Review the notes
Discussion

Overview of:

- Basic Statistical Concepts of population and sample

- Standard distributions- N, t (central, non-central), Chisquare (central, non-central), F- (central & non-central)

- Sampling and Sampling Distributions

- Inferences About the Differences in Means (Randomized Designs)
- Hypothesis Testing
- Confidence Intervals

- Choice of Sample Size;
- Inferences About the Differences in Means (Paired Comparison Designs)

- Inferences About the Variances of Normal Distributions (parametric procedures)

- R/Rmd codes.

Adapted from "Design and Analysis of Experiments, 10th Edition, by Douglas C. Montgomery (John Wiley, 2020).

Some Standard Results on Sampling Distributions: Estimation and tests of Hypotheses

Let $y_1, y_2 \ldots y_n$ r. s. (random sample) from

Population

μ, σ^2

Let $\bar{y} = \frac{\sum y_i}{n}$: sample mean

Then

$$E(\bar{y}) = \mu$$

Let $S^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$

sample variance

Then $E(S^2) = \sigma^2$.

The Central Limit Theorem

If $y_1, y_2, \ldots, y_n$ a r.s.

or a sequence of n

iid r.v. $\sim (\mu, \sigma^2)$

$|\mu| < \infty$

$\sigma < \infty$

and

$x = y_1 + y_2 + \cdots + y_n$

then

$$Z_n = \frac{x - n\mu}{\sqrt{n\sigma^2}}$$

$\sim N(0,1)$ as $n \to \infty$.

$$\text{Or, } Z_n = \frac{\frac{x}{n} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1) \text{ as } n \to \infty$$

Sum of Squares and Degrees of freedom

$y_1, y_2 \ldots \; y_n$ r. s. from (μ, σ^2)

deviations : $y_i - \bar{y}$

SS (using n terms)

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

Note: $\underbrace{\sum_{i=1}^{n} (y_i - \bar{y}) = 0}_{\text{1 linear constraint}}$

$$DF(SS) = n - 1$$

Normal distribution

r. v. Y.

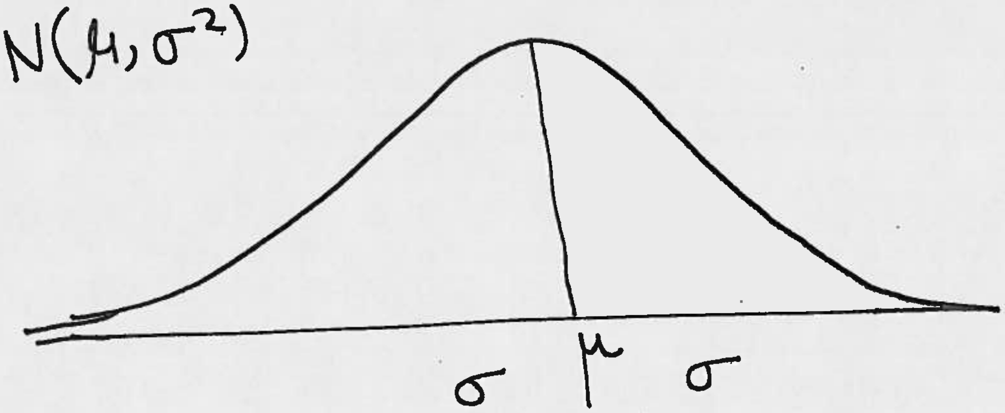
value y

pdf $f_Y(y)$ or simply $f(y)$

$$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2/\sigma^2}$$

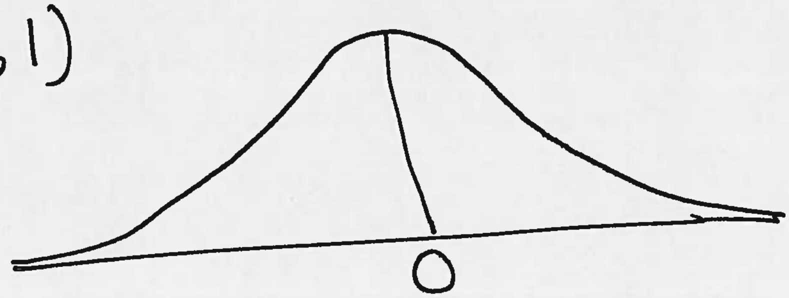
$$\left.\begin{array}{l} -\infty < \mu < \infty \\ -\infty < y < \infty \\ \sigma > 0 \end{array}\right]$$

