Topic: Comparing 2 Samples
(Ch. 11.1 – 11.2)

10-31-22

Setting: $(X_1, \ldots, X_n)$ identically distrib't'ed w/ a continuous density, $f(x;\theta)$

and

$(Y_1, \ldots, Y_m)$ identically distrib't'ed w/ a continuous density, $f(y;\lambda)$

Recall some standard notation:

$$\text{population parameters}\left\{\begin{array}{lll} E(X) = \mu_X & \approx & \hat{\mu}_X = \overline{X} \\ Var(X) = \sigma_X^2 & \approx & \hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ Cov(X,Y) = \sigma_{XY} & \approx & \hat{\sigma}_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \end{array}\right\}\begin{array}{l}\text{sample estimates} \\ \text{from} \\ x_{obs} = (x_1, \ldots, x_n) \\ \text{and} \\ y_{obs} = (y_1, \ldots, y_n)\end{array}$$

And, recall the properties of expectation from our review earlier in the semester. (ie. Ch. 4)

Note: Usually we are interested in whether or not the differences btwn each group are discernably different from zero.

Case 1: Independent Samples

Setting: $(X_1, \ldots, X_n)$ are independent observations
$(Y_1, \ldots, Y_m)$ are independent observations

and

$(X_1, \ldots, X_n)$ is indepent from $(Y_1, \ldots, Y_m)$

Method: Consider the RV $W = \bar{X} - \bar{Y}$ and use properties of independence to characterize the most probable values of $\mu_X - \mu_Y$.

Independence

- Between groups
- Among individual observations w/in each group

Approaches

- Normal Method (parametric)
- Mann-Whitney Test (non-parametric)

Case 2: Dependent (Paired) Samples

Setting: $m=n$ and $(X_1,\dots,X_n)$ and $(Y_1,\dots,Y_m)$ are dependent in a particular way, namely in a way that elements of each can be paired (eg. as before-after observations). And $(X_1,Y_1),\dots,(X_n,Y_n)$ are independent pairs.

Method: Create a data vector of the differences btwn each paired data point

$$d_i = x_i - y_i$$

and proceed w/ 1-sample methods for the IID vector of differences $(D_1,\dots,D_n)$.

Approaches

- Normal Method (parametric)
- Signed-Rank Test (non-parametric)

Group Work - stewardship and inference

11-2-22

Normal Theory for Comparing 2 Samples

General Setting + Relative Effeciency :

For data $X_1, \dots, X_n \overset{IID}{\sim}$ w/ $E(X_i) = \mu_X$, $Var(X_i) = \sigma_X^2$

and $Y_1, \dots, Y_m$ are IID w/ $E(Y_j) = \mu_Y$, $Var(Y_j) = \sigma_Y^2$

the RV $W = \bar{X} - \bar{Y}$ has

$$E(W) = E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - E\left(\frac{1}{m}\sum_{j=1}^{m} Y_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i) - \frac{1}{m}\sum_{i=1}^{m} E(Y_i)$$

$$= \mu_X - \mu_Y$$

and

$$Var(W) = Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) - 2Cov(\bar{X}, \bar{Y})$$

$\sigma_{\bar{X}\bar{Y}} = Cov(\bar{X}, \bar{Y})$

$\rho_{\bar{X}\bar{Y}} = \frac{Cov(\bar{X}, \bar{Y})}{\sigma_{\bar{X}}\,\sigma_{\bar{Y}}}$

$$= \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m} - 2\sigma_{\bar{X}}\sigma_{\bar{Y}}\rho_{\bar{X},\bar{Y}}$$

$\sigma_{\bar{X}} = Var(\bar{X})$

If any X_i are independent from any Y_j then

then $Cov(X_i, Y_j) = 0$ and $Cov(\bar{X}, \bar{Y}) = 0$

$$Var(W) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$$

Note : If the X_i's and Y_j's are positively correlated, and $n = m$, then pairing is more effective than not pairing.

Furthermore, if $(X_1, \dots, X_n)$ are Normally distbted then.

$$\bar{X} \sim N(\mu_x, \sigma_x^2/n)$$

similarly for $(Y_1, \dots, Y_m)$ Normally distbted, $\bar{Y} \sim N(\mu_y, \sigma_y^2/m)$

and

$$W = \bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m} - 2\sigma_{\bar{X}\bar{Y}}\right).$$

This implies that

$$\frac{W - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m} - 2\sigma_{\bar{X}\bar{Y}}}} \sim N(0,1)$$

which is all fine & good unless σ_x^2 and σ_y^2 are unknown.

Q) If σ_x^2 & σ_y^2 are unknown, can we still find some pivot statist w/ W?

Recall : For $X_1, \dots, X_n$ IID $N(\mu, \sigma^2)$ where both μ, σ^2 are unknown

$$\frac{\bar{X} - \mu}{sd(\bar{X})} \sim N(0,1) \quad \text{and} \quad \frac{\bar{X} - \mu}{\frac{S_x}{\sqrt{n}}} \sim t_{(n-1)} .$$

↓ See Corollary B Ch. 6 pg. 198

Case 2: Paired Samples

1-Sample T-test & CI

Note: We are starting w/ case 2

If data pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ are $\perp\!\!\!\perp$ independent

and $D_i = X_i - Y_i$ where $X_1, \dots, X_n$ are IID w/

$$E(X_i) = \mu_X$$
$$Var(X_i) = \sigma_X^2$$

and $Y_1, \dots, Y_n$ are IID w/

$$E(Y_j) = \mu_Y$$
$$Var(Y_j) = \sigma_Y^2$$

then

$$E(D_i) = \mu_X - \mu_Y = \mu_D$$
$$Var(D_i) = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}$$
$$= \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y = \sigma_D^2$$

and $D_1, \dots, D_n$ are independent.

