Example: Poisson Distribution

suppose: $X_1, \cdots X_n \overset{iid}{\sim}$ Poisson (λ)

$P(X_i = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ Question: How to construct CI for λ.

We need to find a point estimate $\hat{\lambda} = \bar{X}$ / unbiased $E\bar{X} = EX_1 = \lambda$

$\sqrt{n}(\bar{X} - \lambda) \xrightarrow{D} N(0, \frac{Var(X_1)}{\lambda})$

Therefore: $\frac{\sqrt{n}(\bar{X}-\lambda)}{\sqrt{\lambda}} \xrightarrow[n\to\infty]{D} N(0,1)$

↳ Asymptotic pivotal.

CI 0.95 CI ± 1.96

Two ways: $\bar{X} \pm 1.96\frac{\sqrt{\lambda}}{\sqrt{n}}$ $\bar{X}$

Variance stabblization transformation.

Suppose θ is an unknown parameter and $\hat{\theta}$ is an estimate of θ

In many situations, we have the CLT: $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, \sigma^2(\theta))$

Example 1: Poisson $\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{D} N(0, \lambda)$

Example 2: If $X_1 \cdots X_n \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$ $\sqrt{n}(\bar{x} - p) \xrightarrow{D} N(0, p(1-p))$

General idea: to find a function g such that

$\sqrt{n}\{g(\hat{\theta}) - g(\theta)\} \xrightarrow{D} N(0, S^2)$ is independent on θ.

How to construct g? By taylor's expansion

$g(\hat{\theta}) - g(\theta) \simeq (\hat{\theta} - \theta)g'(\theta) +$ higher-oder terms.

$\Rightarrow Var(g(\hat{\theta}) - g(\theta)) \asymp g'^2(\theta)(Var(\hat{\theta} - \theta)) \simeq g'^2(\theta)\cdot\frac{\sigma^2(\theta)}{n}$

hence we require $g'^2(\theta)\sigma^2(\theta) =$ const.

$\Rightarrow Var(g(\hat{\theta}) - g(\theta)) =$ const.

$g'(\theta)\sigma(\theta) = \text{const} \Rightarrow g'(\theta) = \frac{\text{const}}{\sigma(\theta)}$

$\Rightarrow g(\theta) = \text{const}\int\frac{1}{\sigma(\theta)}d\theta$

Example 1 (revisted) $\sigma(\lambda)=\sqrt{\lambda}$ then the transformation

$$g(\lambda) = \int \frac{1}{\sqrt{\lambda}} d\lambda = 2\lambda^{\frac{1}{2}} \qquad \text{choose } g(\lambda) = \lambda^{\frac{1}{2}}$$

Then $\sqrt{n}\,(\sqrt{\bar{X}} - \sqrt{\lambda}) \xrightarrow{D} N(0, g'^2(\lambda)\,\sigma^2(\lambda))$

$$= N(0, (\tfrac{1}{2}\lambda^{-\frac{1}{2}})^2 \lambda)$$

$$= N(0, \tfrac{1}{4})$$

Hence, the asymptotic Variance is a constant $\frac{1}{4}$, Then we can use this fact to construct CI: $\sqrt{n}\,(\sqrt{\bar{X}} - \sqrt{\lambda}) \xrightarrow{D} N(0, \frac{1}{4})$ cut off $\pm \frac{1.96}{2}$

$$-\frac{1.96}{2} \le \sqrt{n}\,(\sqrt{\bar{X}} - \sqrt{\lambda}) \le \frac{1.96}{2}$$

$$\Rightarrow \sqrt{\bar{X}} - \frac{1.96}{2\sqrt{n}} \le \sqrt{\lambda} \le \frac{1.96}{2\sqrt{n}} + \sqrt{\bar{X}}$$

Example 2: (revisited) We need to find $g(\cdot)$ such that

$\sqrt{n}\,\{ g(\hat{p}) - g(p) \} \xrightarrow{D} N(0, \sigma^2)$ doesn't depend on p.

$g'(p)\,\sigma(p) = g'(p)\sqrt{p(1-p)} = 1$ let const be 1.

$$g'(p) = \int \frac{1}{\sqrt{p(1-p)}}\, dp = \int \frac{2\sin\theta\cos\theta}{\sqrt{\sin^2\theta\cos^2\theta}}\, d\theta = \int 2\, d\theta = 2\theta = 2\arcsin\sqrt{p}$$

$(p = \sin^2\theta)$

Example (Artificial)

	# car accidents
2006	600
2007	540

Hypothesis: is the drop in # of car accidents significant?

2006 Poisson (λ_1)

2007 Poisson (λ_2)

Null Hypothesis: $\lambda_1 = \lambda_2$ vs. $H_A = \lambda_1 \neq \lambda_2$

$\sqrt{n}\,(\underset{\substack{\uparrow \\ 2006}}{\sqrt{\bar{X}}} - \sqrt{\lambda_1}) \xrightarrow{D} N(0, \frac{1}{4})$ $\qquad H_0: \sqrt{\lambda_1} = \sqrt{\lambda_2}$

$\qquad H_A: \sqrt{\lambda_1} \neq \sqrt{\lambda_2}$

$\sqrt{n}\,(\sqrt{\bar{Y}} - \sqrt{\lambda_2}) \xrightarrow{D} N(0, \frac{1}{4})$

Under H_0: $\sqrt{n}\,(\sqrt{\bar{X}} - \sqrt{\bar{Y}}) \xrightarrow{D} N(0, \frac{1}{2})$.

$\sqrt{n\bar{X}} = \sqrt{600}$

$\sqrt{n\bar{Y}} = \sqrt{540}$

Compare $\underset{\substack{\| \\ 1.266}}{\left| \sqrt{600} - \sqrt{540} \right|}$ with $\underset{1.4}{\frac{1.96}{\sqrt{2}}}$. $\quad 1.266 < 1.4$

Therefore: the drop is not significant.

Example 3: Exponential r.vs $X_1 \cdots X_n \sim \exp(\lambda)$ $\quad \lambda e^{-\lambda x},\ x>0$.

How to estimate λ? $\quad EX_1 = \frac{1}{\lambda} = E\bar{X}$

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

In this case: $\sqrt{n}\left(\frac{1}{\bar{X}} - \lambda\right) \xrightarrow{D} N(0, \sigma^2(\lambda))$

$$\sqrt{n}\left(\frac{1-\lambda\bar{X}}{\bar{X}}\right) = -\lambda\sqrt{n}\,\frac{(\bar{X}-\frac{1}{\lambda})}{\bar{X}}$$

By CLT: $\sqrt{n}(\bar{X}-\frac{1}{\lambda}) \xrightarrow{D} N(0, \frac{1}{\lambda^2})$

$$\approx \frac{\lambda}{\frac{1}{\lambda}} N(0, \tfrac{1}{\lambda^2}) = \lambda^2 N(0, \tfrac{1}{\lambda^2}) = N(0, \lambda^2)$$

let $\bar{X} \approx \frac{1}{\lambda} \to \frac{1}{\lambda}$

We need to find $\underbrace{g \text{ such that}}$ $\quad g'(\lambda)\cdot\lambda = \text{const}$. $\quad g'^2(\lambda)\lambda^2$

$$\Rightarrow g(\lambda) = \ln\lambda \Rightarrow \sqrt{n}\left(\ln\left(\tfrac{1}{\bar{X}}\right) - \ln(\lambda)\right) \xrightarrow{D} N(0, 1)$$

Change of variables: suppose $\binom{X_1}{X_2}$ is a bivariate random vector

We need to find properties of $Y_1 = g(X_1, X_2)$

$Y_2 = h(X_1, X_2)$

How to obtain distribution of $\binom{Y_1}{Y_2}$

Example 1: Assume that $X_1 \sim N(0, \sigma_1^2)$ & $X_2 \sim N(0, \sigma_2^2)$

indep

$$g(X_1, X_2) = a_{11}X_1 + a_{12}X_2$$

$$h(X_1, X_2) = a_{21}X_1 + a_{22}X_2$$

$$\Rightarrow \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \underbrace{\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}}_{A} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \text{ is a bivariate normal.}$$

$$\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, A\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}A^T\right)$$

Generally we need to introduce Jacobian matrix to obtain the joint density of $\binom{Y_1}{Y_2}$

Recall in the one-dimensional case $Y = g(X)$ $\quad X \sim f_X$

$f_Y(y)dy \qquad f_X(x)dx$

$\frac{dy}{dx} = g'(x) \Rightarrow f_Y(y) = f_X(x) \frac{dx}{dy} = \frac{f_X(x)}{g'(x)}$ where $x = g^{-1}(y)$

Example : If $g(x) = x^3$

$$f_Y(y) = \frac{f_X(x)}{3x^2} = \frac{f_X(\sqrt[3]{y})}{3(\sqrt[3]{y})^2}$$

Example : ① If $g(x) = x^2$ $\quad x = \pm\sqrt{y}$ Then density for $Y = x^2$

$$\sum_{x: g(x)=y} \left| \frac{f_X(x)}{g'(x)} \right| = \left| \frac{f_X(\sqrt{y})}{2\sqrt{y}} \right| + \left| \frac{f_X(-\sqrt{y})}{-2\sqrt{y}} \right| = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}$$

χ^2_1

② $y = \sin x$ The density for $Y = \sin x$

$$= \sum_{x: \sin x = y} \left| \frac{f_X(x)}{\cos x} \right|$$

x: $\arcsin(y) + 2k\pi$

$\pi - \arcsin(y) + 2k\pi$

Change of variables.

basic setup: suppose we know the random vector (X_1, X_2) has the density function of $f_{X_1,X_2}(x_1,x_2)$, We want to find the density function of (Y_1, Y_2) where $Y_1 = g(X_1, X_2)$ $Y_2 = h(X_1, X_2)$

Recall in the one-dimensional case $Y = g(X)$, then the density for Y can be computed in the following way $f_Y(y)\,dy = f_X(x)\,dx$.

$$f_Y(y) = f_X(x)\frac{dx}{dy} = \frac{f_X(x)}{g'(x)} \quad \text{where } x \text{ solves } g(x) = y.$$

In the bivariate case, we can do something similar

$$1 = \iint f_{Y_1Y_2}(y_1, y_2)\,dy_1\,dy_2 \qquad \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} g(x_1, x_2) \\ h(x_1, x_2) \end{pmatrix}$$

$$1 = \iint f_{X_1X_2}(x_1, x_2)\,dx_1\,dx_2$$

A simple way to identify $f_{Y_1,Y_2}(y_1,y_2)$ is via $f_{Y_1Y_2}(y_1,y_2)\,dy_1\,dy_2$ $= f_{X_1,X_2}(x_1,x_2)\,dx_1\,dx_2$

$$\Rightarrow \quad f_{Y_1Y_2}(y_1,y_2)\,\frac{dy_1\,dy_2}{dx_1\,dx_2} = f_{X_1,X_2}(x_1,x_2)$$

$$\hookrightarrow \left|\frac{\partial(y_1,y_2)}{\partial(x_1,x_2)}\right| = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{vmatrix} = \frac{\partial g}{\partial x_1}\frac{\partial h}{\partial x_2} - \frac{\partial g}{\partial x_2}\frac{\partial h}{\partial x_1}$$

Hence the joint density: $$f_{Y_1Y_2}(y_1,y_2) = \frac{f_{X_1X_2}(x_1,x_2)}{\left|\frac{\partial g}{\partial x_1}\frac{\partial h}{\partial x_2} - \frac{\partial g}{\partial x_2}\frac{\partial h}{\partial x_1}\right|}$$

where (x_1, x_2) solves $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} g(x_1,x_2) \\ h(x_1,x_2) \end{pmatrix}$

Example 1: let $Y_1 = g(x_1, x_2) = x_1 + x_2$
$Y_2 = x_2 = h(x_1, x_2)$ suppose $(x_1, x_2) \sim f_{X_1X_2}(x_1 x_2)$

How to compute the joint density of (Y_1, Y_2)?