$Y \sim N(\mu, \sigma^2)$



$$\downarrow \quad Z = \frac{Y-\mu}{\sigma}$$

$Z \sim N(0,1)$



Chi-square or χ^2 distribution

Central χ^2 distribution

If $Z_1, Z_2, \ldots Z_k$ are iid $N(0,1)$

then

$$x = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

$$\sim \chi^2_k \quad \text{(central } \chi^2\text{-dist)}$$

$$\mu = E(x) = k, \quad \sigma^2 = V(x) = 2k$$

pdf of X

$$f(x) = \frac{1}{2^{k/2}\,\Gamma(k/2)}\, x^{\frac{k}{2}-1}\, e^{-x/2} \qquad x > 0$$

k: degrees of freedom

Example Application

$$\frac{SS}{\sigma^2} = \frac{\sum (y_i - \bar{y})^2}{\sigma^2} \sim \chi^2_{n-1}$$

Non-central χ^2

let $X_1 \dots X_k$ be indep & normal

$$X_i \sim N(\mu_i, 1)$$

Then

$$X = \sum_{i=1}^{k} X_i^2$$

$\sim$ non-central χ^2 with k degrees of freedom and $\lambda = \sum_{i=1}^{k} \mu_i^2$ non-centrality parameter.

pdf of X

$$f_X(x; k, \lambda) = \sum_{i=0}^{\infty} \frac{e^{-\lambda/2} (\lambda/2)^i}{i!} f_{Y_{k+2i}}(x)$$

where $Y_{k+2i} \sim$ (central) $\chi^2(k+2i$ d.f.)

t-distribution

If $Z \sim N(0,1)$

$\chi^2_k \sim \chi^2$ with k df

Z and χ^2_k independent

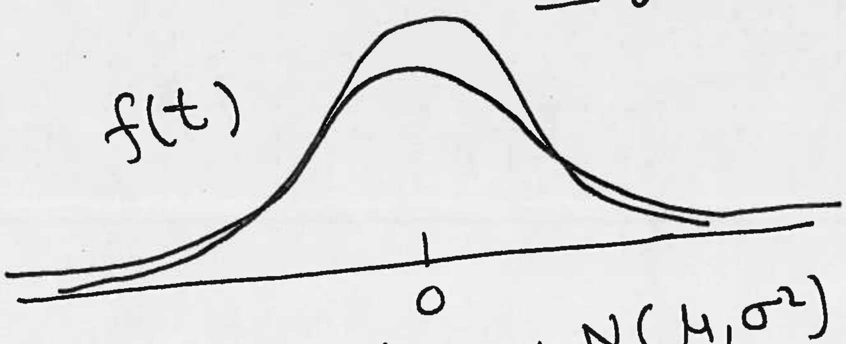
Then

$$t_k = \frac{Z}{\sqrt{\frac{\chi^2_k}{k}}} \sim t \text{ with } k \text{ df}$$

<u>pdf</u> of t_k

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)} \frac{1}{\left[(t^2/k)+1\right]^{\frac{k+1}{2}}}$$

$-\infty < t < \infty$



Application

<u>Example</u>

$y_1, y_2 \cdots y_n \sim N(\mu, \sigma^2)$

$t = \frac{\bar{y}-\mu}{s/\sqrt{n}} \sim t$ with $n-1$ d.f.

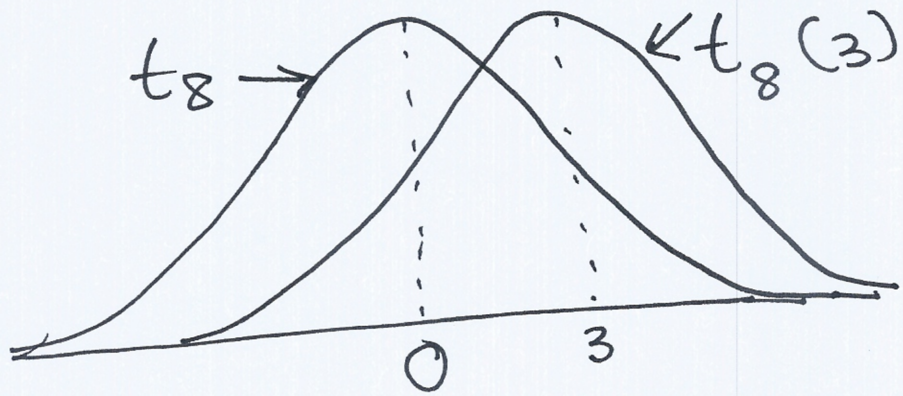
Non-central t-distribution

$$t_k(\vartheta) = \frac{Z + \vartheta}{\sqrt{\frac{\chi^2_k}{k}}} \qquad \vartheta \neq 0$$

