}-\mu)/\sigma\sqrt{n}}{\sqrt{\frac{(n-1)s^2}{\sigma^2}/n-1}} = \frac{\bar{X}-\mu}{s/\sqrt{n}}$

$$\frac{X^* - \bar{X}}{s\sqrt{1+\frac{1}{n}}} \overset{d}{\sim} t_{n-1}, \quad X^* = \bar{X} \pm t_{n-1}^{\alpha/2} \cdot s\sqrt{1+\frac{1}{n}}$$

③ $U \overset{d}{=} \chi_m^2$, $V \overset{d}{=} \chi_n^2$

$$\frac{U/m}{V/n} \overset{d}{=} F_{m,n}$$

④ 1' $\quad \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{d}{\sim} N(0,1)$

2' $\quad \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{d}{\sim} t_{n_1+n_2-2}$

assume equal Var

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

3' know ratio of $\sigma_1 = k\sigma_2$

$$\frac{[(\bar{X}_1 - \bar{X}_2) - (\mu_1-\mu_2)]/\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}{\sqrt{U/n_1+n_2-2}} \overset{d}{=} t_{n_1+n_2-2}$$

$$U = \frac{(n_1-1)s_1^2}{\sigma_1^2} + \frac{(n_2-1)s_2^2}{\sigma_2^2} \overset{d}{=} \chi_{n_1+n_2-2}^2$$

method of moment

$$E(x) = \bar{x} = \int x \cdot f(x)\, dx$$

$$= \theta^2 + b$$

$$\bar{\theta}^2 = \bar{x} - b \Rightarrow \bar{\theta} = \pm\sqrt{\bar{x}-b}$$

MLE estimator

Unbiasedness : $E(T(n)) = \theta$

Consistency : $E(T(n)) \to \theta$, $\text{Var}(T_n) \to 0$

Efficiency. $\text{Var}(T_n) \ge \frac{1}{I(\theta)}$ $\quad \text{eff}(T_n) = \frac{MVB}{\text{Var}(T_n)} \le 1$

$D_1 = \frac{\partial}{\partial\theta}\log L$, $\int \frac{\partial}{\partial\theta}\log L(x)\cdot L(x)\, dx$

$$= \frac{\partial}{\partial\theta}\int L(x)\, dx = 0$$

$$= E(D_1)$$

$$= \int \frac{\partial^2}{\partial\theta^2}\log L(x)\cdot L(x)\, dx + \int \frac{\partial}{\partial\theta}\log L(x)\frac{\partial}{\partial\theta}L(x)\, dx$$

$= E(D_2) + E(D_1^2)$ $\quad E(T) = \theta$

$= 0$

$\text{Cov}(T, D_1) = E(TD_1) - E(T) \cdot E(D_1)$

$= \int T(x) \frac{\partial \log L(x)}{\partial \theta} L(x) dx$

$= 1$

$\frac{\text{Cov}(T, D_1)}{\text{Var}\, T \cdot \text{Var}\, D_1} = \frac{1}{\text{Var}\, T \cdot \text{Var}\, D_1} \leq 1$

$\Rightarrow \text{Var}\, T \geq \frac{1}{\text{Var}\, D_1} = \frac{1}{I(\theta)}$

	$I(\theta)$
$N(0, \theta)$	$n/2\theta^2$
$N(\theta, \sigma^2)$	n/σ^2

$\text{Var}\, T_{MLE} \approx \frac{1}{I(\theta)}$

$I(\theta) = -E(D_2) = E(D_1^2) = \text{Var}(D_1)$

log ratio test

$\frac{L_1}{L_0} = \exp\left(\frac{1}{2} n (\bar{x} - \theta_0)^2\right)$

$\log \frac{L_1}{L_0} = \frac{1}{2} n (\bar{x} - \theta_0)^2 = \frac{1}{2} \frac{(\bar{x} - \theta_0)^2}{(1/\sqrt{n})^2} \overset{d}{=} \frac{1}{2} \chi_1^2$

Reject H_0 if $\log \frac{L_1}{L_0} > \frac{1}{2} (t_{0.95}(\chi_1^2))$

Decision making: the ultimate goal of using statistics is to make decisions. The framework we use for this, is hypothesis testing

$\sum_{i=1}^{n} N(0,1)^2 \sim \chi^2_n$

$\frac{N(0,1)}{\sqrt{\chi^2_{n-1}/n-1}} \sim t_{n-1} \Leftrightarrow \frac{\bar{x}-\mu}{s/\sqrt{n}}$