By (1): $\frac{\partial g}{\partial x_1}\frac{\partial h}{\partial x_2} - \frac{\partial g}{\partial x_2}\frac{\partial h}{\partial x_1} = 1$ we need to solve $(x_1 x_2)$ from $\begin{cases} y_1 = x_1 + x_2 \\ y_2 = x_2 \end{cases}$

$\Rightarrow f_{Y_1Y_2}(y_1, y_2) = f_{X_1X_2}(x_1, x_2) = f_{X_1X_2}(y_1 - y_2, y_2)$

Density for Y_1: $f_{Y_1}(y_1) = \int f_{Y_1 Y_2}(y_1, y_2)\, dy_2 = \int f_{X_1 X_2}(y_1 - y_2, y_2)\, dy_2$

In the special case that $X_1 \perp X_2$

$$= \int f_{X_1}(y_1 - y_2)\, f_{X_2}(y_2)\, dy_2$$

Convolution: $Y_1 = X_1 + X_2$

Density for χ^2 distribution:

Definition: Suppose $z_1 \cdots z_n$ are iid standard normal random variables.

then we say that the square sum $z_1^2 + \cdots + z_n^2$ χ^2-distribution with df n.

How to obtain density of χ_n^2?

$\underline{n=1}$: $Y = Z_1^2$ $\quad z_1: p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

$$f_Y(y) = \sum_{z: g(z)=y} \frac{f_Z(z)}{|\underbrace{g'(z)}_{2z}|} = \frac{f_Z(\sqrt{y})}{|2\sqrt{y}|} + \frac{f_Z(-\sqrt{y})}{|-2\sqrt{y}|} = \frac{1}{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}}$$

Generally speaking: χ_n^2 has density: $\frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$ where $\Gamma(\lambda) = \int_0^\infty u^{\lambda-1} e^{-u}\, du$

Method of Induction:

Assume it's true for $n=k$;

then we only need to show that it holds for $n=k+1$

$$\int_0^\infty u^{-\frac{1}{2}} e^{-u}\, du$$

$$u = \frac{v^2}{2}$$

$$\int \frac{\sqrt{2}}{v} e^{-\frac{v^2}{2}} v\, dv.$$

$$\chi_{k+1}^2 = \underbrace{z_1^2 + \cdots + z_k^2}_{\text{density: } \frac{1}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}} + z_{k+1}^2 \longrightarrow \frac{1}{\sqrt{2}\,\Gamma(\frac{1}{2})} x^{-\frac{1}{2}} e^{-\frac{x}{2}}$$

Therefore the density for $\chi_{k+1}^2 = \int_0^u \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} \frac{(u-x)^{-\frac{1}{2}} e^{-\frac{u-x}{2}}}{\sqrt{2}\,\Gamma(\frac{1}{2})}\, dx.$

$$\overset{?}{=} \frac{u^{\frac{k+1}{2}-1} e^{-\frac{u}{2}}}{2^{\frac{k+1}{2}} \Gamma(\frac{k+1}{2})}$$

$$B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

$$\frac{\int_0^u x^{\frac{k}{2}-1} (u-x)^{-\frac{1}{2}}\, du}{\Gamma(\frac{k}{2}) \Gamma(\frac{1}{2})} \overset{?}{=} \frac{u^{\frac{k+1}{2}-1}}{\Gamma(\frac{k+1}{2})}$$

$x = ut \qquad 0 \le t \le 1$

$$\underbrace{\int_0^1 t^{\frac{k}{2}-1} (1-t)^{-\frac{1}{2}}\, dt}_{B(\frac{k}{2}, \frac{1}{2})} \overset{?}{=} \frac{\Gamma(\frac{k}{2}) \Gamma(\frac{1}{2})}{\Gamma(\frac{k+1}{2})}$$

Remark 1: If $n=2$, then the density $\frac{1}{2\Gamma(1)} e^{-\frac{x}{2}} = \frac{e^{-\frac{x}{2}}}{2}$

$\Gamma(1)=1$

e^{-y}

* $\frac{\chi_2^2}{2} \to$ exponential (standard).

Remark 2: n the degrees of freedom doesn't need to be an integer.

Remark 3: $Y = X\beta + \varepsilon \qquad \hat{Y} = HY \qquad H = X(X^T X)^{-1} X^T$

$df = tr(I-H)$

t-distribution: suppose $Z \sim N(0,1)$ & $Y \sim \chi_n^2$

(Z, Y) are indep.

then $T = \frac{Z}{\sqrt{Y/n}}$ has t-distribution with df. n.

Density of T? $\begin{pmatrix} T \\ W \end{pmatrix} = \begin{pmatrix} \frac{Z}{\sqrt{Y/n}} \\ Z \end{pmatrix}$ need to solve (y, z) from

$$f_{T,W}(t,w) \left| \frac{\partial(t,w)}{\partial(z,y)} \right| = f_{Z,Y}(z,y).$$

$$\left| \frac{\partial(t,w)}{\partial(y,z)} \right| = \begin{vmatrix} \frac{\partial t}{\partial y} & \frac{\partial t}{\partial z} \\ \frac{\partial w}{\partial y} & \frac{\partial w}{\partial z} \end{vmatrix} = \begin{vmatrix} z\sqrt{n}\left(-\frac{1}{2}\right) y^{-\frac{3}{2}} & \frac{1}{\sqrt{y/n}} \\ 0 & 1 \end{vmatrix} = z\sqrt{n} \frac{y^{-\frac{3}{2}}}{-2}$$

$$f_{T,W}(t,w) = \frac{f_{Z,Y}(z,y)}{z\sqrt{n} \frac{y^{-\frac{3}{2}}}{2}} = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}}{z\sqrt{n} \frac{y^{-\frac{3}{2}}}{2}}$$

$$\begin{cases} z = w \\ y = n\frac{w^2}{t^2} \end{cases}$$

$$= \frac{2 \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}}}{w\sqrt{n}\, 2^{\frac{n}{2}} \Gamma(\frac{n}{2})} \left(\frac{nw^2}{t^2}\right)^{\frac{n+1}{2}} e^{-\frac{nw^2}{2t^2}}$$

$$f_T(t) = \int_{-\infty}^{\infty} f_{T,W}(t,w)\, dw = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\, \Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

Remark 1: If $n=1$ the density $= \frac{1}{\pi}(1+t^2)^{-1}$ Cauchy.

If $n \to \infty$ $f_T(t) \to \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$

Review: $\chi_n^2 = Z_1^2 + \cdots + Z_n^2$, where Z_i are iid $N(0,1)$

the density is: $\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$, $x>0$.

t-distribution $T = \frac{X}{\sqrt{Y/n}}$ where $X \sim N(0,1)$ $Y \sim \chi^2_n$ & $X \perp\!\!\!\perp Y$

For example, suppose $Z_1, \cdots Z_{n+1}$ are iid $N(0,1)$

Then $\frac{Z_1}{\sqrt{(Z_2^2+\cdots+Z_{n+1}^2)/n}} \sim t_n$

However, $\frac{Z_1}{\sqrt{(Z_1^2+\cdots+Z_n^2)/n}} \not\sim t_n$.

Review the density for t_n $\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$

What happens if $n \to \infty$

Recall Stirling's formula $m! \sim \left(\frac{m}{e}\right)^m \sqrt{2\pi m}$ $\left(\lim_{m\to\infty} \frac{m!}{\left(\frac{m}{e}\right)^m\sqrt{2\pi m}} = 1\right)$

$\Gamma(m+1) \sim \left(\frac{m}{e}\right)^m \sqrt{2\pi m}$ $(m \in \mathbb{N})$

$$\sim \frac{\left(\frac{\frac{n+1}{2}-1}{e}\right)^{\frac{n+1}{2}-1}\sqrt{\cancel{2\pi}\left(\frac{n+1}{2}-1\right)}}{\sqrt{n\pi}\left(\frac{\frac{n}{2}-1}{e}\right)^{\frac{n}{2}-1}\sqrt{\cancel{2\pi}\left(\frac{n}{2}-1\right)}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

$$\sim \frac{\left(\frac{n-1}{2e}\right)^{\frac{n}{2}-1}\left(\frac{n-1}{2e}\right)^{\frac{1}{2}}}{\sqrt{n\pi}\left(\frac{n-2}{2e}\right)^{\frac{n}{2}-1}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \sim \left(\frac{n-1}{n-2}\right)^{\frac{n}{2}-1}\frac{1}{\sqrt{2\pi e}}\underbrace{\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}}_{e^{-\frac{t^2}{2}}}$$

$$\downarrow$$

$$\left(1+\frac{1}{n-2}\right)^{\frac{n-2}{2}} \to \sqrt{e}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

F-distribution: Suppose $X \sim \chi^2_{n_1}$, $Y \sim \chi^2_{n_2}$ require that $X \perp\!\!\!\perp Y$

$F = \frac{X/n_1}{Y/n_2}$ is called F-distribution with df n_1, n_2

Properties: Density? Outline: let $\begin{cases} U = \dfrac{X/n_1}{Y/n_2} \\ V = X \end{cases}$

$$f_{U,V}(u,v)\,du\,dv = f_{X,Y}(x,y)\,dx\,dy \qquad \text{the Jacobian} = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix}$$

$$= \begin{vmatrix} \frac{1/n_1}{Y/n_1} & -\frac{x/n_1}{y^2/n_2} \\ 1 & 0 \end{vmatrix} = \frac{X/n_1}{y^2/n_2} \qquad f_U(u) = \int f_{U,V}(u,v)\,dv$$

$$= \frac{\Gamma(\frac{n_1+n_2}{2})(\frac{n_1}{n_2})^{\frac{n_1}{2}} \cdot u^{\frac{n_1}{2}-1}}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})(1+\frac{n_1}{n_2}u)^{\frac{n_1+n_2}{2}}}$$

Remark 1: If $n_1 = 1$, $F = \dfrac{X}{Y/n_2} = T_{n_2}^2$

consequently T-test, F-test equivalent.

Remark 2: what happens if $n_2 \to \infty$

$$Y = Z_1^2 + \cdots + Z_{n_2}^2 \qquad \frac{Y}{n_2} \to E(Z_1^2) = 1$$

So $F \to X/n_1 \qquad X \to \chi^2_{n_1}$

Remark 3: If $n_1 = n_2 \qquad F = \dfrac{X}{Y}$

Moment and Variance of χ^2, t, & F-distribution.

Suppose $X \sim \chi^2_n \Rightarrow EX = n \qquad \operatorname{Var} X = \sum_{j=1}^{n} \operatorname{Var} Z_j^2 = 2n$

$$\operatorname{Var} Z_j^2 = \underbrace{EZ_j^4}_{3} - (\underbrace{EZ_j^2}_{1})^2$$

By the CLT: $\dfrac{Z_1^2 + \cdots + Z_n^2 - n}{\sqrt{2n}} \to N(0,1)$

1.96.