is (called) non-central t r.v. with

k d.f.

ϑ noncentrality parameter



R codes

```
dt(x, df, ncp, ...)
pt(q, df, ncp, ...)
qt(p, df, ncp, ...)
rt(n, df, ncp)
```

F-distribution

If $\chi^2_u \sim \chi^2$ with u df

$\chi^2_v \sim \chi^2$ with v df

$\chi^2 u$ & $\chi^2 v$ indep.

Then

$$F_{u,v} = \frac{\chi^2_u / u}{\chi^2_v / v}$$

$\sim F$ with u, v d.f.

numerator   Denominator

d.f

pdf of F

$$h(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{u/2} x^{\frac{u}{2}-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x+1\right]^{(u+v)/2}}$$

$x > 0$

Example / Application

$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}$$

Inference on parameters of a population

Population

μ, σ^2

r.s.

$y_1 \cdots y_n$

- Estimation

  Point

  confidence interval

- Hypothesis testing

$\mu \longrightarrow \hat{\mu} = \hat{\mu}(y_1 \cdots y_n)$

$CI(\mu) = (L, U)$

$P(L < \mu < U) = 1-\alpha$

L, U are function of $y_1 \cdots y_n$.

Hypothesis testing

Hypothesis: A statement about the parameter(s) of a statistical popn.

<u>Illustration</u>:

Popn μ, σ^2

1. **Population ~ Normal, sample size: any n;**

or

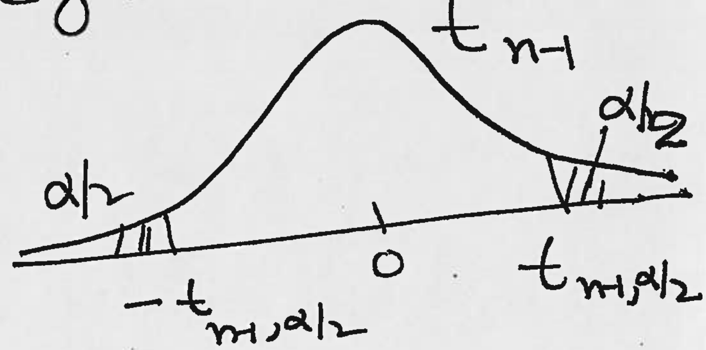
2. **Any population (var finite) but large n**

1. $H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$

2. Test statistic $TS \equiv \frac{\bar{y}-\mu_0}{s/\sqrt{n}} = t$

3. Critical region



4. Conclusion

Reject H_0 if $|t| > t_{n-1;\alpha/2}$

Type I error

$\alpha = \text{Prob}(\text{Type I error})$

$= \text{Prob}(\text{reject } H_0 \mid H_0 \text{ is true})$

Type II error

$\beta = P(\text{Type II error}) = P(\text{fail to reject } H_0 \mid H_0 \text{ is false})$

$\text{Power} = 1-\beta = P(\text{reject } H_0 \mid H_0 \text{ is false})$

α: significance level; is fixed in advance.

p-value

$p = \text{Prob}(TS > TS_{obs}$ or more extreme).