$\frac{\chi^2_{m-1}/m-1}{\chi^2_{n-1}/n-1} \sim F_{m-1,n-1}$

MM & MLE

MM: $E(X) = \mu = \frac{1}{n}\sum X_i$

$E(X^2) = \sigma^2+\mu^2 = \frac{1}{n}\sum X_i^2$

$\Rightarrow \mu = \sim \quad \sigma^2 = \sim$

MLE: $\frac{\partial L}{\partial \mu} = 0 \quad \frac{\partial L}{\partial \sigma^2} = 0$

$\Rightarrow \mu \sim \quad \sigma^2 = \sim$

MVB = Var $T_{MLE} = \frac{1}{I(\theta)} = -\frac{1}{E(D_2)}$

likelihood ratio test

test statistic $\frac{L(\theta_1)}{L(\theta_0)}$ or $\log L(\theta_1) - \log L(\theta_0)$ — $L(\theta)$ is a function of S.S — need to assume a dist first

MC: Inverse-tranform method

$X = F^{-1}(u) \sim F(x)$

$\Rightarrow P(X<x) = P(F^{-1}(U) \le x)$

$= P(U \le F(x))$

$= F(x)$

First: work out $F(x)=u$ (sometimes impossible)

Second: $X = F^{-1}(u)$

ex: $F_x(x) = 1-e^{-x} = u$

$x = -\log(1-u) \quad u \in (0,1)$

The AR algorithm

① Generate Y and U. $U \in [0,1]$

② if $U < \frac{f(Y)}{Mg(y)}$ accept, or X

③ generate a new Y (random walk)

$M = \max_y \frac{f(y)}{g(y)}$

prob of being accepted

$P(U \le \frac{f(y)}{M \cdot g(y)})$

$= \int P(U \le \frac{f(y)}{M \cdot g(y)}) \cdot g(y) dy$

$= \int \frac{f(y)}{M \cdot g(y)} \cdot g(y) dy$

$= 1/M$

Distribution free method

derive test statistic without assume dist

χ^2 test: χ^2_{k-p-1} $\quad \sum_{i=1}^{k} \frac{(obs - nP_i^{exp})^2}{nP_i^{exp}} = u$ test stat

$\chi^2_{(r-1)(c-1)}$ independent test

median inf test: Sign test. $H_0: m = m_0$

T.S.: $Z = freq(X \le m_0)$

$Z|_{H_0} \sim N(\frac{n}{2}, \frac{n}{4})$ or $Z|_{H_0} \sim Bi(n, 0.5)$

Pvalue $= 2 \times P(Z \ge \#obs)$

Rank test, Wilcoxon

$H_0: X_{median} = X^0$

rank$(x<x^0)$ − rank$(x>x^0)$

$\sim N(0, \frac{n(n+1)(2n+1)}{6})$

median equality test

H_0 Median $X = Y$

Contingency table, χ^2 test

$H_0: m_x = m_y \Rightarrow m^*$ coverall

	$\le m^*$	$> m^*$	
X			n_1 exp to be equal
Y			n_2

Sufficient Statistics

Fisher-Neyman factorisation theorem,

$L(\vec{x}) = k_1(\underbrace{t(\vec{x})}_{S.S}|\theta) \cdot k_2(\vec{x})$

Exponential: $f(x|\theta) = \exp(p(\theta)\cdot k(x) + H(x) + q(\theta))$

$Y = \sum k(X_i)$. S.S

$EY = -n\frac{q'(\theta)}{p'(\theta)}$; if $E(Y) = n\theta \Rightarrow \bar{Y}_n \Rightarrow$ MVB

Var $\hat{\theta}$ ← Var $Y = \frac{n}{p'(\theta)^3}[p''(\theta)q'(\theta) - q''(\theta)p'(\theta)]$

Completeness: A S.S in complete if it comes from a family of dist such that iff for all θ, $E(u(y)) = 0 \Rightarrow u = 0$

Joint test equality of all means

Assume: k group: $\bar{X}_{ks}, S^2_k$ one way $(X_{kj} = \mu_k + a_j, \sum a_j = 0)$ $\quad \frac{S^2_k \cdot (n-1)}{\sigma^2} \sim \chi^2_{n-1}$

true "S" $= W = \sum_{i=1}^{k}(n-1)S_k \sim \sigma^2 \cdot \chi_{N-k}$ (always)

$(S_B^2 = \frac{1}{k-1}\sum(\bar{X}_i - \bar{\bar{X}})^2)$ $\quad H_0: \bar{x}$ for all group

$B = n(k-1)S_B^2 \sim \sigma^2 \chi^2_{k-1}$ if H_0 is true

$T = W + B$

one way ANOVA: $\frac{B/k-1}{W/N-k} \sim F_{k-1,N-k}$. reject if P is < 0.01

rank test for not for normal, Comparative inf

W_x (Sum of ranks from X)