For t_n distribution: $E(T) = \infty$ if $n=1$ (cauchy)

$T = \dfrac{X}{\sqrt{Y/n}}$ $\qquad E(T) = 0$ if $n \geq 2$

Variance? $E(T^2) = E\left(\dfrac{X^2}{Y/n}\right) = \underbrace{E(X^2)}_{1} E\left(\dfrac{1}{Y/n}\right)$ independence

$$= n\,E\left(\frac{1}{Y}\right) = n\int_0^\infty \frac{1}{y} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} e^{-\frac{y}{2}} y^{\frac{n}{2}-1}\,dy$$

$$= \frac{n}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \int_0^\infty e^{-\frac{y}{2}} y^{\frac{n}{2}-2}\,dy \qquad y = 2u$$

$$= \frac{n}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \cdot 2^{\frac{n}{2}-2+1} \overbrace{\int_0^\infty e^{-u} u^{\frac{n}{2}-2}\,du}^{\Gamma(\frac{n}{2}-1)}$$

Recall $\Gamma(a+1) = a\Gamma(a)$.

$$= \frac{n \cdot 2^{\frac{n}{2}-1}\cancel{\Gamma(\frac{n}{2}-1)}}{2^{\frac{n}{2}}(\frac{n}{2}-1)\cancel{\Gamma(\frac{n}{2}-1)}} = \frac{n}{n-2} = \text{Variance of T-distribution} \begin{cases} \infty & \text{if } n=2 \\ <\infty & \text{if } n>2 \end{cases}$$

Expectation of $F = \dfrac{X/n_1}{Y/n_2}$ $\quad E(F) = E(X/n_1)\, E\left(\dfrac{1}{Y/n_2}\right) = \dfrac{n_2}{n_2-2}$

Bivariate / Multivariate Normal-distribution

Suppose $X_1\ X_2$ iid $N(0,1)$ $\quad \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} a_{11}X_1 + a_{12}X_2 \\ a_{21}X_1 + a_{22}X_2 \end{pmatrix} = A\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, where $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$

$$f_{Y_1,Y_2}(y_1, y_2)\, dy_1\, dy_2 = f_{X_1X_2}(x_1,x_2)\, dx_1\, dx_2$$

$$\left|\frac{\partial(y_1, y_2)}{\partial(x_1,x_2)}\right| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = |A|$$

Therefore: the joint density: $f_{Y_1Y_2}(y_1 y_2) = \dfrac{f_{X_1}(x_1) f_{X_2}(x_2)}{\det(A)} = \dfrac{\frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_2^2}{2}}}{|\det A|}$

$$= \frac{1}{2\pi|\det A|} e^{-\frac{1}{2}(x_1\ x_2)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}$$

We need to solve

$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = A\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \Rightarrow A^{-1}\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$

$$= \frac{1}{2\pi|\det(A)|} e^{-\frac{1}{2}(Y_1\ Y_2)(A^{-1})^T A^{-1}\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}}$$

$\Downarrow$

covariance matrix

$\Sigma = E\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}(Y_1\ Y_2)$

$$= \frac{1}{2\pi|\det(A)|} e^{-\frac{1}{2}(y_1\ y_2)\Sigma^{-1}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}}$$

$= A\, E\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}(X_1\ X_2)A^T = AA^T$

$\Rightarrow \Sigma^{-1} = (A^T)^{-1}A^{-1}$

Also $\det \Sigma = (\det A)(\det A^T) = (\det(A))^2$

$$= \frac{1}{2\pi\sqrt{(\det\Sigma)}} e^{-\frac{1}{2}(y_1\ y_2)\Sigma^{-1}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}}$$

Multivariate Normal: $f(y_1 \cdots y_p) = \dfrac{1}{(2\pi)^{p/2}\sqrt{\det\Sigma}} e^{-\frac{1}{2}(y_1 \cdots y_p)\Sigma^{-1}\begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}}$

where Σ is the covariance matrix of $\begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix}$

Let $\sigma_1^2 = EY_1^2 = E(a_{11}X_1 + a_{12}X_2)^2 = a_{11}^2 + a_{12}^2$

$\sigma_2^2 = E(Y_2^2) = a_{12}^2 + a_{22}^2$

$\rho = \dfrac{EY_1Y_2}{\sigma_1\sigma_2} = \dfrac{a_{11}a_{21} + a_{12}a_{22}}{\sqrt{(a_{11}^2+a_{12}^2)(a_{21}^2+a_{22}^2)}}$

The bivariate Normal: (based on correlation)

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}\left(\frac{y_1^2}{\sigma_1^2} + \frac{y_2^2}{\sigma_2^2} - 2\rho\frac{y_1\, y_2}{\sigma_1\sigma_2}\right)/1-\rho^2}$$

Same by letting $\Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\rho & 1 \end{pmatrix}$

Note that $\Sigma = AA^T = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1 & a_{11}a_{21} + a_{12}a_{22} \\ \uparrow & \sigma_2^2 \end{pmatrix}$

$a_{11}a_{21} + a_{12}a_{22}$

$= \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

Sampling distribution:

Assume that $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) \quad \text{and} \quad S_n^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)$$

Theorem: 1) $\bar{X}$ and S_n^2 are independent;

2) $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$

Therefore $\frac{\sqrt{n}(\bar{X} - \mu)}{S_n} \sim t_{n-1}$ $\quad \frac{W}{\sqrt{V/(n-1)}}$ where $W \sim N(0,1)$ & $V \sim \chi^2_{n-1}$, W, V indep.

Consider the vector

$$U = \begin{pmatrix} X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{pmatrix} = \begin{pmatrix} (1 - \frac{1}{n})X_1 - \frac{X_2 + \cdots + X_n}{n} \\ -\frac{X_1}{n} + (1 - \frac{1}{n})X_2 - \frac{X_3 + \cdots + X_n}{n} \\ \vdots \\ -\frac{X_1 + \cdots + X_{n-1}}{n} + (1 - \frac{1}{n})X_n \end{pmatrix} = A \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

U doesn't depend on μ.

where
$$A = \begin{pmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & & 1 - \frac{1}{n} \end{pmatrix}$$

U is a Gaussian random vector$_{n \times 1}$ with mean $\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$ and covariance matrix

$$\text{cov}(U) = A \,\text{cov}\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} A^T = \sigma^2 A^2$$

Why $\bar{X}$ and U are independent?

1908 William Gosset

It suffices to show that $E(\bar{X} U) = 0$.

$$E[\bar{X}(X_1 - \bar{X})] = \underbrace{E(\bar{X} X_1)}_{\frac{\sigma^2}{n}} - \underbrace{E(\bar{X}^2)}_{\frac{\sigma^2}{n}} = 0.$$

Since $S_n^2 = \frac{1}{n-1} U^T U$ is a function of U, $\bar{X}$ and S_n^2 are independent.

Why $\frac{U^T U}{\sigma^2} \sim \chi^2_{n-1}$?

$$U^T U = (X_1 \cdots X_n) A^T A \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

$$A = I - \frac{1}{n} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (1 \cdots 1)$$

Eigen-decomposition $A = Q \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{pmatrix} Q^T$ $\quad QQ^T = I$

$$U^T U = (x_1 \cdots x_n)\, Q \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{pmatrix} \underbrace{Q^T \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}}_{Z}$$

let $Z = Q^T \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$

So $U^T U = Z^T \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{pmatrix} Z = z_1^2 + \cdots + z_{n-1}^2 \sim \sigma^2 \chi^2_{n-1}$

provided that: $z_1 \cdots z_{n-1}$ are independent.

The covariance matrix of $Z = Q^T \operatorname{cov}\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} Q = \sigma^2 Q^T Q = \sigma^2 I$

One-sample t-test: $x_1 \cdots x_n \sim N(\mu, \sigma^2)$

We want to test the hypothesis ~~that~~ $H_0: \mu = \mu_0$ $\hookrightarrow$ pre-assigned level. $H_A: \mu \neq \mu_0$.

If σ^2 is known, then $Z = \dfrac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$

If $|Z| > 1.96$ then we reject H_0 at $\alpha = 0.05$.

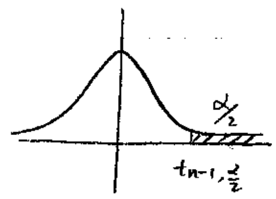
In practice, σ^2 is unknown, it can be estimated by.

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$$

So we can use $T = \dfrac{\sqrt{n}(\bar{X} - \mu_0)}{S_n} \sim t_{n-1}$

If $|T| > t_{n-1, \frac{\alpha}{2}}$

then we reject H_0 at level α



$\alpha = 0.05$ $n \to \infty$, $t_{n-1,\frac{\alpha}{2}} \to 1.96$.

We can construct confidence intervals based on t-test, $1-\alpha$ CI for μ?

$$-t_{n-1,\frac{\alpha}{2}} < \frac{\sqrt{n}(\bar{X} - \mu)}{S_n} < t_{n-1,\frac{\alpha}{2}}$$

The CI for μ is $\bar{X} \pm \dfrac{S_n}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}$

Remark: If σ^2 is known, then CI is $\bar{X} \pm \dfrac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}$ shorter

Power of t-test:

Type I error: H_0 is true, However our test procedure reject it.

Type II error: H_0 is false, However our test fails to reject it.

type II error → H_0 錯了 沒拒絕

Power of z-test: $\mu \neq \mu_0$

$$1 - P\left(\frac{\sqrt{n}\,|\bar{X} - \mu_0|}{\sigma} < 1.96\right)$$

$$= P\left(\frac{\sqrt{n}\,|\bar{X} - \mu_0|}{\sigma} \geq 1.96\right)$$

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

$= \mu + \frac{\sigma}{\sqrt{n}} z$, $z \sim N(0,1)$

$$= P\left(\frac{\sqrt{n}\,|\mu + \frac{\sigma}{\sqrt{n}} z - \mu_0|}{\sigma} \geq 1.96\right)$$

$$= 1 - \Phi\Big(1.96 - \underbrace{\frac{\sqrt{n}(\mu - \mu_0)}{\sigma}}_{\delta = 0}\Big) + \Phi\Big(-1.96 - \underbrace{\frac{\sqrt{n}(\mu - \mu_0)}{\sigma}}_{\delta = 0}\Big)$$

$$= 0.05 \quad \text{if } \delta = 0.$$

for δ is large. Power $\to 1$ large prob. to reject H_0

$H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$.

If $\mu - \mu_0$ is larger, easier to reject H_0

If n is larger, easier to reject H_0

If σ is larger, more difficult to reject H_0

For t-test, the power is the probability

$$P\left(\frac{\sqrt{n}\,|\bar{X} - \mu|}{S_n} \geq t_{n-1, \frac{\alpha}{2}}\right)$$

$$= P\left(\frac{\sqrt{n}(\mu - \mu_0)}{S_n} + \frac{T}{S_n} \leq -t_{n-1, \frac{\alpha}{2}}\right)$$

$$+ P\left(\quad - \quad - \quad - \quad > t_{n-1, \frac{\alpha}{2}}\right)$$

.iew: One sample t-test

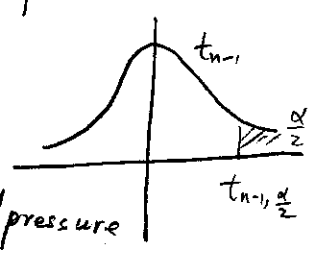
Basic set-up: suppose $X_1, X_2 \cdots X_n \sim N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Test the hypothesis: $H_0: \mu = \mu_0$ vs $H_A: \mu \neq \mu_0$

We can use $T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n}$, where $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

Under H_0: $T \sim t_{n-1}$

Reject H_0 if $|T| \geq t_{n-1, 1-\frac{\alpha}{2}}$ → $(1-\frac{\alpha}{2})^{th}$ quantile of t_{n-1}

(R) ≥ qt(0.975, df)



Variat: $H_0: \mu = \mu_0$ vs $H_A: \mu > \mu_0$

μ: bloodpressure

$H_A: \mu < \mu_0$ test whether can reduce blood pressure.

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n}$$

For one-sided test, we accept H_A if T is too large.

Intuition: under $H_A: \mu > \mu_0$ $\bar{X}_n \approx \mu$

So the numerator in T $\sqrt{n}(\bar{X}_n - \mu_0) \approx \sqrt{n}\underbrace{(\mu - \mu_0)}_{\text{positive}}$.

At level α, we accept H_A, if $T > t_{n-1, 1-\alpha}$

O/w we reject H_A, doesn't automatically mean that we accept H_0.