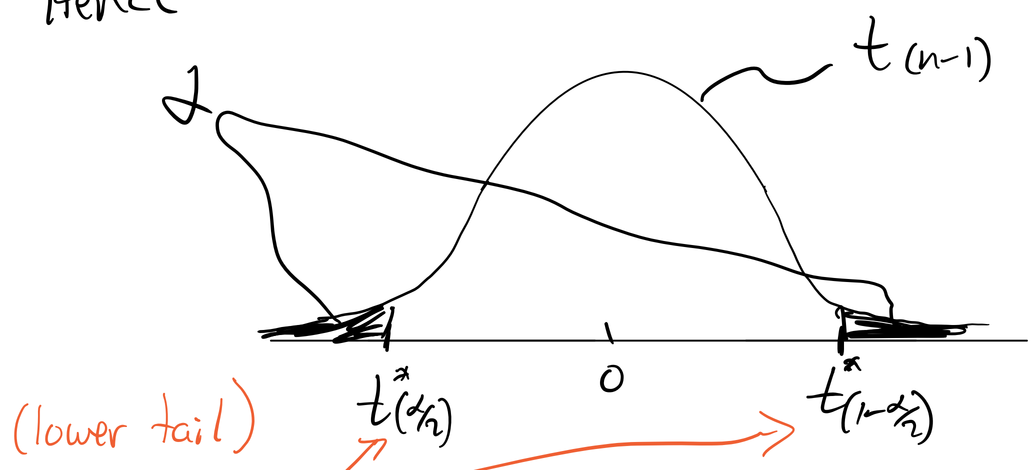
Furthermore, If $D_1, \dots, D_n$ are Normally distb'ted

then

$$t = \frac{\bar{D} - \mu_D}{S_{\bar{D}}} \sim t_{(n-1)}$$

where $S_{\bar{D}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2}$ is an estimate for σ_D.

Hence



note : $t_{(\alpha/2)} = -t_{(1-\alpha/2)}$ by symmetry

and we have that

$$1-\alpha = \Pr\left(t^*_{(\alpha/2)} \le \frac{\bar{D}-\mu_D}{s_{\bar{D}}} \le t^*_{(1-\alpha/2)}\right)$$

$$= \Pr\left(t^*_{(\alpha/2)} s_{\bar{D}} \le \bar{D}-\mu_D \le t^*_{(1-\alpha/2)} s_{\bar{D}}\right)$$

$$= \Pr\left(-\bar{D} + t^*_{(\alpha/2)} s_{\bar{D}} \le -\mu_D \le -\bar{D} + t^*_{(1-\alpha/2)} s_{\bar{D}}\right)$$

$$= \Pr\left(\bar{D} - t^*_{(\alpha/2)} s_{\bar{D}} \ge \mu_D \ge \bar{D} - t^*_{(1-\alpha/2)} s_{\bar{D}}\right)$$

$$= \Pr\left(\bar{D} + \boxed{t^*_{(1-\alpha/2)}} s_{\bar{D}} \ge \mu_D \ge \bar{D} - t^*_{(1-\alpha/2)} s_{\bar{D}}\right)$$

for small $\alpha < 1/2$ this is > 0

Q) What is random?

So a $(1-\alpha)100\%$ CI for μ_D is:

$$\bar{D}_{obs} \pm \left[t^*_{(1-\alpha/2;\, df=n-1)} \times S_{\bar{D}} \right]$$

And an α-level significance test of
$H_0: \mu_D = 0$ vs. $H_1: \mu_D \neq 0$

using test statistic $t = \dfrac{\bar{D} - \mu_D}{S_{\bar{D}}} \overset{H_0}{\sim} t_{(n-1)}$

will reject H_0 for

$$\left\{ \underset{\sim}{D} : |\bar{D}| > t^*_{(1-\alpha/2;\, df=n-1)} \times S_{\bar{D}} \right\}.$$

Case 1: Independent Samples

Two-Sample T-test & CI for $(\mu_x-\mu_y)$

Now, suppose any of the $X_1,\dots,X_n$ are $\perp\!\!\!\perp$ of any of the $Y_1,\dots,Y_m$ and m may differ from n.

We still have $W \sim N\left(\mu_x-\mu_y, \frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}-\cancel{2\sigma_{XY}}\right)$.

and

$$\frac{W-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}-\cancel{2\sigma_{XY}}}} \sim N(0,1)$$

Which we could use to find a test or CI if we knew σ_x^2 and σ_y^2.

$$W=\bar{X}-\bar{Y}$$

$X_1,\dots,X_n$
$Y_1,\dots,Y_m$

Q) How can we estimate $Var(W)$, (which is a weighted average of each samples variance)?

One idea is to approximate

$$\mathrm{Var}(W) = \mathrm{Var}(\bar{X}-\bar{Y}) \approx \frac{S_x^2}{n} + \frac{S_y^2}{m}$$

This is helpful if $\sigma_x \neq \sigma_y$ but it is challenging to find the distb'n of:

$$\frac{W-(\mu_x-\mu_y)}{\sqrt{S_x^2/n + S_y^2/n}} \approx t_{(\nu)}$$

where $$\nu = \frac{\left[\left(S_x^2/n\right) + \left(S_y^2/m\right)\right]}{\frac{S_x^2/n}{n-1} + \frac{S_y^2/m}{m+1}}$$

is rounded to the nearest integer.

(called Satterwaite's approximation)

Q) What would an ethical stat. practioner do before using this to conduct a test or find a CI for $\mu_x - \mu_y$?