$\sim N(\frac{1}{2}n_1(n_1+n_2+1), \frac{1}{12}n_1n_2(n_1+n_2+1))$

(to compare if X, Y are different)

$H_0: X = Y$

① summary of distribution

② rank test more powerful

$\|X_i - m_0\|$	x_1	x_2	...	x_m
rank	1	2	...	m

ANOVA test. for size α

Bonferroni method

$\alpha_E = 1-(1-\alpha_c)^m$

Tukey method

$\bar{x}_i - \bar{x}_j \pm C_{0.95}(Q_{k,N-k})\frac{S}{\sqrt{n}}$

For general linear model (easier to solve)

$y = A\theta + e$ ($\mu=0$) $\quad y = \mu + A'\theta + e$, $e = y_i - \hat{y}_i$

$A^TA\hat{\theta} = A^Ty \Rightarrow \hat{\theta} = (A^TA)^{-1}A^Ty$

$\hat{\theta} \overset{d}{=} N_p(\theta, \sigma^2(A^TA)^{-1})$ $\quad \mathrm{Var}(\hat{\theta}) = \begin{bmatrix} \mathrm{Var}(\hat{\alpha}) & \mathrm{Cov}(\hat{\alpha},\hat{\beta}) \\ \mathrm{Cov}(\hat{\alpha},\hat{\beta}) & \mathrm{Var}(\hat{\beta}) \end{bmatrix}$

$\hat{\theta}^T\hat{e} = \sum(\hat{\alpha}+\hat{\beta}u_i)(y_i - \hat{\alpha} - \hat{\beta}u_i)$

$A^T\cdot e = A^T(y - A\theta)$

$E(A^Te) = A^T(y - A\theta) = 0$

$\hat{\theta} = (A^TA)^{-1}A^Ty$

$S^2 = \frac{1}{n-p}(y^Ty - \hat{\theta}^TA^Ty)$ for σ^2

$\frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p}$

$E(y_i) = A\cdot\theta$, $\mathrm{Var}(y_i) = \mathrm{Var}(A\cdot\theta)$ ex.

$\mathrm{Var}(\alpha+2\beta) = \mathrm{Var}(\alpha) + 4\mathrm{Var}(\beta) + 4\mathrm{Cov}(\alpha,\beta)$

Two way ANOVA & Additive model

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

where $\sum_{i=1}^{a}\alpha_i = 0$ $\quad \sum\beta = 0$

$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}$ $\begin{matrix} \}G1 \\ \}G2 \end{matrix}$ $\quad k$

invasible if

$T = R + C + I + W = B + W$

Between the Row

$B = R + C + I$

if we only one obs, $\Rightarrow C = 1$ $\quad$ set

$\Downarrow$ I not exist

$T = R + C + W$

$B = R + C$

and $W =$

$\bar{x}_{...} \Rightarrow \bar{x}_.$

$R = bc\sum_{\text{Row}}^{a}(\bar{x}_{i..} - \bar{x}_{...})^2$ $\quad \alpha_i$

$C = ac\sum_{\text{Col}}^{b}(\bar{x}_{.j.} - \bar{x}_{...})^2$ $\quad \beta_j$

$I = c\sum_{\text{Row}}^{a}\sum_{\text{Col}}^{b}(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...})^2$ $\quad \epsilon_{ij}$

$W = \sum\sum\sum(x_{ijk} - \bar{x}_{ij.})^2$

$T = \sum\sum\sum(x_{ijk} - \bar{x}_{...})^2$

$\frac{R/a-1}{W/ab(c-1)}$ test row, $\frac{C/b-1}{W/ab(c-1)}$ test col, $\frac{I/ab-a-b+1}{W/ab(c-1)} \overset{(m)}{\sim} F_{m,n}$ test Inte $\overset{(n)}{=}$ test R.C.I $= 0$

if $c \neq 1$, and we use $T = R + C + W$

Then, $W = I + W$

24-6

$\ln y \sim \chi$, $W_1 = (n_1=10)/S^2 = 6.5$

$\ln y \sim x \cdot f$ $\quad W_2 = 6 = (n_2=8)/S^2$

$\frac{(W_1 - W_2)/(n_1 - n_2)}{W_2/n_2}$

Linear regress.

model: $y = a_0 + \beta(x - \bar{x}) = a_0 + \beta u = \underline{a + \beta x}$ (LS linear)

$a = a_0 - \beta\bar{x}$

$\hat{a}_0 = \bar{y}$, $\hat{A}_0$, $\hat{B}$ for $\hat{a}_0$ and $\hat{B}$

$\hat{\beta} = \frac{\sum u_iy_i}{\sum u_i^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