A more reasonal formulation is $H_0: \mu \leq \mu_0$ vs $H_A: \mu > \mu_0$

if $T > t_{n-1, 1-\alpha}$ then we accept H_A

$<$ reject H_A (accept H_0)

Remark: before applying t-test, what issues should we pay ~~attention to~~?

$X_i \sim$ Normal, iid (independence)

non Gaussian ⟶ Non-parametric (Rank-based)

$\boxed{\text{Comparing two samples (Chapter 11)}}$

I comparing paired samples. Suppose we have n pairs (X_1, Y_1) blood pressure before, blood pressure after, the same person

iid $\vdots$

(X_n, Y_n)

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right)$$

$\mu_x, \mu_y, \sigma_x, \sigma_y, \rho$

$H_0: \mu_x = \mu_y$ vs $H_A: \mu_x \neq \mu_y$

We introduce $D_i = X_i - Y_i$ $\quad \mu_D = \mu_x - \mu_y$

$H_0: \mu_D = 0$ vs $H_A: \mu_D \neq 0$

Then essentially, we are dealing with one-sample t-test.

$$T = \frac{\sqrt{n}\bar{D}}{\sqrt{\frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}}} \sim t_{n-1}$$

Suppose, $T = 2.8$, then the p-value $= P(|t_{n-1}| \geq 2.8) = 2P(t_{n-1} \geq 2.8)$

If p-value > 0.05, then accept H_0

" " $<$ " " " reject "

Similarly, we can test $H_0: \mu_D \leq 0$ vs $H_A: \mu_D > 0$.

II) comparing independent samples (with the same variance)

Suppose $X_1, X_2 \cdots X_n \overset{iid}{\sim} N(\mu_x, \sigma^2)$

$Y_1, \cdots\cdots Y_m \overset{iid}{\sim} N(\mu_y, \sigma^2)$

Also suppose $X_1, \cdots X_n$, $Y_1, \cdots Y_m$ are independent

How to test $H_0: \mu_x = \mu_y$ vs $H_A: \mu_x \neq \mu_y$

Should compare $\bar{X} - \bar{Y}$

under H_0: $\bar{X} - \bar{Y} \sim N(0, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\sigma^2(\frac{1}{n} + \frac{1}{m})}} \sim N(0, 1)$$

If σ^2 is known, we can apply z-test

In most situations, σ^2 is unknown

based on $X_1, \cdots, X_n$

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

The first estimation $\hat{\sigma}_1^2 = \frac{S_X^2}{n-1}$

The second estimate $\hat{\sigma}_2^2 = \frac{S_Y^2}{m-1}$ where $S_Y^2 = \sum_{j=1}^{m} (Y_j - \bar{Y})^2$

the 3rd estimate: $\hat{\sigma}_3 = \frac{S_X^2 + S_Y^2}{n-1+m-1} = \frac{(n-1)\hat{\sigma}_1^2 + (m-1)\hat{\sigma}_2^2}{n-1+m-1}$ $\quad E(\hat{\sigma}_3^2) = \sigma^2$

3rd $\quad T = \frac{\bar{X} - \bar{Y}}{\sqrt{\hat{\sigma}_3^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n+m-2}$

2nd $\quad \frac{\bar{X} - \bar{Y}}{\sqrt{\hat{\sigma}_2^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{m-1}$

1st $\quad \frac{\bar{X} - \bar{Y}}{\sqrt{\hat{\sigma}_1^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n-1}$

Thm: If $X_1, \cdots, X_n$, $Y_1, \cdots, Y_m$ are indep, $X_i \sim N(\mu_x, \sigma^2)$, $Y_i \sim N(\mu_y, \sigma^2)$

then $t = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n+m-2}$ $\quad S_p^2 = \frac{\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2}{n+m-2}$

pooled

Confidence Interval for $\mu_x - \mu_y$:

$$\bar{X} - \bar{Y} \pm t_{n+m-2, 1-\frac{\alpha}{2}} \cdot \sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}$$

Pooling vs Pairing: $(X_1, Y_1), \cdots, (X_n, Y_n) \sim N_2\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \sigma^2, \rho\right)$

equal variance

$Var(\bar{X} - \bar{Y})$

$= \frac{1}{n} Var(X_1 - Y_1)$

$= \frac{1}{n} [\sigma^2 + \sigma^2 - 2\sigma^2 \rho] = \frac{2\sigma^2}{n}(1-\rho)$

In the Pooling, $X_1, \cdots X_n \sim N(\mu_x, \sigma^2)$, $Y_1 \cdots Y_n \sim N(\mu_y, \sigma^2)$

$Var(\bar{X} - \bar{Y}) = 2\sigma^2/n$

Pairing : n people

Pooling : $2n$ (random)

If $\rho = \frac{3}{4}$

then $var(\bar{X} - \bar{Y}) = \frac{2\sigma^2}{n} \frac{1}{4} \ll$ pooling.

pairing is preferred $\leftrightarrow \rho > 0$

pooling perferred $\leftrightarrow \rho < 0$

Comparing 2 samples

pooling vs pairing.

suppose we have 2 samples $X_1, \cdots X_n \sim N(\mu_1, \sigma^2)$, $Y_1, \cdots Y_n \sim N(\mu_2, \sigma^2)$ every thing independent.

$$Var(\bar{X} - \bar{Y}) = 2\frac{\sigma^2}{n}$$

For pairing, $(X_1, Y_1) \cdots (X_n, Y_n) \overset{\text{pair}}{\underset{\text{iid}}{\sim}} N(\mu_1, \mu_2, \sigma^2, \sigma^2, \rho)$

$$Var(\bar{X} - \bar{Y}) = \frac{2\sigma^2}{n}(1-\rho)$$

Pearson's Correlation 1857 -1936

How to estimate ρ?

Setup, $(X_1, Y_1) \cdots (X_n, Y_n) \sim N\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var X \cdot Var Y}} = \frac{E[(X-EX)(Y-EY)]}{\sqrt{E(X-EX)^2 \, E(Y-EY)^2}}$$

$$\hat{\mu}_1 = \bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

$$\hat{\mu}_2 = \bar{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

$E(X-EX)(Y-EY)$ can be estimated by $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$

$\sigma_1^2 = Var X = E(X-EX)^2$ can — — — $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ MLE

$\sigma_2^2 = Var Y = E(Y-EY)^2$ — — — $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ MLE

Put them together we have $\hat{\rho} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$

Remark: $\hat{\sigma}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ unbiased

$\tilde{\sigma}_1^2 = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

properties of $\hat{\rho}$: $\sqrt{n}(\hat{\rho} - \rho) \xrightarrow[n\to\infty]{D} N(0, (1-\rho^2)^2)$

(CLT)

We can use variance transformation

$\sqrt{n}(g(\hat{\rho}) - g(\rho)) \xrightarrow{D} N(0, v)$ → independent

$$g'(\rho)^2(1-\rho^2)^2 = \text{const} \qquad g'(\rho)(1-\rho^2) = 1$$

$$\Rightarrow g'(\rho) = \frac{1}{1-\rho^2} \Rightarrow g(\rho) = \int_0^\rho \frac{1}{1-u^2}\,du \Rightarrow g(\rho) = \frac{1}{2}\log\frac{1+\rho}{1-\rho} = z$$

$$g(\hat\rho) - g(\rho) \sim N(0, \tfrac{1}{n}) \qquad \rho = \frac{e^{2z}}{1+e^{2z}}$$

CI: $\hat\rho = 0.6$, $n=30$

then $g(\rho)$ $\quad g(0.6) \pm \frac{1.96}{\sqrt{30}}$ $\qquad$ fisher's z-transformation.

Improved version: $g(\hat\rho) - g(\rho) \sim N(0, \frac{1}{n-3})$

Comparing 2 samples (unequal variance)

Suppose $X_1, \dots, X_n \sim N(\mu_1, \sigma_1^2)$

$Y_1 \cdots Y_m \sim N(\mu_2, \sigma_2^2)$ $\qquad \sigma_1^2 \neq \sigma_2^2$

Question: How to test $H_0: \mu_1 = \mu_2$?

We can use $\bar X - \bar Y$ $\qquad \hat\sigma_1^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar X)^2$

$Var(\bar X - \bar Y) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$ $\qquad \hat\sigma_2^2 = \frac{1}{m-1}\sum_{i=1}^m (Y_i - \bar Y)^2$

So we can propose:

$$T = \frac{\bar X - \bar Y}{\sqrt{\frac{1}{n}\hat\sigma_1^2 + \frac{1}{m}\hat\sigma_2^2}} = \frac{\bar X - \bar Y}{\sqrt{\frac{1}{n}\left(\frac{1}{n-1}\sum (X_i-\bar X)^2\right) + \frac{1}{m}\left(\frac{1}{m-1}\sum (Y_j - \bar Y)^2\right)}}$$

$$= \frac{\bar X - \bar Y}{\sqrt{\frac{(n-1)\hat\sigma_1^2}{\sigma_1^2} \times \frac{1}{(n-1)n}\sigma_1^2 + \frac{(m-1)\hat\sigma_2^2}{\sigma_2^2} \times \frac{1}{(m-1)m}\sigma_2^2}}$$

$$\frac{(n-1)\hat\sigma_1^2}{\sigma_1^2} \sim \chi^2_{n-1} \qquad \frac{(m-1)\hat\sigma_2^2}{\sigma_2^2} \sim \chi^2_{m-1}$$

The denominator $= \frac{\sigma_1^2}{n(n-1)}\chi^2_{n-1} + \frac{\sigma_2^2}{m(m-1)}\chi^2_{m-1}$ $\quad$ not χ^2-distribution.

T is NOT a t-distribution random variable.

However, T can be approximated by t-distribution with df $\nu = \left\lfloor \frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{m}\right)^2}{\frac{(S_x^2/n)^2}{n-1} + \frac{(S_y^2/m)^2}{m-1}} \right\rfloor$

where $S_x^2 = \hat\sigma_1^2$

$S_y^2 = \hat\sigma_2^2$ $\qquad \lfloor\cdot\rfloor$ floor function

$\lfloor 7.8 \rfloor = 7$

Welch B.L. 1947
The generalization of "Student's" problem when several different population are involved Biometrika 34, 28-35

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

Welch formula

Welch- Satterthwaite equation

Multiple testing:

Suppose we have 3 samples: $X_1 \cdots X_n \sim N(\mu_1, \sigma^2)$

$Y_1 \cdots Y_m \sim N(\mu_2, \sigma^2)$

$Z_1 \cdots Z_l \sim N(\mu_3, \sigma^2)$

$H_0: \mu_1 = \mu_2 = \mu_3$

Suppose we test: $H_1, \cdots, H_k$

Let the test level be $\alpha \in (0, 1)$

$p(\text{rejecting } H_i \mid H_i \text{ is true}) = \alpha$ type I error.

$p(\text{rejecting at least one of these } k \text{ hypotheses} \mid \text{all } H_i \text{ is true})$

$= 1 - p(\text{accept } H_1, H_2, \cdots, H_k)$

Assuming that these k tests are independent.

$= 1 - p(\text{accept } H_1) \cdot p(\text{accept } H_2) \cdots$

$= 1 - (1-\alpha)^k$

If $\alpha = 0.01$, $k = 100$

then $p(\text{reject at least one hypothesis} \mid \text{all true})$

$= 1 - 0.99^{100} = 0.64$

false discovery - false rejection

To remedy this problem, we need to choose α^* such that

$1 - (1-\alpha^*)^{100} = 0.01 \longrightarrow$ level want to control.

$\Rightarrow \alpha^* = 1 - 0.99^{\frac{1}{100}} = 0.0001$

Suppose all tests are based on normal distribution.

How to deal with the case in which tests are dependent?