It turns out that, if we can assume $\sigma_X = \sigma_Y$, then using

$$Var(W) = Var(\bar{X} - \bar{Y}) \approx S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)$$

where

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}$$

yields* a pivot statistic:

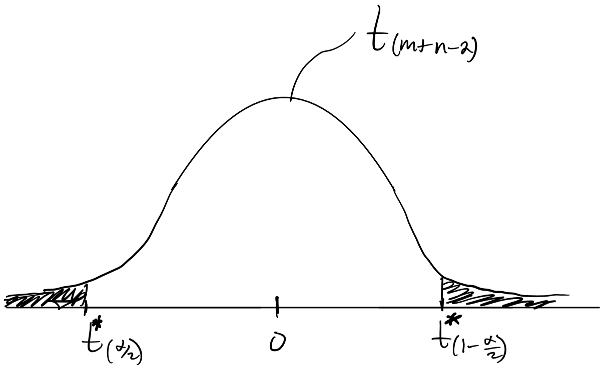
$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{(m+n-2)}$$

*read proof in textbook pg. 422

Useful Identity

$$\sum(X_i - \bar{X})^2 + \sum(Y_j - \bar{Y})^2 + \frac{mn}{m+n}(\bar{X} - \bar{Y})^2 = \sum(X_i - \hat{\mu}_0)^2 + \sum(Y_j - \hat{\mu}_0)^2$$

where $\hat{\mu}_0 = \frac{1}{m+n}\left(\sum X_i + \sum Y_j\right)$



And so

$$1-\alpha = \Pr\left(t^*_{(\alpha/2)} \leq \frac{(\overline{X}-\overline{Y}) - (\mu_x - \mu_y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \leq t^*_{(1-\alpha/2)} \right)$$

$$= \Pr\left(t^*_{(\alpha/2)} * S_p\sqrt{\frac{1}{n}+\frac{1}{m}} \leq (\overline{X}-Y) - (\mu_x-\mu_y) \leq t^*_{(1-\alpha/2)} * S_p\sqrt{\frac{1}{n}+\frac{1}{m}} \right)$$

$$\vdots$$

$$= \Pr\left((\bar{x}-\overline{Y}) + t^*_{(1-\frac{\alpha}{2})} * S_p\sqrt{\frac{1}{n}+\frac{1}{m}} \geq (\mu_x - \mu_y) \geq (\bar{x}-\overline{Y}) - t^*_{(1-\frac{}{2})} \sqrt{\frac{1}{n}+\frac{1}{m}} \right)$$

Thus for $(\bar{x}_{obs} - \bar{y}_{obs})$, a $(1-\alpha)100\%$ CI for $(\mu_x - \mu_y)$ is:

$$(\bar{x}_{obs} - \bar{y}_{obs}) \pm \left[t^*_{(1-\frac{\alpha}{2})} * S_p\sqrt{\frac{1}{n}+\frac{1}{m}} \right]$$

And for $H_0: \mu_x - \mu_y = 0$ w/ one of

$H_1: \mu_x - \mu_y > 0$ or $H_2: \mu_x - \mu_y < 0$ or $H_3: \mu_x - \mu_y \neq 0$

as the alternative, we can use the test statistic

$$t = \frac{(\bar{X} - \bar{Y}) - 0}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \overset{H_0}{\sim} t_{(n+m-2)}$$

w/ corresponding rejection regions:

$H_1: \{(\underset{\sim}{x}, \underset{\sim}{y}) : t > \boxed{t_{(1-\alpha)}}\}$ or $H_2: \{(\underset{\sim}{x}, \underset{\sim}{y}) : t < \boxed{t_{(\alpha)}}\}$ or

$H_3: \{(\underset{\sim}{x}, \underset{\sim}{y}) : |t| > t_{(1-\frac{\alpha}{2})}\}$

for small $\alpha < 0.5$ this will be > 0

for small α this will be < 0

Note:

This is actually the (generalized) LHR test for $H_0: \mu_x - \mu_y = 0$ vs. $H_1: \mu_x - \mu_y \neq 0$!

[See proof on pg 4/26]

Power of a 2-Sample t-test

$$\text{Power} = \Pr(\text{Reject } H_0 \,/\, H_1 \text{ is true})$$

power analysis is a crucial part of planning an experiment. Typically, this involves solving "sample size determination" questions, before any data is collected.

Useful facts about the power of this test:

1) The larger the true difference $|\mu_x - \mu_y|$ the greater the power

2) The larger the sig. level α, the greater the power

3) The larger the sample sizes n & m, the greater the power.

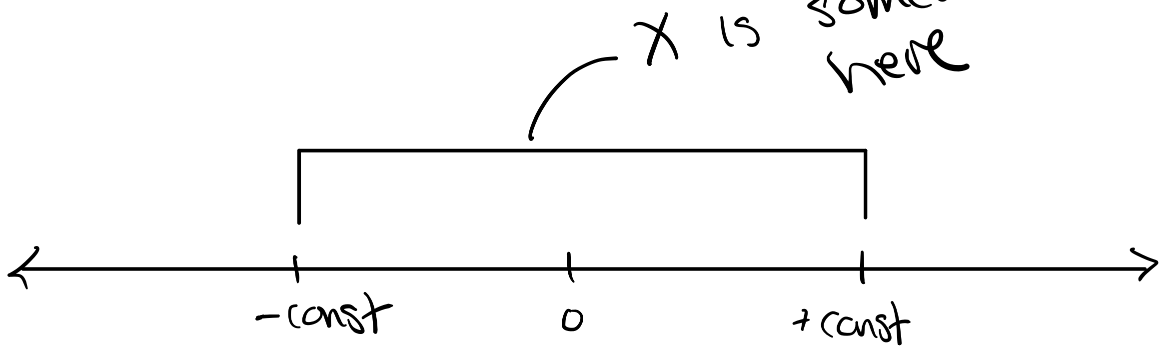
4) If $Var(X_i) = Var(Y_j) = \sigma^2$, the smaller the value of σ^2 corresponds to greater power.