$\hat{\mathrm{Cov}}(\hat{A}_0, \hat{\beta}) = \mathrm{Cov}(\frac{1}{n}\sum y_i - \frac{1}{\sum u_i^2}\sum u_iy_i)$

$\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = \frac{1}{n}\frac{1}{\sum u_i^2}\sum u_i\mathrm{Cov}(y_i, y_i) = 0$

$S^2 = \frac{1}{n-2}\sum(y_i - \hat{A}_0 - \hat{B}u_i)^2 \Rightarrow$

$\frac{S^2(n-2)}{\sigma^2} \sim \chi^2_{n-2}$

$\frac{\hat{A}_0 - \hat{a}_0}{S/\sqrt{n}} \overset{d}{=} t_{n-2}$, $\frac{\hat{B} - \beta}{S/\sqrt{\sum u^2}} \overset{d}{=} t_{n-2}$

$\mathrm{Var}\,\hat{a}_0 = \frac{\sigma^2}{n}$ $\quad \mathrm{Var}\,\hat{B} = \frac{\sigma^2}{\sum u^{2-2}}$

$\mathrm{Var}\,\hat{a} = (\frac{1}{n} + \frac{\bar{x}^2}{\sum u^2})\sigma^2$ $\quad \mathrm{Cov}(a,\beta) = -\frac{\bar{x}^{-2}}{\sum u^2}\sigma^2$

$\hat{y} \overset{d}{=} N(\alpha_0 + \beta u, \frac{\sigma^2}{n} + \frac{(x_0 - \bar{x})^2\cdot\sigma^2}{\sum u_i^2})$

$\sum(y_i - \bar{y})^2 = R^2\sum(y_i - \bar{y})^2 + S^2(n-2)$

$\sum(y_i - \bar{y})^2 = \underbrace{\beta^2\sum u_i^2}_{B} + \underbrace{S^2\cdot(n-2)}_{W}$

ANOVA test $\beta = 0$: $\frac{B/1}{W/n-2} \sim F_{1,n-2}$

$= \frac{R^2(n-2)}{1-R^2}$

$\Rightarrow \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t_{n-2}$

Point estimation ①

Interval estimation

If you have estimator, for a value you know its distribution, Then you can infer a Interval estimation, it is a reasonable estimation.

Unbiased estimator → Sufficient statistics

does Unbiased estimator must be sufficient Statistics ?

MLE

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{d}{=} N(0,1)$$

$$\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \overset{d}{=} t_{n-1}$$

$$U \overset{d}{=} \chi^2_m$$

$$V \overset{d}{=} \chi^2_n$$

$$Z = \frac{U/m}{V/n}$$

$$\frac{(n-1)\cdot S^2}{\sigma^2} \overset{d}{=} \chi_{n-1}$$

$$Z \overset{d}{=} F_{m,n}$$

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \overset{d}{=} F_{n_1-1,\,n_2-1}$$

$$\text{eff}(T_n) = \frac{\text{MVB}}{\text{Var}(T_n)}$$

If $\text{eff}(T_n) = 1$ then we say T_n is an efficient estimator

estimator

- MLE
  - often biased
    - but asymptotically, actually, the best!
      - If $\phi = \psi(\theta)$, MLE of ϕ $\Rightarrow \hat{\phi} = \psi(\hat{\theta})$
      - MLE are consistent, $n \to \infty$
      - efficiency
        - $\frac{\partial \log L}{\partial \theta}$ D_1、D_2
        - $E(D_1) = 0$
        - $E(D_2) = -E(D_1^2)$
        - $E(D_1^2) = \text{Var}(D_1)$
        - Information function (all inform) $= -E(D_2) = I(\theta)$
- Unbiased
  - exist a lower bound for var of any unbiased estimator T
    - $\text{Var}(T) \geq \frac{[k'(\theta)]^2}{I(\theta)}$
    - $n \to \infty$
  - $T_{MLE} \overset{d}{\to} N(\theta, \frac{1}{I(\theta)})$ Theorem
- sufficient Statistics
  - we define a test for a parameter, the test is based on a function of data → S.S.
  - $E(Y_2|Y_1) = \phi(Y)$, $\text{Var}(Y_2) \geq \text{Var}(\phi(Y_1))$
  - MVB estimator must be a function of S.S.
  - Y. S.S for θ, MLE for $\hat{\theta}$ exist and unique, $\hat{\theta}$ is a function of Y, Y is complete S.S

$f(x|\theta) = \exp(p(\theta)k(x) + H(x) + q(\theta))$

$Y = \sum_{i=1}^{n} k(X_i)$ S.S (2)

$E(Y) = -n\frac{q'(\theta)}{p'(\theta)}$

$Var(Y) = \frac{n}{p'(\theta)^3}[p''(\theta)q'(\theta) - q''(\theta)p'(\theta)]$

If $Var(Y) = \frac{1}{I(\theta)}$ then → MVB → $\frac{1}{I(\theta)}$

do not need to know the dist of Y to find MVB.