$P(\text{rejecting at least one hypothesis} \mid \text{all true}) = P(\text{reject } H_1 \text{ OR reject } H_2 \text{ OR} \cdots \mid \text{All true})$

$\leq P(\text{reject } H_1 \mid \text{True}) + P(\text{reject } H_2 \mid \text{True}) \cdots = k\alpha$

Conclusion: To control the prob of false rejection @ level, α_0.
We need to conduct each individual test @ level $\frac{\alpha_0}{k}$.

Bonferroni Correction.

Test $\mu_1 = \mu_2 = \mu_3$

$X_1 \cdots X_n \sim N(\mu_1, \sigma^2)$

$Y_1 \cdots Y_m \sim N(\mu_2, \sigma^2)$

$Z_1 \cdots Z_l \sim N(\mu_3, \sigma^2)$

$\alpha_0 = 0.05$

$H_1 : \mu_1 = \mu_2 \quad H_2 : \mu_2 = \mu_3 \quad H_3 : \mu_3 = \mu_1$

$\frac{0.05}{3} \quad \frac{0.05}{3} \quad \frac{0.05}{3}$

If we were to use 0.05, then we would have too many false rejection

Summarizing data

Chap 10.3

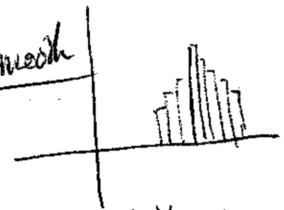
Checking normality

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

QQ plot.

- Graphical method 1. histogram — not smooth

  hist(X) — Kernel density Estimate



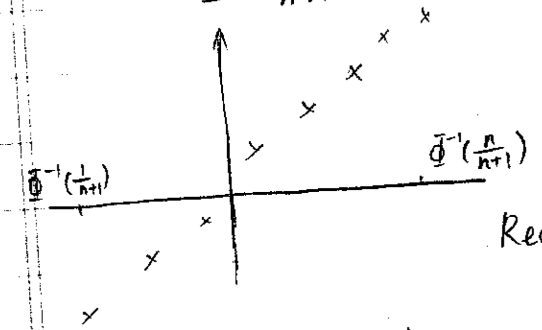
2. QQ plot.

"Comparing distribution functions"

Take the order Statistics $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$

plot the order statistics vs normal quantiles

$\Phi^{-1}\left(\frac{i}{n+1}\right), \quad i = 1, 2, \cdots n.$



If we see a line, then the normality assumption is valid.

Reason: For a standard normal random variable

$Z = \Phi^{-1}(U)$

where $U \sim \text{Unif}(0, 1)$

$$\left[\begin{aligned} \text{why: } \quad & P(Z \le z) \\ &= P(\Phi^{-1}(U) \le z) \\ &= P(U \le \Phi(z)) \\ &= \Phi(z) \end{aligned}\right]$$

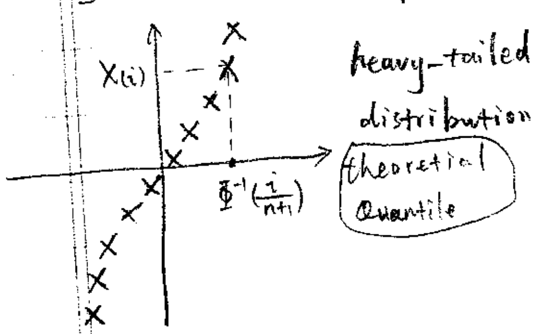
iid standard normal $\begin{cases} X_1 = \Phi^{-1}(U_1) \\ \vdots \\ X_n = \Phi^{-1}(U_n) \end{cases}$ iid standard uniforms

$X_{(1)} \le \cdots \le X_{(n)}$

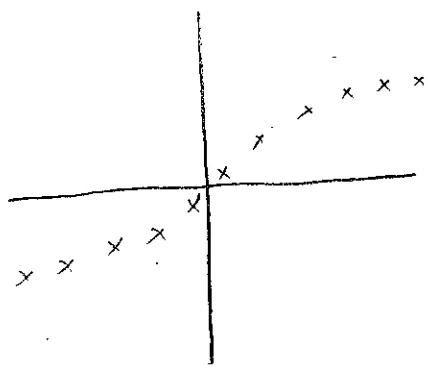
$\Phi^{-1}(U_{(1)}) \le \cdots \le \Phi^{-1}(U_{(n)})$ where $U_{(1)} \le \cdots \le U_{(n)}$ are order statistics of $U_1, \cdots, U_n$.

$U_{(k)} \approx \frac{k}{n+1}$

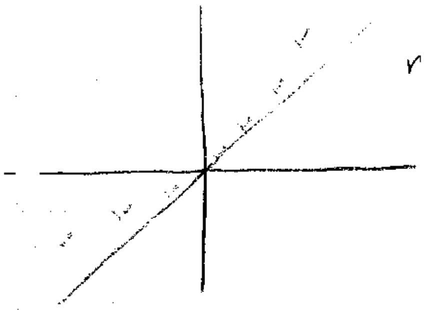
Some non-normal patterns



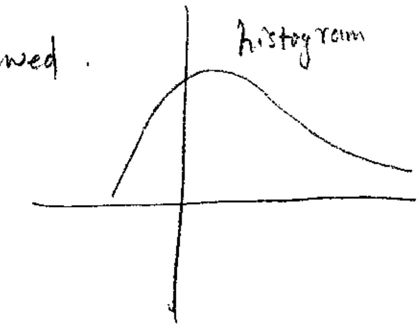
heavy-tailed distribution.



light-tailed



right-skewed.



Skewness: $\gamma_1 = \dfrac{E[(x-\mu)^3]}{\sigma^3}$, where μ is the mean; σ is the sd.

for normal distribution $\gamma_1 = 0$.

Kurtosis: $\gamma_2 = \dfrac{E[(x-\mu)^4]}{\sigma^4}$ for normal $\gamma_2 = 3$

Formal method K-Smirnov test

Suppose $X_1, \cdots, X_n$ are iid $N(0,1)$ then $F_n(x) = \frac{1}{n} \# \{ i : X_i \le x \}$

$$P\left(\sup_{x \in \mathbb{R}} \sqrt{n} \left| F_n(x) - \Phi(x) \right| \ge u \right) \longrightarrow \sum_{k \in \mathbb{Z}} e^{-(2k+1)^2 u^2}$$

find a $\underline{\underline{u}}$ s.t. $= \alpha$.

If $\sup \sqrt{n} \left| F_n(x) - \Phi(x) \right| \le u$ then accept H_0 (normality)

Measures of Location

Normal

Median.

center of a batch of numbers

Trimmed Mean

M-estimate

One-way ANOVA (chap 12) Ron. Fisher 1918

statistical methods for Research Workers

Suppose we have I groups of observations

For i^{th} group, there are J_i observations

$$y_{i1}, y_{i2}, \cdots y_{iJ_i} \sim N(\mu_i, \sigma^2)$$

We also assume independence

Want to test $H_0: \mu_1 = \cdots = \mu_I$

Notation: $\overline{y}_{i\cdot} = \frac{1}{J_i} \sum_{l=1}^{J_i} y_{il}$ (group mean), also $\overline{y}_{\cdot\cdot} = \frac{1}{J_\cdot} \sum_{i=1}^{I} \sum_{l=1}^{J_i} y_{il}$ (overall mean), where $J_\cdot = J_1 + \cdots + J_I$

$SS_T = \sum_{i=1}^{I} \sum_{l=1}^{J_i} (y_{il} - \overline{y}_{\cdot\cdot})^2$ (total sum of squares)

$SS_W = \sum_{i=1}^{I} \sum_{l=1}^{J_i} (y_{il} - \overline{y}_{i\cdot})^2$ (within group sum of squares)

$SS_B = \sum_{i=1}^{I} J_i (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2$ (between group sum of squares)

Thm $SS_T = SS_W + SS_B$

How to apply this result? $\dfrac{SS_W}{\sigma^2} \sim \chi^2_{\sum_i (J_i - 1)} = \chi^2_{J_\cdot - I}$

Assume H_0 is true then $\dfrac{SS_B/(I-1)}{SS_W/(J_\cdot - I)} \sim F_{I-1, J_\cdot - I}$

One-way ANOVA

letting we have I groups

independent:

Group 1 $Y_{11}, \cdots Y_{1J_1} \overset{iid}{\sim} N(\mu_1, \sigma^2)$

$\vdots$ same

Group I $Y_{I1} \cdots Y_{IJ_I} \overset{iid}{\sim} N(\mu_I, \sigma^2)$

To test the null hypothesis: $H_0: \mu_1 = \quad = \mu_I$

Introduce: $\underbrace{\overline{Y}_{i\cdot}}_{\text{Group mean}} = $ Ave of the ith group $\frac{1}{J_i}\sum_{l=1}^{J_i} Y_{il}$

Grand mean: $\overline{Y}_{\cdot\cdot} = \frac{1}{J_\cdot}\sum_{i=1}^{I}\sum_{l=1}^{J_i} Y_{il}$, where $J_\cdot = \sum_{i=1}^{I} J_i$

$$SS_T = \sum_{i=1}^{I}\sum_{l=1}^{J_i}(Y_{il} - \overline{Y}_{\cdot\cdot})^2$$

Theorem: $SS_T = \underbrace{SS_W}_{\sum_{i=1}^{I}\sum_{l=1}^{J_i}(Y_{il}-\overline{Y}_{i\cdot})^2} + SS_B = \sum_{i=1}^{I} J_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$

proof: $\sum_{l=1}^{J_i}(Y_{il} - \overline{Y}_{\cdot\cdot})^2 = \sum_{l=1}^{J_i}(Y_{il} - \overline{Y}_{i\cdot})^2 + \underbrace{\sum_{l=1}^{J_i}(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}_{= J_i \cdot (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2} + \underbrace{2\sum_{l=1}^{J_i}(Y_{il} - \overline{Y}_{i\cdot})(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})}_{=0}$

So $\sum_{i=1}^{I}\sum_{l=1}^{J_i}(Y_{il} - \overline{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{I}\sum_{l=1}^{J_i}(Y_{il} - \overline{Y}_{i\cdot})^2 + \sum_{i=1}^{I} J_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$

$$SS_T \quad = \quad SS_W \quad + \quad SS_B.$$

Apply of Thm to test $H_0: \mu_1 = \mu_2 = \cdots = \mu_I$

$\frac{SS_W}{\sigma^2}$ Recall that $\frac{\sum_{l=1}^{J_i}(Y_{il} - \overline{Y}_{i\cdot})^2}{\sigma^2} \sim \chi^2_{J_i - 1}$

Also, for different i, $\sum_{l=1}^{J_i}(Y_{il} - \overline{Y}_{i\cdot})^2$ are independent

$= \chi^2_{\sum_{i=1}^{I}(J_i - 1) = J_\cdot - I}$

$SS_B = \sum_{i=1}^{I} J_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$ Under $H_0: \mu_1 = \mu_2 = \cdots = \mu_I$

For simplicity, we assume $J_i = J$ (balance design)

$\bar{Y}_{i\cdot} \sim N(\mu, \frac{\sigma^2}{J})$ ⟵ (under H_0)

$\sqrt{J}\ (\bar{Y}_{i\cdot} - \mu) \sim N(0, \sigma^2)$

Notice that $\bar{Y}_{\cdot\cdot}$ is the average of $\bar{Y}_{i\cdot}$ for $i = 1, \cdots, I$

So $\frac{SS_B}{\sigma^2} \sim \chi^2_{I-1}$

$\bar{Y}_{\cdot\cdot} = \frac{1}{J_\cdot}\sum_{i=1}^{I}\sum_{l=1}^{J_i} Y_{il} = \frac{1}{J_\cdot}\sum_{i=1}^{I} J_i \bar{Y}_{i\cdot}$

weighted average of $\bar{Y}_{i\cdot}$

$J_\cdot = IJ$

Also, since $\bar{Y}_{i\cdot}$ is independent of $\sum_{l=1}^{J_i}(Y_{il} - \bar{Y}_{i\cdot})^2$

Hence, SS_B is independent of SS_W

Therefore, we can propose the test statistic $F = \frac{SS_B/(I-1)}{SS_W/(J_\cdot - I)}$

under H_0: we know that $F \sim F_{I-1, J_\cdot - I}$

We reject H_0 if $F > F_{I-1, J_\cdot - I, 1-\alpha}$ ⟶ qf(I-1, J.-I, 1-α)

What happen if H_0 is Not **True**? $SS_B/(I-1)$ will be too large.