See pg 433 & 434 for power calculation assuming $n = m$ and that n is large enough for the CLT to apply

Detour : Absolute Value Heuristic

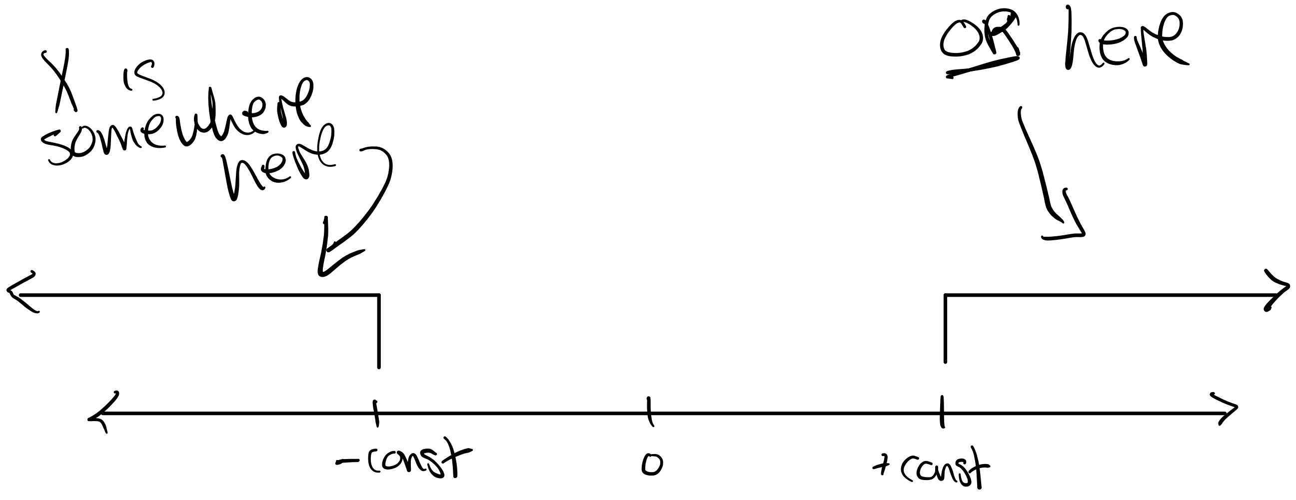
Less than - less thand :

If $|X| \leq \text{const}$ then



Great OR than :

If $|X| \geq \text{const}$ then



11-4-22

Topic: Experimental Design (ch 11.4)

"The proper design of a scientific study is far more important than the specific technique used in the analysis."

"a well-designed study is typically simple to analyze... a poorly-designed study or a botched expt often cannot be salvaged even w/ the most sophisticated analysis"

[ALSM pg 642]

Types of Statistical Studies:

		Random Sampling	
		yes	no
Random assignment	no	observational association, population	association, sample
	yes	experiment causal, population	causal, sample

	experiment	observational
data	experimental units	observational units
variables	experimental group(s) treatment levels • control group • treatments assigned via randomization • blinding	observational factor(s) simple (one variable) or comparative (> 1 varb) • randomization of treatment levels is not possible (ethical and/or practical reasons)
examples	completely randomized, factorial, repeated measures, nested design	cross-sectional, retrospective, prospective/cohort studies

The random assignment of treatment levels also randomizes confounding variables, creating a strong basis for cause-and-effect conclusions.

No random assignments means possible confounders are not controlled.

Study requires additional external evidence before cause-and-effect conclusions are justified.

Also, studies can be a mixture of <u>both</u> experiments + observational studies

ex) blocked experimental study

Random Sampling Selection

This is what is meant by "random sample of a population". In theory, random sampling is the best way to ensure the sample of data is representative of the population of interest. Random sampling mitigates any overt or unintentional selection bias and ensures any confounding features are also randomly distributed throughout the sample.

In practice, random selection is rarely possible and various sampling strategies are used to obtain

"pseudo-random" samples

or "representative" samples

instead. A complete discussion of thes sampling strategies is beyond the scope of this class but it is crucial to be aware of this because most statistical theory relies upon the assumption of IID (random) data.

Non-parametric Approaches

Case 1: Independent Samples

The Mann-Whitney Test
(AKA Wilcoxon Rank Sum test)

If $\mu_x - \mu_y = 0$ is actually true, and if data are randomly assigned a treatment, then any observed difference in $\bar{X}_{obs} - \bar{y}_{obs}$ is due to chance/luck and not the treatment.

Procedure:

1) Group both samples together to rank from least to greatest.

2) Add the ranks of each data value that is from the first treatment group.

3) We have evidence against $H_0: \mu_x - \mu_y = 0$ if the summed rank is extreme.

Case 2: Paired Samples

The Signed-Rank Test
(AKA Wilcoxon signed rank test)

Using the same idea as for the Mann-Whitney test, we consider ranks of the data rather than the observed values of the data.

Proceedure:

1) Calculate the vecter of paired differences $\underset{\sim}{D} = (D_1, \dots, D_n)$
   then rank the magnitude of the differences from least to greatest

2) Calculate the sum of the (magnitude) ranks of $\underset{\sim}{D}$ that are positive.

If $\mu_x - \mu_y = 0$ is correct, then we'd expect about half of the differences to be positive, and half negative. The sum in (2) will not be too extreme in this case.