Application in test → "distribution free" method

fundamental thing we define S. statistic, (central limit theorem)

if
- MLE case
- if the test is about param in distr.

such func of S.S if usually a estimator for that parameter, Since MLE is the best, and also a func of S.S ⇒ $ MLE

dist of the

other func of S.S.

dist of that

key in define threshold → concerned with α, β

the dist.

to test some property about

not to test a param of a dist

case1

① test H_0: a uniform dist

② test H_0: a poisson dist

I_m^2 test $U = \sum_{i=1}^{k} \frac{(f_i - nP_i)^2}{nP_i} \xrightarrow{d} \chi^2_{k-p-1}$

case2 ① test independency

r×c contingency table

$\chi^2_{(r-1)(c-1)}$

case3

case3

Equality of medians

χ^2 test

comparative inference

inference on median

① sign test (about the value of median)

a' $Z = freq(X<m)$

$Z \sim bin(n, 0.5)$

b' $Z \sim N(\frac{n}{2}, \frac{n}{4})$

based on the confidence interval of Z

② ~~Rank test~~ & Rank test

$W_1 \overset{d}{=} N(\frac{1}{2}n_1(n_1+n_2+1), \frac{1}{12}n_1n_2(n_1+n_2+1))$

$W = \sum_{i=1}^{k} (n-1) S_i^2$, $S_i^2 = \frac{1}{n-1}\sum (X_i - \bar{X}_i)^2 \Rightarrow$ $W = \sum_{i=1}^{k} \cdot \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2$

$B = n(k-1) S_B^2$, $S_B^2 = \frac{1}{k-1}\sum (\bar{X}_i - \bar{\bar{X}})^2 \Rightarrow$ $B = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{\bar{X}})^2$

(3)

estimator for $S^2 = \frac{W}{(n-1)\cdot k}$

head to estimate parameter $(r-1)\times(c-1)$

test $\mu_i = \mu_j$

$t = \frac{\bar{\mu}_i - \bar{\mu}_j}{S\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$ compare to $t_{(n-1)\cdot k}$

$\frac{\mu_i - \mu_j}{\cdots} \sim t_{(n-1)k}$

test $\mu_0 = \mu_1 \cdots \mu_k$

Test stat $= \frac{B/k-1}{W/N-k} \sim F_{k-1,\, N-k}$

test μ_i CI

$\frac{\mu_i - \hat{\mu}_i}{S\sqrt{\frac{1}{n_i}}} \sim t_{(n-1)k}$

$\mu_i \in \hat{\mu}_i \pm \left(t_{N-k}^{0.975}\sqrt{\frac{S^2}{n_i}}\right)$

$t_{N-k}^{0.975}\sqrt{\frac{S^2}{n_i}}$

χ^2 test

χ^2_{k-p-1} test $\sum_{i=1}^{k} \frac{(obs - nP_i)^2}{nP_i}$

$\chi^2_{(r-1)(c-1)}$ independency test (contingency table) $r \times c$

χ^2 test (2×2 con. table) Equality of median

rank test Equality of median

$W_1 \overset{d}{=} N\left(\frac{1}{2}(n_1+n_2+1)n_1 , \frac{1}{12} n_1 n_2 (n_1+n_2+1)\right)$

sign test

$Z = \text{freq}(x \le m)$

Under H_0: $Z \sim \text{bin}(n, 0.5)$

$Z \notin C_{0.95}\ \text{bin}(n, 0.5)$

or $Z \notin C_{0.95}\ N\left(\frac{n}{2}, \frac{n}{4}\right)$

test whether true or not

or $P = P(Z \ge obs) < 0.05$ reject H_0

If $E(u(z)) = 0$

proof $\int_{-\infty}^{+\infty} u(z)\cdot f(z)\,dz = 0$

$\Rightarrow$ **$u(z) = 0$**

if $f(z) = \theta$

$\Rightarrow \theta \cdot \int_z u(z) = 0$

: if $u(z)$ intergrate to 0

then, $u(z)$ can be

non-zero

completeness

$z(\vec{x}) = k_1(\underbrace{\vec{x}}_{t(\vec{x})} | \theta)\cdot k_2(\vec{x})$ **S.S**