Compute $E\frac{SS_B}{I-1}$ Just Assume $J_i = J$

$Z_i =: \bar{Y}_{i\cdot} \sim N(\mu_i, \frac{\sigma^2}{J})$

$\bar{Z} = \frac{Z_1 + \cdots + Z_I}{I}$ with mean $\frac{\mu_1 + \cdots + \mu_I}{I} = \bar{\mu}$

$SS_B = \sum_{i=1}^{I} J(Z_i - \bar{Z})^2 \qquad E(SS_B) = J\sum_{i=1}^{I} E(Z_i - \bar{Z})^2 = J\sum_{i=1}^{I}\{\underbrace{Var(Z_i - \bar{Z})}_{(1-\frac{1}{I})\frac{\sigma^2}{J}} + (\mu_i - \bar{\mu})^2\}$

$= (I-1)\sigma^2 + J\sum_{i=1}^{I}(\mu_i - \bar{\mu})^2$

$\xrightarrow{\text{So}} E\frac{SS_B}{I-1} = \sigma^2 + \frac{J}{I-1}\sum_{i=1}^{J}(\mu_i - \bar{\mu})^2$

If H_0 is not true $\sum(\mu_i - \bar{\mu})^2$ will be large

$\frac{SS_B}{I-1}$ will be large.

So we expect large $F = \frac{SS_B/(I-1)}{SS_W/(J_\cdot - I)} > F_{I-1, J_\cdot - I, 1-\alpha}$.

ANOVA Table:

Source of Variation	df	SS	MS	F-statistic	P-value
Between groups (组)	$I-1$	SS_B	$\frac{SS_B}{I-1}$	$\frac{SS_B/I-1}{SS_T/J_\cdot-1}$	$P(F>\cdot)$
Within group (errors)	$J_\cdot-I$	SS_W	$\frac{SS_W}{J_\cdot-I}$	$\frac{SS_W/(J_\cdot-I)}{SS_T/(J_\cdot-I)}$	
Total	$J_\cdot-1$	SS_T	$\frac{SS_T}{J_\cdot-1}$		

Remark: $I=2$ then F-test = T-test (equivalent)

Different Parameterization:

$Y_{il} \sim N(\mu_i, \sigma^2)$ $\quad\quad \hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$

$y_{il} = \overbrace{\underset{\downarrow}{\underline{\mu}} + \alpha_i}^{\mu_i} + \varepsilon_{il}$ where α_i statisfies $\sum_{i=1}^{I} \alpha_i = 0$.

overall level. $\quad$ α_i: difference between the ith group level and the overall level.

$y_{il} = \mu + \alpha_i + \varepsilon_{il}$ where $\alpha_1 = 0$.

We treat the first group as benchmark

α_i: differente between the i^{th} group level and the first group level.

If σ^2 not the same, Behrens-Fisher. ??

If Y_{ij} not normal. Kruskell-Wallis Checking normality

Simultaneous confidence intervals:

How to construct SCI for $\mu_i - \mu_j$ $\quad 1 \le i < j \le I \longrightarrow \binom{I}{2}$

Single CI $\quad \mu_i - \mu_j \Leftarrow \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}$

$Var(\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}) = \frac{\sigma^2}{J_i} + \frac{\sigma^2}{J_j}$

$s_p^2 = \frac{SS_W}{J_\cdot - I}$

(pool all these group together)

$$P\left(\mu_i - \mu_j \in \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm s_p \sqrt{\frac{1}{J_i} + \frac{1}{J_j}}\, t_{J_\cdot - I, \frac{\alpha}{2}}\right) = 1-\alpha$$

Studentized Range distribution

SCI: Using Bonferroni Method

$$P\left(\mu_i - \mu_j \in \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm s_p \sqrt{\frac{1}{J_i} + \frac{1}{J_j}}\, t_{J_\cdot - I, \frac{\alpha}{2}\frac{1}{\binom{I}{2}}} \text{ for all } i<j\right) \ge 1-\alpha$$

Under H_0: $F \sim F_{I-1,\, J.-I}$

$$E\, SS_B/(I-1) = \sigma^2 + \frac{J}{I-1}\sum_{i=1}^{I}(\mu_i - \bar{\mu})^2$$

Simultaneous Confidence intervals: for $\mu_i - \mu_j$ $\quad 1 \le i < j \le I$

Bonferroni-type method: $P\left(\mu_i - \mu_j \in \bar{Y}_{i.} - \bar{Y}_{j.} \pm S_p\sqrt{\frac{1}{J_i} + \frac{1}{J_j}}\; t_{J.-I,\, \frac{\alpha}{2}/\binom{I}{2}} \quad i<j\right)$ $\ge 1-\alpha$

Conservative

$$S_p^2 = \frac{SS_W}{J.-I}$$

Another way to construct SCI:

Assume $J_i = J$ (balanced design)

studentized range distribution

$$S_{I,\, IJ.-I} = \max_{1\le i<j\le I} \frac{\sqrt{J}\left|\bar{Y}_{i.} - \bar{Y}_{j.} - (\mu_i - \mu_j)\right|}{S_p}$$ is a pivotal

Suppose $z_1 \cdots z_I \overset{iid}{\sim} N(0,1)$ indepent with $\chi^2_{J.-I}$

$$S_{I,\, IJ.-I} \overset{\mathcal{D}}{=} \frac{\max_{i<j}|z_i - z_j|}{\sqrt{\chi^2_{J.-I}/(J.-I)}} \qquad S_p^2 = \frac{SS_W}{J.-I} \qquad \frac{SS_W}{\sigma^2} \sim \chi^2_{J.-I}$$

$J. = IJ$

let $z_{(1)} \le z_{(2)} \cdots \le z_{(n)}$ be the order statistics of $z_1 \cdots z_n$.

$$\max_{i<j}|z_i - z_j| = \underbrace{z_{(n)} - z_{(1)}}_{\text{range}}$$

Let $S_{I,IJ-I,1-\alpha}$ be the $(1-\alpha)^{th}$ quantile of $S_{I,IJ-I}$

Then the S.C.I. for $\mu_i - \mu_j$

$$P\left(\mu_i-\mu_j \in \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm S_p \sqrt{\frac{1}{J}+\frac{1}{J}}\, S_{I,IJ-I,(1-\alpha)} \text{ for all } i<j\right) = 1-\alpha$$

$\downarrow$ exact. no longer conservative. (bonferroni case $\geq$)

Remark:

① J_i the same

. for studentized range distribution $J_i \equiv J$

②. the length of the CI almost <u>the same</u>

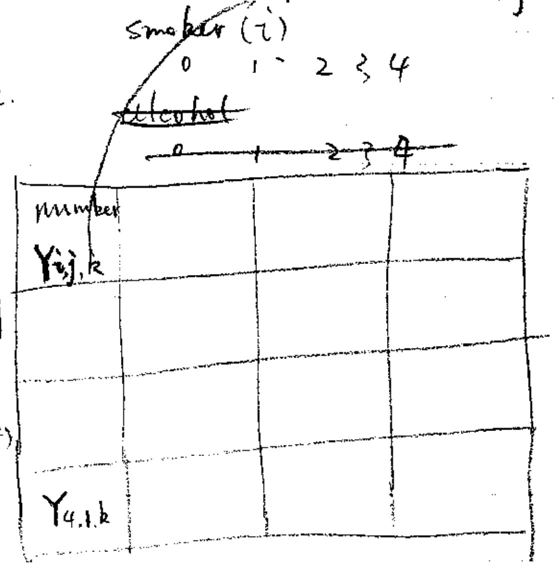
Two-way ANOVA

2 factors

heart disease.

Y_{ijk} degree of heart diseases

$k = 1, 2 \cdots k_{ij}$ number of people in this category



$Y_{ij1}, Y_{ij2} \cdots Y_{ijk_{ij}} \overset{iid}{\sim} N(\mu_{ij}, \sigma^2)$

$i = 1, 2, \cdots I$

$j = 1, 2, \cdots J$

$Y_{ijk} = \mu_{ij} + \underset{N(0,\sigma^2)}{\overset{iid}{\varepsilon_{ijk}}}$

If $k_{ij} \equiv k \Rightarrow$ balanced design.

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \delta_{ij} \longrightarrow \text{interaction effect.}$$

μ: grand mean (overall)

α_i: effect of first factor, $i=1,2,\cdots I$

β_j: effect of second factor, $j=1,2,\cdots J$

$\delta_{ij} = 0$ means additive model : $\mu_{ij} = \mu + \alpha_i + \beta_j$

The effect of alcohol on Y is independent of the other factor smoking

$\delta_{ij} \neq 0$. means — — dependent — —

Estimate μ, α_i, β_j and δ_{ij} → row effect → column effect.

$$\underbrace{\mu_{ij}}_{I\times J} = \underbrace{\mu}_{1} + \underbrace{\alpha_i}_{I} + \underbrace{\beta_j}_{J} + \underbrace{\delta_{ij}}_{I\times J}$$

parameters

need to impose constraints : $\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$.

$\begin{cases} \sum_{i=1}^{I} \delta_{ij} = 0 \text{ for all } j = 1, \cdots J \\ \sum_{j=1}^{J} \delta_{ij} = 0 \text{ for all } i = 1, \cdots I-1 \end{cases}$

one way: $\mu_i = \mu + \alpha_i$, $\sum \alpha_i = 0$. or $\alpha_1 = 0$.