Applicatims

1. The two-sample t-test

Popn 1

Treatment 1

$N(\mu_1, \sigma_1^2)$

$\downarrow$

$y_{11}\ y_{12} \cdots y_{1n_1}$

$\bar{y}_1,\ S_1^2$

Popn 2

Treatment 2

$N(\mu_2, \sigma_2^2)$

$y_{21}, y_{22} \cdots y_{2n_2}$

$\bar{y}_2,\ S_2^2$

Assumptim : $\sigma_1^2 = \sigma_2^2$

$H_0: \quad \mu_1 = \mu_2$

$H_1: \quad \mu_1 \neq \mu_2$

TS

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

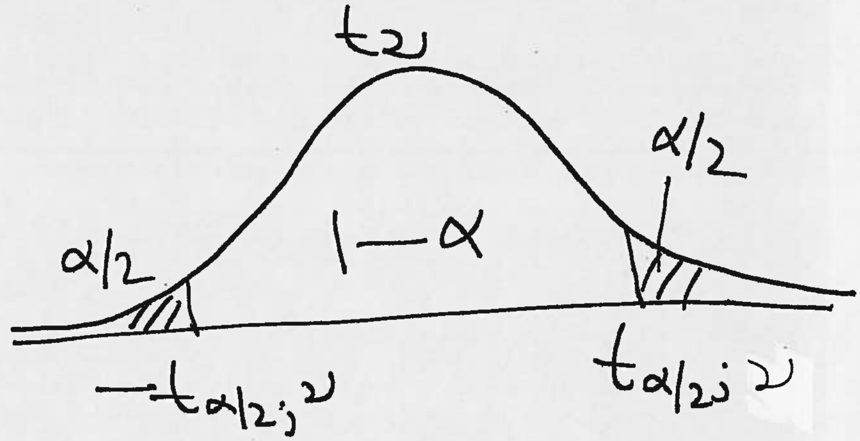
$$\sim t_{n_1+n_2-2} \text{ under } H_0$$

where

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

CRITICAL REGION $\quad \nu = n_1+n_2-2$

$$|t_0| > t_{\frac{\alpha}{2};\nu}$$



Conclusion Reject H_0: if $|t_0| > t_{\frac{\alpha}{2};\nu}$

$$\text{P-value} = P\left[\,|t_\nu| > |t_0|\,\right]$$

$$= 2P\left[\,t_\nu > |t_0|\,\right]$$

Normality Assumption

Normal probability plot

$y_1, y_2 \cdots y_n$ r.s.

$\rightarrow$ order statistics

$y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$

- Observed cumulative frequency
  j-th ordered observation:

$$\frac{j - .5}{n}$$

- Get the normal score z_j

$$\int_{-\infty}^{z_j} \phi(z)\,dz = \frac{j - .5}{n}$$

- Or, Get the cumulative prob from assumed normality of y's.

$$\Phi\left(\frac{y_{(j)} - \bar{y}}{s}\right)$$

- Plot: y_axis : $\frac{j-.5}{n}$, $\Phi(\;)$
  x-axis : $y_{(j)}$ values

2.4.2 Choice of Sample Size

Situation

$H_0: \quad \mu_1 = \mu_2$

$H_1: \quad \mu_1 \neq \mu_2$

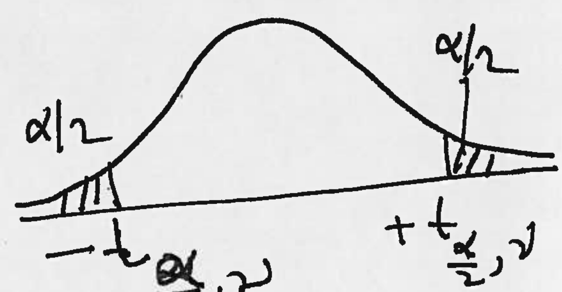
Given prob of Type I error $= \alpha$
Type II error $= \beta$

Power $= 1 - \beta$

Difference to be detected $\mu_1 - \mu_2 = \delta$

Distributions: Normal
$\sigma_1^2 = \sigma_2^2 = \sigma^2$ unknown.

$$TS = \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = t^*$$



$\nu = n_1 + n_2 - 2.$

Let $n_1 = n_2 = n$.