Group Worksheet + Review

Topic: Comparing > 2 means w/ ANOVA

11-7-22

One-Way ANOVA

Setting + Notation:

Data

→ qualitative/categorical "treatment levels" $(i = 1, \dots, I)$ α_i

→ numeric, continuous response/measurement $(j = 1, \dots, J_i)$ Y_{ij}

RV! (noise)

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \text{where } \varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

$$\text{and } \sum_{i=1}^{I} \alpha_i = 0$$

unknown constant "error variance"

j^{th} observation of i^{th} trt.

i^{th} trt "effect"

overall mean

So $E[Y_{ij}] = \mu + \alpha_i$ and $\alpha_i - \alpha_k = \Big\{$ difference btwn expected values of the response under trt i and k

We say the model is balanced when

$$J_1 = J_2 = \dots = J_I = J$$

Our discussion will focus on the balanced 1-way ANOVA model. The unbalanced model is similar but requires careful notation.

Analysis of Variance:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{..})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i.})^2 + J\sum_{i=1}^{I}(\overline{Y}_{i.}-\overline{Y}_{..})^2$$

see pg. 479 for proof

$$SS_{Tot} = SS_{Trt} + SS_{Error}$$

$$= SS_W + SS_B$$

w/in trt levels — between/among trt levels

$$\overline{Y}_{i.} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$$

$$\overline{Y}_{..} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}$$

And we estimate $Var(\varepsilon_{ij}) = \sigma^2$ w/ S_p^2 where:

$$S_p^2 = \frac{SS_W}{I(J-1)}$$

or equivalently,

$$S_p^2 = \frac{1}{I}\sum_{i=1}^{I} S_i^2$$ where S_i^2 = sample variance of i^{th} group

(Balanced) ANOVA Table:

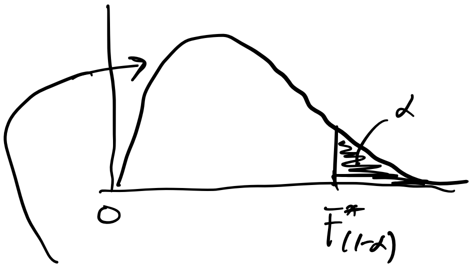
Source (of variation)	df	Sum of Squares	Mean Squares	F-statistic
Treatment	$I-1$	$SS_W = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i.})^2$	$MS_W = \frac{SS_W}{df_W}$	$F = \frac{MS_B}{MS_W}$
Error	$I(J-1)$	$SS_B = J\sum_{i=1}^{I}(\overline{Y}_{i.}-\overline{Y}_{..})^2$	$MS_B = \frac{SS_B}{df_B}$	
Total	$IJ-1$	$SS_{Tot} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{..})^2$		

These were swapped (incorrectly) in a previous version of the notes

Overall F-test :

$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_I = 0$

H_1 : At least one α_i is not equal to zero.



Test statistic : $F = \frac{MS_B}{MS_W} = \frac{SS_B/(I-1)}{SS_W/(I(J-1))} \overset{H_0}{\sim} F_{(I-1, I(J-1))}$

Rejection Region : If Ho is true then $\frac{SS_B}{I-1} \approx \frac{SS_W}{I(J-1)}$

If Ho is false then $\frac{SS_B}{I-1} > \frac{SS_W}{I(J-1)}$

so $A_\alpha = \left\{ y_{ij} : \frac{SS_B/I-1}{SS_W/I(J-1)} > F^*_{(1-\alpha; I-1, I(J-1))} \right\}$

Theorem:

If $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$, $\sum_{i=1}^{I} \alpha_i = 0$

then $E(SS_W) = I(J-1)\sigma^2$

$E(SS_B) = J \cdot \sum_{i=1}^{I} \alpha_i^2 + (I-1)\sigma^2$

Proof

- $E(SS_W) = E\left[\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i.})^2\right] = \sum_{i=1}^{I}\sum_{j=1}^{J} E(Y_{ij} - \bar{Y}_{i.})^2 = \overset{(*)}{\cdots} = I(J-1)\sigma^2$
- $E(SS_B) = E\left[J\sum_{i=1}^{I}(\bar{Y}_{i.} - \bar{Y}_{..})^2\right] = J\sum_{i=1}^{I} E(Y_{i\cdot} - \bar{Y}_{..})^2$

$= \cdots (*)$

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

Where (*) is established by lemma A pg. 480

∎

Theorem:

If $\quad Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad , \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \; \sum_{i=1}^{I} \alpha_i = 0$

then $$\frac{SS_W}{\sigma^2} \sim \chi^2_{(I(J-1))}.$$

Furthermore, if $\quad \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$

then $\quad \dfrac{SS_B}{\sigma^2} \sim \chi^2_{(I-1)} \quad$ and $\quad \dfrac{SS_B}{\sigma^2} \perp\!\!\!\perp \dfrac{SS_W}{\sigma^2}$.

Proof

Part 1:

*Relevant: HW 4, ch. 6

Since $\left[\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)} \right.$ for $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ where $\left. X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \right]^*$

we have that
$$\frac{1}{\sigma^2}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i\cdot})^2 \sim \chi^2_{(J-1)} \sim \text{Gamma}\left(\frac{J-1}{2}, \frac{1}{2}\right).$$

If $\quad W = \frac{1}{\sigma^2}\chi^2_{(J-1)} \quad$ then $\quad W \sim \text{Gamma}\left(\frac{J-1}{2}, \frac{\sigma^2}{2}\right)$

and $\quad \sum_{i=1}^{I} W_i \sim \text{Gamma}\left(I \cdot \frac{J-1}{2}, \frac{\sigma^2}{2}\right)$ for iid W_i.