$t(\vec{x})$ S.S

$f(x|\theta) = \exp(p(\theta)\cdot k(x) + H(x) + q(\theta))$

$Y = \sum_{i=1}^{n} k(x_i)$ S.S

$E(Y) = -n\frac{q'(\theta)}{p'(\theta)}$: if $E(Y) = n\theta \Rightarrow$ MVB

$Var(Y) = \frac{n}{p'(\theta)^3}[p''(\theta)q'(\theta) - q''(\theta)p'(\theta)]$

Seldom used

for complete S.S Y, **MLE** is a func of Y

Sign test

H_0 $m = 3$

H_1 $m > 3$

Let $Z = \text{freq}(X > 3)$

Under H_0, $Z \overset{d}{=} \text{Bin}(77, 0.5)$

$z = 45$ (observed)

$\Rightarrow p = 2 \times P(Z \geq 45)$

If $p < 0.05 \Rightarrow$ reject H_0

Comparative Hypo.

① observations are independent

② Obs are normally dist with same var

③ mean differs by group

$f_X(x) = e^{-x}$

$F_X(x) = \int_0^x e^{-x}\,dx$

$= -e^{-x}\Big|_0^x$

$= 1 - e^{-x}$

④

MC

① Simulate observation from $F_Y(y)$

given $f_X(x) = \frac{1}{\pi\sqrt{y(1-y)}}$ $\Rightarrow F_X(x) = \frac{2}{\pi}\arcsin(\sqrt{y})$ $\Rightarrow y = \sin^2\frac{\pi}{2}u$, $u \in [0,1]$

② how to use a-r algo

if $U \leq \frac{f_X(Y)}{M \cdot f_Y(Y)}$, accept Y. $U \in [0,1]$. if rejected, start again.

③ $M = \max_x \frac{f_X(x)}{f_Y(x)} = \max e.x. \frac{\pi}{1.4}\sin x$ or $h(x) = \frac{f_X(x)}{f_Y(x)} \Rightarrow h'(x) = 0 \Rightarrow h(x^*) = M$

④ probability of being accepted

$P\left(U \leq \frac{f_X(Y)}{M \cdot f_Y(Y)}\right)$

$= \int P\left(U \leq \frac{f_X(Y)}{M \cdot f_Y(Y)}\right) \cdot f_Y(y)\,dy$

$= \int \frac{f_X(Y)}{M \cdot f_Y(Y)} \cdot f_Y(y)\,dy$

$= \frac{1}{M}$

Simulate example

for each simulated y, allow a 50/50 chance of setting it to be $-y$ or y.

Multiple comparisons problem (5)

$H_0 : \mu_1 = \mu_2 = \dots = \mu_n$

rewrite this as a joint hypothesis of pairwise hypothesis

① individually a size of 0.05

② joint size of 0.05

Bonferroni's correction:

$\alpha_E = 1 - (1 - \alpha_c)^m$

α_E experimental — for single comparison

reasonable approximate

$\alpha_c = \frac{\alpha_E}{m}$

⇒ the actual α would be much less than α_E

Tukey method.

1. standardised range distribution

$Z_1, \dots Z_k \sim N(\mu, \sigma^2)$

$R_k = Z_{(k)} - Z_{(1)}$

S_ν^2 is an estimator of σ^2

and $\frac{\nu S^2}{\sigma^2} \overset{d}{=} \chi^2_\nu$

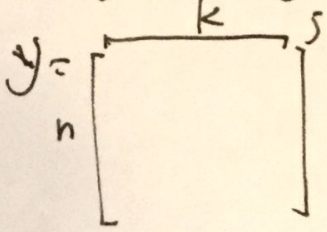
Set $Q_{k,\nu} = \frac{R_k}{S_\nu}$ ⇒ standardised rank distribution with k and ν degree of freedom.

2. Tukey CI

$\bar{X}_i - \bar{X}_j \pm C_{0.95}(Q_{k,N-k}) \frac{S}{\sqrt{n}}$

Model interpretation

$y_{ij} = \mu_j + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$



one-way

(fit a different mean for each diff kategorical factor) / level

μ_j

~~Additive models~~

if more than one variable

Disaggression

additive mode

each level of factor simply adds something to it's expected value

textbook contrast

let $\alpha_i = \mu_i - \mu$, $\beta_j = \mu_j - \mu$ (6)

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

"simply add"

where $\sum_{i=1}^{a} \alpha_i = 0$ $\sum_{j=1}^{b} \beta_j = 0$

y: k × k' (in total k'·k group)

$T = \sum\sum(X_{ij} - \bar{X}_{..})^2 = T = \sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{X}_{i.} - \bar{X}_{..})^2 +$ BR (between the Row)

$\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{X}_{.j} - \bar{X}_{..})^2 +$ BC (between the column)

$\sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij} - X_{i.} - X_{.j} + \bar{X}_{..})$ Within the cell