Equation for n : iterative solution

$\beta = P(\text{fail to reject } H_0 \mid H_0 \text{ is false})$
$= 1 - P(\text{reject } H_0 \mid H_1 \text{ is true})$

$$\beta = 1 - P\left(|t| > t_{\frac{\alpha}{2},\nu} \mid \mu_1 - \mu_2 = \delta\right)$$

$$= P\left(-t_{\alpha/2,\nu} \le t \le t_{\alpha/2,\nu} \mid \mu_1 - \mu_2 = \delta\right)$$

$$= P\left(-t_{\alpha/2,\nu} \le \frac{y_1 - y_2 - \delta + \delta}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \le t_{\alpha/2,\nu}\right)$$

$$= P\left(-t_{\alpha/2,\nu} \le t + \frac{\delta}{s_p\sqrt{\frac{2}{n}}} \le t_{\alpha/2,\nu}\right)$$

Approximation:

$t \sim t_\nu$ for any δ.

$$= P\left(-t_{\frac{\alpha}{2},\nu} - \frac{\delta}{s_p\sqrt{\frac{2}{n}}} \le t \le t_{\frac{\alpha}{2},\nu} - \frac{\delta}{s_p\sqrt{\frac{2}{n}}}\right)$$

If S_p is constant, e.g., for large n, then

$$\beta = P\left(t_\nu \le t_{\frac{\alpha}{2},\nu} - \frac{\delta}{s_p\sqrt{\frac{2}{n}}}\right) - P\left(t_\nu \le -t_{\frac{\alpha}{2}} - \frac{\delta}{s_p\sqrt{\frac{2}{n}}}\right)$$

a function of (δ, n) for given α & β.

Solution iterative.

The exact computation of the power will be done using non-central t-distribution. See the supplementary notes for NT Section 5.4.

An other situation; for an explicit solution of n.

Single population, normal, σ known, one-sided alternative.

H_0: $\mu = \mu_0$

H_1: $\mu > \mu_0$

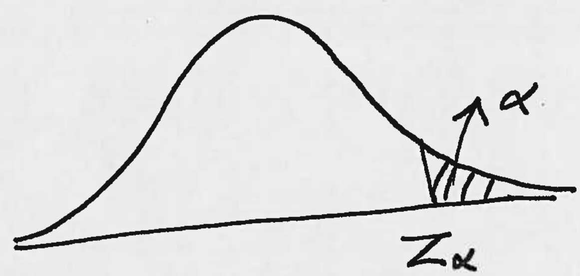
or $\mu = \mu_0 + \delta$, $\delta > 0$

TS
$$Z^* = \frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \text{ under } H_0$$

CR

Reject Ho

if $Z^* > Z_\alpha$



Here

$$\beta = 1 - P(\text{reject } H_0 \mid H_1 \text{ true})$$
$$= 1 - P(Z^* > Z_\alpha \mid H_1 \text{ true})$$
$$= 1 - P\left(\frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > Z_\alpha \;\middle|\; \mu_1 = \mu_0 + \delta \right)$$

$$\beta = 1 - P\left(\frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > Z_\alpha \,\middle|\, \bar{y} \sim N\left(\mu_1 = \mu_0 + \delta, \frac{\sigma^2}{n}\right)\right)$$

$$= 1 - P\left(\frac{\bar{y} - \mu_1 + \mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}} > Z_\alpha \,\middle|\, \bar{y} \sim N\left(\mu_0 + \delta, \frac{\sigma^2}{n}\right)\right)$$

$$= 1 - P\left(Z + \frac{\delta}{\frac{\sigma}{\sqrt{n}}} > Z_\alpha \right)$$

$$= 1 - P\left(Z > Z_\alpha - \frac{\delta\sqrt{n}}{\sigma} \right)$$

$$\beta = P\left(Z < Z_\alpha - \frac{\delta\sqrt{n}}{\sigma} \right)$$

$$= \Phi\left(Z_\alpha - \frac{\delta\sqrt{n}}{\sigma} \right).$$

or,

$$Z_\alpha - \frac{\delta\sqrt{n}}{\sigma} = \Phi^{-1}(\beta)$$

$$n = \left[\left(Z_\alpha - \Phi^{-1}(\beta) \right) \frac{\sigma}{\delta} \right]^2$$

If $\alpha = 0.05$, $\beta = 0.10$ [i.e. power $= 0.90$]

then $$n = \Big[\big(\underset{Z_\alpha}{1.64} - \underset{\Phi^{-1}(.1)}{(-1.28)} \big) \frac{\sigma}{\delta} \Big]^2$$

$$= 8.526\ \sigma^2/\delta^2.$$

Confidence Intervals $1-\alpha$ c.c.