Hence $$\frac{SS_W}{\sigma^2} = \frac{\sum_{i=1}^{I}\left[\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i\cdot})^2\right]}{\sigma^2} \sim \chi^2_{(I(J-1))}.$$

Part 2:

Note that

$$\begin{aligned} Var(\overline{Y}_{i\cdot}) &= Var\left(\frac{1}{J}\sum_{j=1}^{J} Y_{ij}\right) \\ &= \left(\frac{1}{J}\right)^2 \sum_{j=1}^{J} Var(Y_{ij}) \quad \text{since } Y_{ij} \text{ are indep} \\ &= \frac{1}{J} Var(Y_{11}) \quad \text{since } Y_{ij} \text{ are identically distbted} \\ &= \frac{1}{J} Var(\mu + \alpha_1 + \varepsilon_1) \\ &= \frac{1}{J} Var(\varepsilon_1) \\ &= \frac{\sigma^2}{J}. \end{aligned}$$

Using the same fact as in Part 1, if $\alpha_1 = \ldots = \alpha_I = 0$ then $Y_{ij} \overset{iid}{\sim} N(\mu, \sigma^2)$ and we have that

$$\frac{\sum_{i=1}^{I} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}{\sigma^2/J} \sim \chi^2_{(I-1)} .$$

Hence,

$$\frac{SS_B}{\sigma^2} = \frac{J\sum_{i=1}^{I} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}{\sigma^2} \sim \chi^2_{(I-1)} .$$

Part 3

It suffices to show that data vectors

$$[(Y_{1j}-\bar{Y}_{1\cdot}), (Y_{2j}-\bar{Y}_{2\cdot}), \ldots, (Y_{Ij}-\bar{Y}_{I\cdot})]$$

and

$$[\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \ldots, \bar{Y}_{I\cdot}]$$

are independent for each $j=1,\ldots,J$.

For any $i_1 \neq i_2$, $Y_{i_1 j}-\bar{Y}_{i_1\cdot} \perp\!\!\!\perp \bar{Y}_{i_2\cdot}$ since these are functions of different, independent error terms, ε_{ij}.

Recall, for $X_1,\ldots,X_n$ IID $N(\mu,\sigma^2)$, we have that $\bar{X} \perp\!\!\!\perp [(X_1-\bar{X}),(X_2-\bar{X}),\ldots,(X_n-\bar{X})]$.*

Hence each $\bar{Y}_{i\cdot}$ is independent of

$[(Y_{1j}-\bar{Y}_{1\cdot}), (Y_{2j}-\bar{Y}_{2\cdot}), \ldots, (Y_{Ij}-\bar{Y}_{I\cdot})]$ and the result holds.

$\blacksquare$

(*) Ch. 6 Thm A

Multiple Comparisons

If we reject $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_I = 0$, we still do not know

- how many treatment effects are non-zero
- which treatment effects are non-zero

One idea is to use the 2-sample t-test (or confidence interval) to test each

$$H_0: \alpha_i - \alpha_k = 0 \quad \text{for all } i \neq k \in \{1, 2, \dots, I\}$$

however, if we proceed w/out making any adjustments, the more tests we run on the same data set inflates the probability of a Type I error.

Some of the earliest adjustments for multiplicity include Tukey's method and the Bonferroni correction. Although these methods vary greatly in terms of their usefulness and there are more sophisticated ways to adjust for multiple comparisons today, they are are relatively straightforward and provide a foundation for understanding modern methods of error control.

Tukey's Method for simultaneous $(1-\alpha)100\%$ confidence intervals estimates each $(\alpha_i - \alpha_k)$ w/

$$(\bar{Y}_{i\cdot} - \bar{Y}_{k\cdot}) \pm \left[\underbrace{q_{I,I(J-1)}}^{(\alpha)} \times \frac{S_p}{\sqrt{J}} \right]$$

$\uparrow$

This is an α^{th} quantile of the "studentized range" distb'n

Bonferroni Adjustment for simultaneous $(1-\alpha)100\%$ confidence interval estimates each $(\alpha_i - \alpha_k)$ w/

$$(\bar{Y}_{i\cdot} - \bar{Y}_{k\cdot}) \pm \left[t^{*}_{I(J-1)}\left(\frac{\alpha}{m}\right) \times \frac{S_p}{\sqrt{J}} \right] \quad \text{where } m = \binom{I}{2}$$

Generalized LHR: 1 Samp T-test

11-9-22

Ex) Suppose $X_1,\dots,X_n \overset{iid}{\sim} N(\mu,\sigma^2)$ where μ,σ^2 both unknown

Show that the generalized LHR test of $H_0: \mu=\mu_0$ vs. $H_1: \mu\neq\mu_0$ is the same as the one-sample t-test

$$\text{Likelihood}: L(\mu,\sigma^2) = \prod_{i=1}^{n} \underbrace{\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sigma}\cdot\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}}_{\text{density for } X_i}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n\cdot\left(\frac{1}{\sigma}\right)^n\cdot\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

$$\omega_0 = \{\mu_0\}\cup(0,\infty) \qquad \Omega = (-\infty,\infty)\cup(0,\infty)$$

$$\Lambda^* = \frac{\max_{\mu,\sigma\in\omega_0} L(\mu,\sigma^2)}{\max_{\mu,\sigma\in\Omega} L(\mu,\sigma^2)}$$

we need $\hat{\sigma}^2_{MLE}$ in numerator & denominator and $\hat{\mu}_{MLE}$ in denominator

Find the MLEs:

$$\ell(\mu,\sigma^2) = \ln\left(\left(\frac{1}{\sqrt{2\pi}}\right)^n\right) + n\ln\left(1/\sigma\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

$$= n\ln\left(1/\sqrt{2\pi}\right) + n\left(\ln(1) - \ln(\sigma)\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

$$= n\ln\left(1/\sqrt{2\pi}\right) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{\partial}{\partial\mu}\ell(\mu,\sigma) = \frac{-1}{2\sigma^2}\sum_{i=1}^{n}2(X_i-\mu)\cdot -1 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \overset{\text{set}}{=} 0$$

solve for $\hat{\mu}$: $\sum_{i=1}^{n}(x_i-\mu) = 0$

$$\Rightarrow \sum x_i - n\mu = 0 \Rightarrow \hat{\mu}_{MLE} = \bar{X}$$

$$\frac{\partial}{\partial\sigma}\ell(\mu,\sigma) = \frac{-n}{\sigma} - \frac{1}{2}\left(-2\sigma^{-3}\right)\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum(x_i-\mu)^2 \overset{\text{set}}{=} 0$$

solve for $\hat{\sigma}^2$ (subbing in $\hat{\mu}$): $\sum(X_i-\hat{\mu}_{MLE})^2 = \frac{n}{\sigma}\sigma^3$

$$\Rightarrow \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum(X_i-\hat{\mu}_{MLE})^2$$

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

So now we can evaluate the LHR test statistic.