$k' = 2$, $k = 2$

$y = \begin{matrix} & F & M \\ Am & & \\ Aus & & \end{matrix}$

Hypothesis test

- Does 'row factor' has effect
- — — column — — —

Total SS = R + C + W

B row Bcol W within the cell

actually, "intuitively"

B = R + C

Interaction term

T = R + C + I + W (B = R + C + I)

Hypothesis testing: F-test

Estimation: MLE, row average, column average

$T = \sum\sum\sum(X_{ijk} - \bar{X}_{...})^2$

$= bc\sum_{i=1}^{a}(\bar{X}_{i..} - \bar{X}_{...}) + ac\sum_{j=1}^{b}(\bar{X}_{.j.} - \bar{X}_{...})^2 + c\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{X}_{ij.} - (\bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{..})^2 + \sum\sum\sum(X_{ijk} - \bar{X}_{ij.})^2$

The problem $E(Y|X) = \mu(x)$

Linear regression: $E(Y|x) = \alpha + \beta x$

assume $Var(Y_i) = \sigma^2$

estimate what the value of y would be given different value of X. (have some connect with I (previous))

$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = \min \Rightarrow$ MLE for α and β

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

$\hat{\alpha} = \bar{y} - \beta\bar{x}$

$\hat{\alpha}_0 = \bar{y}$, ~~$E(\hat{\alpha}_0) = \bar{y}$, $Var = \frac{\sigma^2}{n}$~~

~~$E(\hat{\beta}) =$~~

$E(y_i) = \alpha_0 + \beta(x - \bar{x})$

$\alpha_0 = \alpha + \beta \bar{x}$ $\quad u_i = x_i - \bar{x}$

$E(Y_i) = \alpha_0 + \beta u_i$

$\hat{A}_0$ be the estimator of α_0, $\hat{B}$ for β.

$E(\hat{A}_0) = \alpha_0$

$\mathrm{Var}(\hat{A}_0) = \frac{\sigma^2}{n}$

$E(\hat{B}) = \beta$

$\mathrm{Var}(\hat{B}) = \frac{\sigma^2}{\sum_{i=1}^n u_i^2}$

$\mathrm{cov}(\hat{A}_0, \hat{B}) = 0$

$\hat{A}$ to be the estimator of α

$\hat{B}$ for β (x as constant not a variable)

$E(\hat{A}) = \alpha$

$\mathrm{Var}(\hat{A}) = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n u_i^2}\right)\sigma^2$

$\mathrm{Cov}(\hat{A}, \hat{B}) = -\frac{\bar{x}}{\sum_{i=1}^n u^2}\sigma^2$

Let $\hat{m}(x)$ be the estimator for $\mu(x)$

$\Rightarrow \hat{m}(x) = \hat{A}_0 + (x - \bar{x})\hat{B}$

$\hat{A} = \hat{A}_0 - \bar{x}\cdot\hat{B}$

$\sum_{i=1}^n (y_i - \alpha_0 - \beta u_i)^2$

$= \sum_{i=1}^n (y_i - \hat{a}_0 - \hat{\beta} u_i)^2 + n(\hat{a}_0 - a_0)^2 + (\hat{\beta} - \beta)^2 \sum_{i=1}^n u_i^2$

$\Rightarrow \sum_{i=1}^n (y_i - \bar{y})^2 = \hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2 + d^2$

F test

(7)

$E(\hat{m}(x)) = \mu(x)$

$\mathrm{Var}(\hat{m}(x)) = \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n u_i^2}\right)\sigma^2$

$d^2 = \sum_{i=1}^n (y_i - \hat{a}_0 - \hat{\beta} u_i)^2$

$D^2 = \sum_{i=1}^n (Y_i - \hat{A}_0 - \hat{B} u_i)^2$

$E(D^2) = (n-2)\sigma^2$

$S^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{A}_0 - \hat{B} u_i)^2$

unbiased estimator for σ^2

$\hat{A}_0 \overset{d}{=} N(a_0, \frac{\sigma^2}{n})$, $\hat{B} \overset{d}{=} N(\beta, \frac{\sigma^2}{\sum u^2})$

$\frac{D^2}{\sigma^2} \overset{d}{=} \chi^2_{n-2}$

$\Rightarrow \frac{(n-2)S^2}{\sigma^2} \overset{d}{=} \chi^2_{n-2}$

$\Rightarrow \frac{\hat{A}_0 - a_0}{S/\sqrt{n}} \overset{d}{=} t_{n-2}$ $\quad \frac{\hat{B} - \beta}{S/\sqrt{\sum u^2}} \overset{d}{=} t_{n-2}$

Intuitive, the correlation of X, Y has to do with β, if $\beta = 0$, $\Rightarrow$ independent