Case

	Popn 1	Popn 2
	$N(\mu_1, \sigma_1^2)$	$N(\mu_2, \sigma_2^2)$
	$\downarrow$	$\downarrow$
	r. s. n_1	r. s. n_2
	$\bar{y}_1, S_1^2$	$\bar{y}_2, S_2^2$

(i) $\underline{\sigma_1^2 = \sigma_2^2} = \sigma^2$ (unknown)

Result
$$\frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Thus using the statement

$$P\left(-t_{\frac{\alpha}{2},\nu} \le \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \le t_{\frac{\alpha}{2},\nu}\right)$$
$$= 1-\alpha$$

where $\nu = n_1 + n_2 - 2$

This results into

$1-\alpha$ CI $(\mu_1 - \mu_2)$:

$$\bar{y}_1 - \bar{y}_2 \mp t_{\frac{\alpha}{2},\nu}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

(ii) $\underline{\sigma_1^2 \neq \sigma_2^2}$: both unknown.

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$\sim$ approx t_{ν^*}

where $$\nu^* = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

$1-\alpha$ CI $(\mu_1 - \mu_2)$

$$= \bar{y}_1 - \bar{y}_2 \mp t_{\frac{\alpha}{2},\nu^*} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

2.5.1 The Paired Comparison Problem

U_i $\quad$ y_{1i}, y_{2i}

$d_i = y_{1i} - y_{2i}$

y_{1i}, y_{2i}: not independent.

Single sample problem

solution on differences

d_i, $i=1 \ldots n$.

2.6. Inferences about variances

(i) Single population

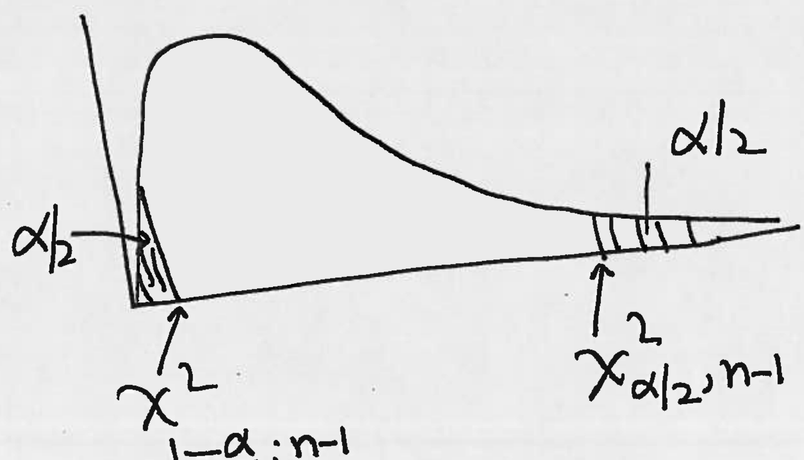
$$y_1, \ldots, y_n \sim N(\mu, \sigma^2)$$

a) Test for $H_0: \sigma^2 = \sigma_0^2$

$H_1: \sigma^2 \neq \sigma_0^2$

TS $$\chi_0^2 = \frac{SS}{\sigma_0^2} = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

<u>CR</u>



Reject H_0 if $\chi_0^2 > \chi^2_{\frac{\alpha}{2}, n-1}$

or $\chi_0^2 < \chi^2_{1-\frac{\alpha}{2}, n-1}$

b) $(1-\alpha)$ CI (σ^2):

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}, n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}}$$

(ii) Two populations

Popn 1: $N(\mu_1, \sigma_1^2)$

Popn 2: $N(\mu_2, \sigma_2^2)$

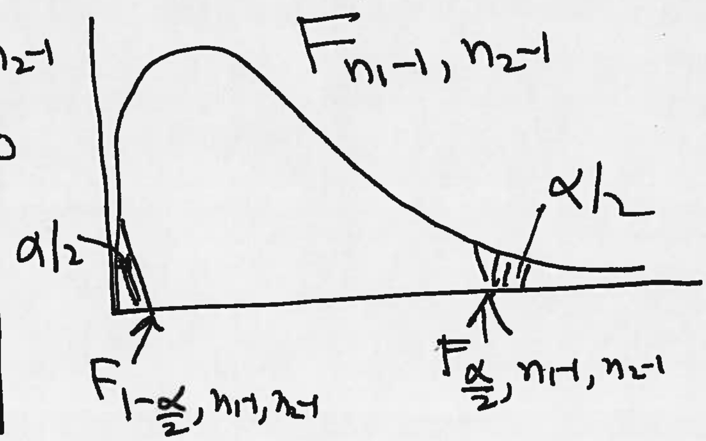
a) Test for $H_0: \sigma_1^2 = \sigma_2^2$

$H_1: \sigma_1^2 \neq \sigma_2^2$

TS $\quad F_0 = \dfrac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}$

CR: Reject Ho: if $F_0 > F_{\frac{\alpha}{2}; n_1-1, n_2-1}$

or $F_0 < F_{1-\frac{\alpha}{2}; n_1-1, n_2-1}$



[Use a relation to use tables on F.