$(x_i-\bar{x})^2$

$$\Lambda^* = \frac{\cancel{\left(\frac{1}{\sqrt{2\pi}}\right)^n} \cancel{\left(\frac{1}{\hat\sigma_{mle}}\right)^n} \exp\left\{-\frac{1}{2\hat\sigma^2_{mle}} \sum_{i=1}^n (x_i-\mu_0)^2\right\}}{\cancel{\left(\frac{1}{\sqrt{2\pi}}\right)^n} \cancel{\left(\frac{1}{\hat\sigma_{mle}}\right)^n} \exp\left\{-\frac{1}{2\hat\sigma^2_{mle}} \sum_{i=1}^n (x_i-\hat\mu_{mle})^2\right\}} = \ldots = \exp\left\{\frac{-1}{2}\cdot\frac{1}{\frac{1}{n}\Sigma(x_i-\bar x)^2}\left[\Sigma(x_i-\mu_0)^2 - \Sigma(x_i-\bar x)^2\right]\right\}$$

sub $\hat\mu = \bar x$, $\hat\sigma^2 = \frac{1}{n}\Sigma$

So the rejection region is

$$A_\alpha = \left\{ \underset{\sim}{x} : \exp\left\{\frac{-1}{2}\cdot\frac{1}{\frac{1}{n}\Sigma(x_i-\bar x)^2}\left[\Sigma(x_i-\mu_0)^2 - \Sigma(x_i-\bar x)^2\right]\right\} < C_\alpha \right\}$$

where C_α solves

$$\alpha = \Pr\left(\exp\left\{\frac{-1}{2}\cdot\frac{1}{\frac{1}{n}\Sigma(x_i-\bar x)^2}\left[\Sigma(x_i-\mu_0)^2 - \Sigma(x_i-\bar x)^2\right]\right\} < C_\alpha \,\middle|\, H_0\right)$$

$$= \Pr\left(\frac{-1}{2}\cdot\frac{1}{\frac{1}{n}\Sigma(x_i-\bar x)^2}\left[\Sigma(x_i-\mu_0)^2 - \Sigma(x_i-\bar x)^2\right] < c'_\alpha \,\middle|\, H_0\right)$$

$$= \Pr\left(\frac{\left[\Sigma(x_i-\mu_0)^2 - \Sigma(x_i-\bar x)^2\right]}{\frac{1}{n}\Sigma(x_i-\bar x)^2} > c''_\alpha \,\middle|\, H_0\right)$$

$$= \Pr\left(n\left[\frac{\sum_{i=1}^n (x_i-\mu_0)^2}{\sum_{i=1}^n (x_i-\bar x)^2} - 1\right] > c''_\alpha \,\middle|\, H_0\right)$$

$$= \Pr\left(n\,\frac{\Sigma(x_i \boxed{-\bar x}\boxed{+\bar x} - \mu_0)^2}{\Sigma(x_i-\bar x)^2} > c''_\alpha \,\middle|\, H_0\right)$$

creative form of zero

careful when applying the sum!

$$= \Pr\left(n\,\frac{\Sigma\left[(x_i-\bar x)^2 + 2(x_i-\bar x)(\bar x-\mu_0) + (\bar x-\mu_0)^2\right]}{\Sigma(x_i-\bar x)^2} \,\middle|\, H_0\right)$$

$$\vdots$$

make sure you understand why!

$$= \Pr\left(n\left[1 + \frac{0}{\Sigma(x_i-\bar x)^2} + \frac{n(\bar x-\mu_0)^2}{\Sigma(x_i-\bar x)^2}\right] > c''_\alpha \,\middle|\, H_0\right)$$

Thus far we have

$$\alpha = \Pr\left(\frac{n^2(\bar{x}-\mu_0)^2}{\sum(x_i-\bar{x})^2} > C_\alpha''' \,\middle|\, H_0\right)$$

$$= \Pr\left(\frac{(\bar{x}-\mu_0)^2}{\sum(x_i-\bar{x})^2} > c_\alpha'''' \,\middle|\, H_0\right)$$

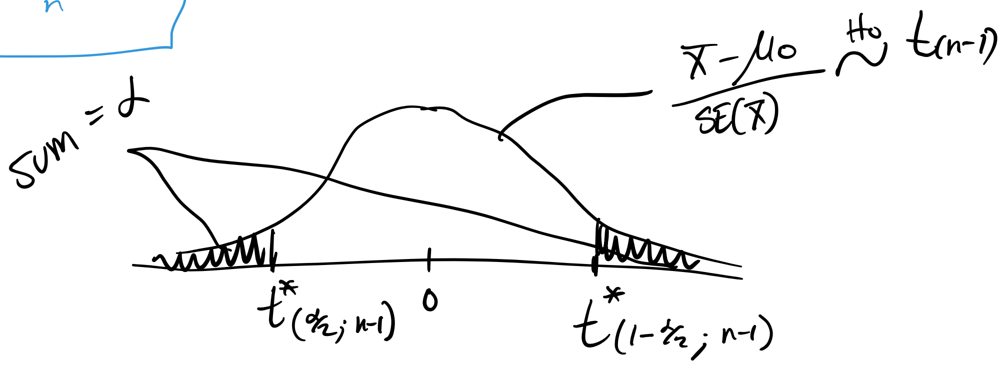
make sure you understand this part!