$1 + I + J + IJ - 1 - 1 - I - J + 1$

$= IJ$

MLE: $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{k_{ij}} \left\{ -\frac{(Y_{ijk} - \mu_{ij})^2}{2\sigma^2} - \frac{1}{2} \log \sigma^2 \right\}$$

$$\hat{\mu}_{ij} = \overline{Y}_{ij\cdot} = \frac{1}{k_{ij}} \sum_{k=1}^{k_{ij}} Y_{ijk}$$

$$\hat{\mu} = \frac{1}{k_{\cdot\cdot}} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{k_{ij}} Y_{ijk} = \overline{Y}_{\cdots}$$

$$\hat{\alpha}_i = \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots} \qquad \sum_{i=1}^{I} \hat{\alpha}_i = 0 \text{ satisfied}$$

$$\hat{\beta}_j = \overline{Y}_{\cdot j \cdot} - \overline{Y}_{\cdots} \qquad \sum \hat{\beta}_j = 0 \text{ satisfied.}$$

$$\hat{\delta}_{ij} = \hat{\mu}_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j = \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}$$

$$\sum_{i=1}^{I} \hat{\delta}_{ij} = 0 \quad \text{for all } j$$

Sums of Squares:

$$SS_A = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{k_{ij}} (\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2$$

(Sum of squares for factor A) $= \sum_{i=1}^{I} \sum_{j=1}^{J} k_{ij} (\underbrace{\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}}_{\hat{\alpha}_i})^2$

$$SS_B = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{k_{ij}} (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})^2$$

for fator B

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} k_{ij} (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})^2$$

$$SS_{AB} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{k_{ij}} \hat{\delta}_{ij}^2$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} k_{ij} \hat{\delta}_{ij}^2$$

Thm: $SS_T = SS_A + SS_B + SS_{AB} + SS_E$

SS_T: $\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{k_{ij}} (Y_{ijk} - \overline{Y}_{\cdots})^2$

SS_E: within group $\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{k_{ij}} (Y_{ijk} - \overline{Y}_{ij\cdot})^2$

Proof: $Y_{ijk} - \overline{Y}_{\cdots} = (\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}) + (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}) + (\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}) + (Y_{ijk} - \overline{Y}_{ij\cdot})$

Two-way-ANOVA

Suppose $Y_{ij} \sim N(\mu_{ijk}, \sigma^2)$ for $k = 1, 2, \cdots k_{ij}$

$i = 1, \cdots, I$; $j = 1, 2, \cdots J$

$\overline{Y_{\cdots}} = \frac{1}{K_{\cdot\cdot}} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{k_{ij}} Y_{ijk}$ $\quad \overline{Y_{i\cdot\cdot}}$

$\overline{Y_{ij\cdot}} = \frac{1}{k_{ij}} \sum_{k=1}^{k_{ij}} Y_{ijk}$ $\quad \overline{Y_{\cdot j\cdot}}$

Theorem. $SS_T = SS_A + SS_B + SS_{AB} + SS_E$ $\qquad K_{i\cdot} = \sum_{j=1}^{J} k_{ij}$

$SS_T = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{k_{ij}} (Y_{ijk} - \overline{Y_{\cdots}})^2$; $SS_A = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{k_{ij}} (\overline{Y_{i\cdot\cdot}} - \overline{Y_{\cdots}})^2 = \sum_{i=1}^{I} k_{i\cdot} (\overline{Y_{i\cdot\cdot}} - \overline{Y_{\cdots}})^2$

$SS_B = \sum_{j=1}^{J} k_{\cdot j} (\overline{Y_{\cdot j\cdot}} - \overline{Y_{\cdots}})^2$ $\qquad SS_{AB} = \sum_{i=1}^{I} \sum_{j=1}^{J} k_{ij} \hat{\delta}_{ij}^2$ where $\hat{\delta}_{ij} = \overline{Y_{i\cdot\cdot}} + \overline{Y_{\cdot j\cdot}} - \overline{Y_{ij\cdot}} - \overline{Y_{\cdots}}$

$SS_E = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{k_{ij}} (\overline{Y_{ij}} - Y_{ijk})^2$

$\mu_{ij} = \mu + \alpha_i + \beta_j + \delta_{ij}$

$\alpha_\cdot = 0 \quad \beta_\cdot = 0 \qquad \delta_{i\cdot} = 0$ for all i

$\delta_{\cdot j} = 0$ for all j

Two-way **ANOVA** Table:

	df	SS	MSS	F-test
Fator A	$I-1$	SS_A	$SS_A/(I-1)$	$\frac{MSS_A}{MSS_E}$
Fator B	$J-1$	SS_B	$SS_B/(J-1)$	MSS_B/MSS_E
Fator AB	$(I-1)(J-1)$	SS_{AB}	$SS_{AB}/[(I-1)(J-1)]$	MSS_{AB}/MSS_E
Error	$df_T - df_A - df_B - df_{AB}$	SS_E	$SS_E/(IJk-IJ)$	
Total	$IJk-1$	SS_T		

$k_{ij} = k$

$IJk - IJ$

Test: $H_0: \alpha_1 = \cdots = \alpha_I = 0$. Then under H_0: $\frac{MSS_A}{MSS_E} \sim F_{I-1, df_E}$

If $> F_{I-1, df_E, 1-\alpha}$ we reject H_0 at α.

Test: $H_0': \beta_1 = \cdots = \beta_J = 0$ $\qquad \frac{MSS_B}{MSS_E} \sim F_{J-1, df_E}$

Test: $H_0'': \delta_{ij} = 0$ (No Interaction) ($\mu_{ij} = \mu + \alpha_i + \beta_j$ additive model)

$$\frac{MSS_{AB}}{MSS_E} \sim F_{df_{AB}, df_E}$$

~~Under H_0''~~: We have $E(MSS_{AB}) = \sigma^2 + \frac{k}{df_{AB}} \sum_{i=1}^{I} \sum_{j=1}^{J} \delta_{ij}^2$ $\qquad \frac{k}{(I-1)(J-1)}$

$k_{ij} \equiv k$

Consider the special case of additive model.

$\mu_{ij} = \mu + \alpha_i + \beta_j \qquad Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$

$\gg 1 - \gg A + SS_B + SS_{AB} + SS_E$

Since the interaction term doesn't exist.

$$\frac{SS_{AB} + SS_E}{\sigma^2} \sim \chi^2_{df_{AB}+df_E}$$

Two-way anova for additive model:

	df	MSS	F-test
SS_A	$df_A = I-1$	SS_A/df_A	$\frac{SS_A/df_A}{\sigma^2_{pooled}} \sim F_{df_A,\, df_{AB}+df_E}$
SS_B	$df_B = J-1$	SS_B/df_B	$\vdots$
$SS_{AB}+SS_E$	$df_{AB}+df_E$	$\frac{SS_{AB}+SS_E}{df_{AB}+df_E} = \sigma^2_{pooled}$	

Ch14 Linear least squares.

Consider iid bivariate normal random vectors (X_i, Y_i) $i=1,\cdots,n$

$$\sim N\left((\mu_x, \mu_y), \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right)$$

where ρ is correlation.

How to estimate $E(Y|X=x)$

$$E(Y|X=x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)$$

$$Var(Y|X=x) = \sigma_y^2(1-\rho^2)$$

Regression equation: Given $(x_1, Y_1) \cdots (X_n, Y_n)$

$$Y = \hat{\mu}_y + \hat{\rho}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}(x-\hat{\mu}_x) \qquad \hat{\mu}_y = \frac{1}{n}\sum_{j=1}^{n} Y_j \quad \hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\rho} = \frac{\sum(X_i-\bar{X})(y_i-\bar{y})}{\sqrt{\sum(X_i-\bar{X})^2\sum(y_i-\bar{y})^2}} \quad \hat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \quad \hat{\sigma}_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

An alternative view: We treat X_i as deterministic

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

The regression equation is $y = \beta_0 + \beta_1 x$

The goal is to estimate β_0 and β_1 from observations (X_i, Y_i) $i=1,\cdots,n$.

explanatory variable / predictor. / independent variable / covariate

response / dependent variable

Assuming that Y_i are independent, then the likelihood function

$$\prod_{i}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{[Y_i - (\beta_0+\beta_1 x_i)]^2}{2\sigma^2}}$$

Therefore, we need to find $(\beta_0, \beta_1, \sigma^2)$ to minimize the likelihood function:

$$\frac{\partial lik}{\partial \beta_0} = \sum_{i=1}^{n} \frac{-2(Y_i - (\beta_0+\beta_1 x))}{\sigma^2} = 0$$

$$\frac{\partial lik}{\partial \beta_1} = \sum_{i=1}^{n} \frac{-2(Y_i-(\beta_0+\beta_1 x))x_i}{\sigma^2} = 0$$

Therefore, $\sum_{i=1}^{n}(Y_i - (\beta_0+\beta_1 x_i)) = 0$ $\quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i(Y_i-\bar{Y})}{\sum_{i=1}^{n} X_i(X_i-\bar{X})}$

$\sum_{i=1}^{n}(Y_i - (\beta_0+\beta_1 x_i))x_i = 0$ $\quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

let. $S_{XX} = \sum_{i=1}^{n}(x_i-\bar{X})^2 = \sum_{i=1}^{n}[(x_i-\bar{x})x_i]$.

$S_{YY} = \sum_{i=1}^{n}(y_i-\bar{y})^2$

$S_{XY} = \sum_{i=1}^{n}(x_i-\bar{X})(Y_i-\bar{Y}) = \sum_{i=1}^{n} x_i(y_i-\bar{y})$

So $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \hat{\rho}\frac{\sqrt{S_{YY}}}{\sqrt{S_{XX}}}$, $\quad \hat{\rho} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$

$y = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$.

(exactly the same as the one obtained assuming the bivariate Normal).

How about σ^2?

$$\frac{\partial \ln}{\partial \sigma^2} = \sum\left[-\frac{[Y_i-(\beta_0+\beta_1 x)]^2}{\sigma^4} + \frac{1}{\sigma^2}\right] = 0.$$

So. $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - (\hat{\beta}_0+\hat{\beta}_1 x_i)\right]^2$ MLE for σ^2

Properties of $\hat{\beta}_0$ $\hat{\beta}_1$ & $\hat{\sigma}^2$:

For $\hat{\beta}_1$: we claim that $E(\hat{\beta}_1) = \beta_1$ unbiased

$E(\hat{\beta}_1) = \frac{\sum x_i E(Y_i-\bar{Y})}{\sum x_i(x_i-\bar{X})} = \beta_1$ $\qquad EY_i = \beta_0 + \beta_1 X_i$ (model) <u>Assumption</u>

$E\bar{Y} = \beta_0 + \beta_1\bar{X}$

whether $\hat{\beta}_1 \to \beta_1$ <u>in P</u>.

Simple linear regression: $\dashrightarrow$ only one predictor (Simple)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i=1,\cdots,n$$

We want to estimate $\beta_0\ \beta_1$ from (x_i, y_i) $i=1,\cdots,n$

The least squared method: Find β_0 & β_1 to minimize

$$\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2$$

$$\hat\beta_1 = \frac{\sum_{i=1}^{n}(x_i-\bar x)(y_i-\bar y)}{\sum_{i=1}^{n}(x_i-\bar x)^2}, \quad \text{also} \quad \hat\beta_0 = \bar y - \hat\beta_1 \bar x$$

The estimated $(\hat\beta_0, \hat\beta_1)$ is unbiased: $E(\hat\beta_1)=\beta_1 \quad E(\hat\beta_0)=\beta_0$

We are now calculate $Var\,\hat\beta_1$

$$\hat\beta_1 = \frac{\sum_{i=1}^{n}(x_i-\bar x)(\beta_0+\beta_1 x_i+\varepsilon_i-(\beta_0+\beta_1\bar x+\bar\varepsilon))}{\sum(x_i-\bar x)^2}$$

$$= \frac{\sum(x_i-\bar x)(\beta_1 x_i-\beta_1\bar x)}{\sum(x_i-\bar x)^2} + \frac{\sum(x_i-\bar x)(\varepsilon_i-\bar\varepsilon)}{\sum(x_i-\bar x)^2}$$

$$= \beta_1 + \frac{\sum(x_i-\bar x)\varepsilon_i}{\sum(x_i-\bar x)^2}$$

ε_i, iid $\sim N(0,\sigma^2)$

Therefore $Var\,\hat\beta_1 = E(\hat\beta_1-\beta_1)^2$

$$= E\left\{\frac{\sum(x_i-\bar x)\varepsilon_i}{\sum(x_i-\bar x)^2}\right\}^2 = \frac{E\{\sum_{i=1}^{n}(x_i-\bar x)\varepsilon_i\}^2}{[\sum(x_i-\bar x)^2]^2} = \frac{\sum(x_i-\bar x)^2\sigma^2}{[\sum(x_i-\bar x)^2]^2}$$

$$= \frac{\sigma^2}{\sum(x_i-\bar x)^2}$$

Corollary: If $\lim_{n\to\infty}\sum_{i=1}^{n}(x_i-\bar x)^2=\infty$ then $E[\hat\beta_1-\beta_1]^2\to 0$.

and consequently $\hat\beta_1\to\beta_1$ in Probability

Our estimate will be close to the true but unknown parameter β_1.

$\hat\beta_0 = \bar y - \hat\beta_1\cdot\bar x$ Since $\bar y = \beta_0+\beta_1\bar x+\bar\varepsilon$

$= \beta_0+\beta_1\bar x+\bar\varepsilon-\hat\beta_1\bar x$ $\qquad \bar\varepsilon = \frac{\sum\varepsilon_i}{n}$.