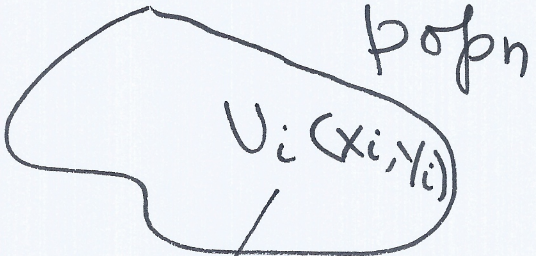
$$F_{1-\alpha; \nu_1, \nu_2} = \frac{1}{F_{\alpha; \nu_2, \nu_1}}$$
]

b) $(1-\alpha)$ CI $\left(\frac{\sigma_1^2}{\sigma_2^2}\right)$

$$\frac{S_1^2}{S_2^2} F_{1-\frac{\alpha}{2}; n_2-1, n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2} F_{\frac{\alpha}{2}, n_2-1, n_1-1}.$$

Linear Regression – selected points

Obs. study



$U_i : \{X_i, Y_i\}$ as they were observed/found

- no intervention to affect X, Y association, relationship

Exp. design

intervention applied through the choice of X_i values/levels

$X_i(U_i)$: Experimenter's choice

$Y_i(U_i)$: effect/response

Simple Linear regression

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- X_i's are fixed by assumption after Y_i's were observed

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- X_i's are fixed by design before Y_is were observed

- $\epsilon_i \sim N(0, \sigma^2)$

General linear model

Several predictors (cause vars)

$X_1, X_2, \cdots X_{p-1}$

one response variable (effect var)

Assumed form of the model to begin with:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_{p-1} X_{p-1\,i} + \epsilon_i$$

- $X_{1i}, \cdots X_{p-1\,i}$ are fixed by assumption or design
- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Purpose Infer $\beta_0, \beta_1 \cdots \beta_{p-1}, \sigma^2$

using Data $= \{(X_{1i}, X_{2i}, \cdots X_{p-1\,i}, Y_i)$
$i = 1 \cdots n\}$.

Notations: vector forms

$$\underline{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1}, \quad \underset{\sim}{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p-1\,1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1n} & & X_{p-1\,n} \end{bmatrix}_{n \times p}$$

$$\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \\ \beta_{p-1} \end{bmatrix}_{p\times 1}, \quad \underline{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

Estimators of $\underline{\beta}$

- LSE : Minimize

$$\phi = \sum_{i=1} (Y_i - \beta_0 - \beta_1 X_{1i} \cdots - \beta_{p-1} X_{p-1\,i})^2$$

$$= (\underline{Y} - \underset{\sim}{X}\underline{\beta})'(\underline{Y} - \underset{\sim}{X}\underline{\beta}).$$

- MLE

Consider LSE:

$$\hat{\underline{\beta}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underline{Y}$$

$$V(\hat{\underline{\beta}}) = \sigma^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$$

Testing the role/influence of $X_1, \cdots X_{p-1}$

H_0 : $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ (no role)

H_1 : at least one of the β_i's $i=1 \ldots p-1$ is nonzero.

Procedure : ANOVA

Test statistic

$$F_0 = \frac{MSR}{MSE}$$

$$\sim F_{p-1,\, n-p} \text{ under } H_0.$$

Testing the role/influence of a single variable, say X_i

$H_0: \quad \beta_i = \beta_{i0}$ (including $\beta_{i0} = 0$)

Test statistic

$$t_0 = \frac{\hat{\beta}_i - \beta_{i0}}{s\sqrt{\{(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\}_{ii}}}$$

↳ for β_i.

$s^2 = MSE$

$t_0 \sim t_{n-p}$ under H_0.

confidence intervals.

R codes, etc.