Sample correlation

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Recall: $\hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (X_i - \bar{x})^2}$

$$\beta = \rho \cdot \underbrace{\frac{\sigma_Y}{\sigma_X}}_{\text{constant}}$$

So, the hypotheses

$H_0: \rho = 0 \iff H_0: \beta = 0$

test $\beta = 0$ by test ρ (easier)

Since: $\sum_{i=1}^n (y_i - \bar{y})^2 = \hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2 + d^2$

$$\beta = \rho \cdot \frac{\sigma_Y}{\sigma_X}$$

$$\Rightarrow \underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{n-1} = \underbrace{D^2}_{n-2} + R^2 \underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{1}$$

$$F = \frac{\text{regression MS}}{\text{residual MS}} = \frac{R^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}{D^2/(n-2)} = \frac{(n-2)R^2}{1-R^2} \overset{d}{=} F_{1,n-2} \quad (t_{n-2}^2 = F_{1,n-2})$$

$\Rightarrow$ test based on $\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \overset{d}{=} t_{n-2}$

The General Linear model

$$\underset{\sim}{Y} = A\underset{\sim}{\theta} + E$$

It is assumed that

$E(\underset{\sim}{E}) = 0$, $D(\underset{\sim}{E}) = \sigma^2 I$

$\underset{\sim}{Y} \overset{d}{=} N_n(A\underset{\sim}{\theta}, \sigma^2 I)$

↳ multivariate normal

$\Rightarrow y_i = A_i\underset{\sim}{\theta} + \epsilon_i$

Goal is to estimate $\underset{\sim}{\theta}$. method of LS

$\min (\underset{\sim}{y} - A\underset{\sim}{\theta})^T (y - A\theta)$

$$\Rightarrow \hat{\underset{\sim}{\theta}} = (A^TA)^{-1}A^T\underset{\sim}{y}$$

$D(\hat{\underset{\sim}{\theta}}) = \sigma^2(A^TA)^{-1}$ find estimator of σ^2

denote $d^2 = (y - A\hat{\underset{\sim}{\theta}})^T(y - A\hat{\underset{\sim}{\theta}})$ $\Rightarrow$ error

$$(\underset{\sim}{y} - A\underset{\sim}{t})^T(y - A\underset{\sim}{t}) = (\underset{\sim}{y} - A\hat{\underset{\sim}{\theta}})^T(\underset{\sim}{y} - A\hat{\underset{\sim}{\theta}}) + (\hat{\underset{\sim}{\theta}} - \underset{\sim}{t})^TA^TA(\hat{\underset{\sim}{\theta}} - \underset{\sim}{t})$$

$$= d^2 + (\hat{\underset{\sim}{\theta}} - \underset{\sim}{t})^TA^TA(\hat{\underset{\sim}{\theta}} - \underset{\sim}{t})$$

$E(D^2) = E[(\underset{\sim}{y} - A\tilde{\theta})^T(\underset{\sim}{Y} - A\tilde{\theta})] = (n-p)\sigma^2$

$\underset{\sim}{S^2 = \frac{D^2}{n-p}}$ (unbiased estimator for σ^2)

H_0, $t = 0$ (under H_0)

$$\Rightarrow y^T \cdot y = d^2 + \underset{\sim}{\hat{\theta}}^T A^T A \underset{\sim}{\hat{\theta}}$$

$$\boxed{A^T \underset{\sim}{y} = A^T (A \underset{\sim}{\hat{\theta}} + \underset{\sim}{\hat{e}}) = A^T A \underset{\sim}{\hat{\theta}}}$$

($A^T \hat{e} = 0$ by orthogonality)

$$\Rightarrow y^T y = d^2 + \underset{\sim}{\hat{\theta}}^T A^T y$$

$$\boxed{\begin{aligned} d^2 &= (y - A\underset{\sim}{\hat{\theta}})^T (y - A\hat{\theta}) \\ &= y^T y - \underset{\sim}{\hat{\theta}}^T A^T y \end{aligned}}$$

$$E(D^2) = (n-p)\,\sigma^2$$

$$\boxed{S^2 = \frac{1}{n-p}\left(\underset{\sim}{y}^T \underset{\sim}{y} - \underset{\sim}{\hat{\theta}}^T A^T \underset{\sim}{y}\right) = \underbrace{\frac{D^2}{n-p}}_{\text{rank of } A^T A}}$$

$$D^2 = S^2 (n-p)$$

$$\boxed{\frac{D^2}{\sigma^2} = \frac{(n-p)S^2}{\sigma^2} \overset{d}{=} \chi^2_{n-p}}$$

* Application of the general linear model:
  - One-way and Two way ANOVA
  - Multiple linear regression