So $Var(\hat\beta_0) = E(\hat\beta_0-\beta_0)^2 = E((\beta_1-\hat\beta_1)\bar x+\bar\varepsilon)^2 = E\left(\bar\varepsilon-\bar x\frac{\sum(x_i-\bar x)\varepsilon_i}{\sum_j(x_j-\bar x)^2}\right)^2$

$$= E\left(\sum_{i=1}^{n}\left[\frac{1}{n}-\frac{\bar x(x_i-\bar x)}{\sum_j(x_j-\bar x)^2}\right]\varepsilon_i\right)^2$$

$$= \sum_{i=1}^{n}\left[\frac{1}{n}-\frac{\bar x(x_i-\bar x)}{\sum_{j=1}^{n}(x_j-\bar x)^2}\right]^2\sigma^2$$

$$= \sigma^2 \left\{ \frac{n}{n^2} + \sum_{i=1}^n \left[\frac{-\bar{x}(x_i-\bar{x})}{\sum_{j=1}^n (x_j-\bar{x})^2} \right]^2 \right\}$$ cross term disappear

$$\Rightarrow \mathrm{Var}(\hat\beta_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{j=1}^n (x_j-\bar{x})^2}\right) \qquad \frac{\bar{x}^2 \sum \cancel{(x_i-\bar{x})^2}}{\left[\sum (x_j-\bar{x})^2\right]^{\cancel{2}}}$$

corollary: If $\lim_{n\to\infty} \frac{\bar{x}^2}{\sum_{j=1}^n (x_j-\bar{x})^2} = 0$ then $\hat\beta_0 \to \beta_0$ in Probability

since $\mathrm{Var}(\hat\beta_0) \to 0$.

$$\mathrm{Cov}(\hat\beta_0, \hat\beta_1) = \mathrm{Cov}(\beta_1\bar{x} + \bar{\varepsilon} - \hat\beta_1\bar{x},\ \hat\beta_1 - \beta_1)$$

$$\hat\beta_1 - \beta_1 = \frac{\sum_{i=1}^n (x_i-\bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i-\bar{x})^2}$$

$$= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i - \bar{x}\frac{\sum_{i=1}^n (x_i-\bar{x})\varepsilon_i}{SSx},\ \frac{\sum_{i=1}^n (x_i-\bar{x})\varepsilon_i}{SSx}\right)$$

$$= \mathrm{Cov}\left(\sum_{i=1}^n \varepsilon_i\left(\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{SSx}\right),\ \sum_{j=1}^n \varepsilon_j \frac{x_j-\bar{x}}{SSx}\right)$$

$$= \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{SSx}\right)\frac{x_i-\bar{x}}{SSx}\sigma^2$$

$$= \sum_{i=1}^n \frac{-\bar{x}(x_i-\bar{x})^2}{SSx^2}\sigma^2 = -\frac{\bar{x}}{SSx}\sigma^2$$

In summary: $$\begin{pmatrix}\hat\beta_0 \\ \hat\beta_1\end{pmatrix} \sim N\left(\begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix}, \sigma^2\begin{pmatrix}\frac{1}{n}+\frac{\bar{x}^2}{SSx} & -\frac{\bar{x}}{SSx} \\ -\frac{\bar{x}}{SSx} & \frac{1}{SSx}\end{pmatrix}\right)$$

The mle $\hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n [y_i - (\hat\beta_0 + \hat\beta_1 x_i)]^2$

Testing: $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$.

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

yield fertilizer

Under H_0: fertilizer is useless

$\mathrm{Var}(\hat\beta_1) = \sigma^2 \frac{1}{SSx}$ If σ^2 known, $\left|\frac{\hat\beta_1 - \beta_1}{\sqrt{\sigma^2/SSx}}\right|$ to 1.96

We should conduct a t-test if σ^2 unknown

Studentized: $$\frac{\hat\beta_1 - \beta_1}{\sqrt{\tilde\sigma^2/SSx}}$$

$$\hat\sigma^2_{mle} = \frac{1}{n}\sum_{i=1}^n (y_i - \hat\beta_0 - \hat\beta_1 x_i)^2$$

use $S_p^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat\beta_0 - \hat\beta_1 x_i)^2$ unbiased estimator of σ^2

It turns out that $\frac{(n-2)S_p^2}{\sigma^2} \sim \chi^2_{n-2}$.

Hence, $$\underbrace{\frac{\hat\beta_1 - \beta_1}{\sqrt{\sigma^2/SSx}}}_{N(0,1)} \Bigg/ \underbrace{\sqrt{S_p^2/\sigma^2}}_{\sqrt{\chi^2_{n-2}/(n-2)}} \sim t_{(n-2)}$$

So the Hypothesis $H_0: \beta_1 = 0$. can be tested by $\left| \dfrac{\hat{\beta}_1}{\sqrt{S_p^2 / SS_X}} \right| > t_{n-2, 1-\frac{\alpha}{2}}$

reject @ level α.

ANOVA: for regression: model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

The total variance: $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$

$SS_E = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

regression or between $SS_B = \sum_{i=1}^{n} (\underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\text{fitted value, } i\text{th item}} - \bar{y})^2$.

fitted value. $\rightarrow$ i^{th} average. $-$ overall average

$SS_T = SS_E + SS_B$

SS	df
SS_T	$n-1$
SS_E	$n-2$
SS_B	1

Recall in the one-way ANOVA case:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_I$

$$\frac{SS_B / df_B}{SS_E / df_E}$$

In the regression case:

$$\frac{SS_B / 1}{SS_E / df_E}$$

testing for H_0: β_1 same for all i

$\Leftrightarrow \beta_1 = 0$

Simple linear regression:

basic setup: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; $i = 1, \cdots, n$ Assume: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$$\begin{pmatrix} \hat\beta_0 \\ \hat\beta_1 \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{SS_x} & -\frac{\bar{x}}{SS_x} \\ -\frac{\bar{x}}{SS_x} & \frac{1}{SS_x} \end{pmatrix} \right), \text{ where } SS_x = \sum_{i=1}^n (x_i - \bar{x})^2$$

So the estimate $\hat\beta_1$ is consistent if $SS_x \longrightarrow \infty$ as $n \to \infty$

$$SS_T = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Where $\hat{y}_i = \hat\beta_0 + \hat\beta_1 x_i$ is the fitted value of (y_i, x_i)

$SS_B = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ regression effect

$SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ F-test statistics: $\dfrac{SS_B}{SS_E/(n-2)}$

S_p^2 = estimate of σ^2

Remark 1: If $\bar{x} = 0$, then the joint distribution: $\begin{pmatrix} \hat\beta_0 \\ \hat\beta_1 \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \sigma^2 \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{SS_x} \end{pmatrix} \right)$

orthogonal design: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

inner product is $0 \Leftrightarrow$ orthogonal

$(x_1 + \cdots + x_n = 0)$

has the following advantage: $\hat\beta_0$ and $\hat\beta_1$ are independent.

Remark 2: $\dfrac{SS_B}{SS_T} = R^2 \leq 1$ percentage being explained by regression.

Remark 3: The error distribution $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ ε_i can be estimated by $\hat\varepsilon_i = y_i - \hat{y}_i = y_i - (\hat\beta_0 + \hat\beta$

Check normality for $\hat\varepsilon_1, \hat\varepsilon_2, \cdots, \hat\varepsilon_n$ residual plot, qqnorm plot

(either line or light tail is fine)

Regression diagnostics

Multiple Regression: Multiple predictors

Statistical Model:

$$y_i = f(x_{i1}, \cdots x_{ip}) + \underbrace{\varepsilon_i}_{\text{error}}$$

Our goal of estimation is to find the function f based on observations

$$\begin{matrix} y_1 & x_{11} & \cdots & x_{1p} \\ y_2 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ y_n & x_{n1} & \cdots & x_{np} \end{matrix}$$

Assumption: f is linear

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Therefore, we only need to estimate unknown parameters $\beta_0, \beta_1, \cdots, \beta_p$

All models are wrong;
but some are useful

Taylor's expansion:

$$f(t) = \underbrace{f(t_0) + f'(t_0)(t-t_0)}_{\text{linear}}$$

If $p=1$, then we can resort to least squares method.

For the multiple linear regression, we can still use least squares method:

$$l(\beta_0, \cdots, \beta_p) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \right)^2$$

The minimizer $(\hat{\beta}_0, \cdots, \hat{\beta}_p)$ solves the equation

$$\frac{\partial l(\beta_0, \cdots, \beta_p)}{\partial \beta_0} = 0, \quad \cdots, \quad \frac{\partial l(\beta_0, \cdots, \beta_p)}{\partial \beta_p} = 0.$$

$$\frac{\partial l}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \right) = 0$$

$$\frac{\partial l}{\partial \beta_1} = -2 \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \right) x_{i1} = 0$$

$$\vdots$$

$$\frac{\partial l}{\partial \beta_p} = -2 \sum (\quad \cdots \quad) x_{ip} = 0.$$

$$\Rightarrow \quad \sum (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) = \sum y_i.$$

$$\sum (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p (x_{ip})) x_{i1} = \sum y_i x_{i1}$$

$$\vdots$$

$$\sum (\quad \quad) x_{ip} = \sum y_i x_{ip}$$

$$\Rightarrow \begin{pmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_{i1} & \cdots & \sum_{i=1}^n x_{ip} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \cdots & \sum_{i=1}^n x_{ip}x_{i1} \\ \vdots & & & \vdots \\ \sum_{i=1}^n x_{ip} & & \cdots & \sum_{i=1}^n x_{ip}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Design Matrix: $X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$

then $\Rightarrow$ $X^T X \underbrace{\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}}_{\beta} = X^T \underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{y}$ $\Rightarrow$ $(X^T X)\beta = X^T y$

Normal equation, therefore

$\hat{\beta} = (X^T X)^{-1} X^T y$ $X^T X$ is not singular

$\hat{\beta} = (X^T X)^{-1} X^T y$ also minimizes

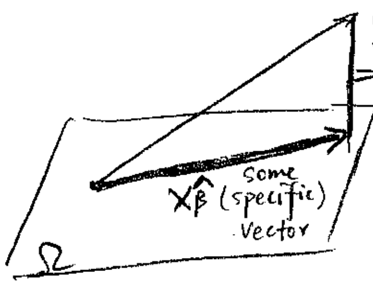
$\|y - X\beta\|^2 = \langle y - X\beta, y - X\beta \rangle$ inner product.

g. $u = (u_1, \cdots, u_n)$

$v = (v_1, \cdots, v_n)$

$\langle u, v \rangle = u_1 v_1 + \cdots + u_n v_n = u v^T$ $\|u\|^2 = u_1^2 + \cdots + u_n^2 = \langle u, u \rangle = u u^T$

Geometric Interpretation for β



$y \rightarrow$ n dimensional space.

$y - X\hat{\beta}$

linear space spanned by column of X.

$\beta_0 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} + \beta_p \begin{pmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{pmatrix}$

find $X\hat{\beta}$ minimize $\|y - X\beta\|^2$ (length)

$X\hat{\beta}$ is the projection of the vector y onto Ω

$= Hy$ H is the projection matrix. $H = X(X^T X)^{-1} X^T \longrightarrow$ hat matrix.

In other words $\frac{RSS}{\sigma^2} = \frac{\varepsilon'\varepsilon}{\sigma^2} \sim \chi^2_{n-p-1}$

So $S_p^2 = \frac{1}{n-p-1} RSS$ is an unbiased estimator of σ^2

Additionally : $$\frac{\hat{\beta}_i - \beta_i}{S_p\sqrt{V_{ii}}} \sim t_{n-p-1}$$