$$= \Pr\left(\left|\frac{\bar{x}-\mu_0}{\sqrt{\sum(x_i-\bar{x})^2}}\right| > C_\alpha^{\text{卌}'} \,\middle|\, H_0\right)$$

$$= \Pr\left(\boxed{\sqrt{n}\left|\frac{\bar{x}-\mu_0}{\sqrt{\sum(x_i-\bar{x})^2}}\right|} > C_\alpha^{\text{卌}''} \,\middle|\, H_0\right)$$

This is the test statistic for the 1-sample t-test!

Note:
$SE(\bar{x}) = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$



Thus $C_\alpha^{\text{卌}''} = t^*_{(1-\alpha/2;\, n-1)}$ and we're done!

∎

Generalized LHR : ANOVA overall F-test

Ex) (from HW 18)

Suppose $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ where $\sum_{i=1}^{I} \alpha_i = 0$

and $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

For $i = 1, \dots, I$

and $j = 1, \dots, J$.

Show that the generalized LHR test of

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_I = 0$ vs. H_1: Not H_0

is the same as the ANOVA overall F-test.

Find the likelihood:

$$L(\mu, \alpha_1, \dots, \alpha_I, \sigma^2) = \prod_{i=1}^{I} \prod_{j=1}^{J} \overbrace{\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2\sigma^2} (y_{ij} - (\mu + \alpha_i))^2 \right\}}^{\text{density for } y_{ij}}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^{IJ} \left(\frac{1}{\sigma}\right)^{IJ} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - (\mu + \alpha_i))^2 \right\}$$

$$\Omega = (-\infty, \infty) \cup \overbrace{(-\infty, \infty) \cup \dots \cup (-\infty, \infty)}^{I \text{ of these}} \cup (0, \infty)$$

$$\omega_0 = (-\infty, \infty) \cup \underbrace{\{0\} \cup \dots \cup \{0\}}_{I \text{ times}} \cup (0, \infty)$$

Find the MLE's:

$$\hat{\mu}_{MLE} = \bar{y}_{\cdot\cdot} \; ; \quad \hat{\alpha}_{i,MLE} = \bar{y}_{i\cdot} \; ; \quad \hat{\sigma}^2_{MLE} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_{i\cdot})^2$$

Note: $\bar{y}_{\cdot\cdot} = \frac{1}{IJ} \sum_i \sum_j y_{ij}$ & $\bar{y}_{i\cdot} = \frac{1}{J} \sum_{j=1}^{J} y_{ij}$

Find the LHR test statistic:

$$\Lambda^* = \frac{L(\hat{\mu}_{MLE}, \alpha_1 = \dots = \alpha_I = 0, \hat{\sigma}^2_{MLE})}{L(\hat{\mu}_{MLE}, \hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_I, \hat{\sigma}^2_{MLE})}$$

$$= \frac{\cancel{\left(\frac{1}{\sqrt{2\pi}}\right)^{IJ}} \cancel{\left(\frac{1}{IJ \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}\right)^{IJ}} \exp\left\{-\frac{1}{2}\left(\frac{1}{IJ \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}\right) \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \bar{y}_{\cdot\cdot})^2\right\}}{\cancel{\left(\frac{1}{\sqrt{2\pi}}\right)^{IJ}} \cancel{\left(\frac{1}{IJ \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}\right)^{IJ}} \exp\left\{-\frac{1}{2}\left(\frac{1}{IJ \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}\right) \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - (\bar{y}_{\cdot\cdot} + \bar{y}_{i\cdot}))^2\right\}}$$

$$= \exp\left\{-\frac{1}{2}\left[\frac{\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2 - \sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot} - \bar{y}_{i\cdot})^2}{IJ \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}\right]\right\}$$

Find the rejection region:

$$A_\alpha = \left\{ \underline{y} : \exp\left\{-\frac{1}{2}\left[\frac{\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2 - \sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot} - \bar{y}_{i\cdot})^2}{IJ \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}\right]\right\} < c_\alpha \right\}$$

where c'_α solves

$$\alpha = P_r\left(\exp\left\{-\frac{1}{2}\left[\frac{\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2 - \sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot} - \bar{y}_{i\cdot})^2}{IJ \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}\right]\right\} < c_\alpha \,\middle|\, H_0 : \alpha_1 = \dots = \alpha_I = 0\right)$$

Now lets find equivalant probability statements until we recognize the distribution of the function of the data!

$$\alpha = P_r\left(\exp\left\{-\frac{1}{2}\left[\frac{\sum_i\sum_j(y_{ij}-\bar{y}_{\cdot\cdot})^2-\sum_i\sum_j(y_{ij}-\bar{y}_{\cdot\cdot}-\bar{y}_{i\cdot})^2}{IJ\sum_i\sum_j(y_{ij}-\bar{y}_{i\cdot})^2}\right]\right\}< c_\alpha \,\middle|\, H_0: \alpha_1=\dots=\alpha_I=0\right)$$

$$\vdots$$

$$= P_r\left(\frac{\sum\sum\left((y_{ij}-\bar{y}_{\cdot\cdot})^2-(y_{ij}-\bar{y}_{\cdot\cdot}-\bar{y}_{i\cdot})^2\right)}{\frac{1}{IJ}\sum_i\sum_j(Y_{ij}-\bar{Y}_{i\cdot})^2} > c'_\alpha \,\middle|\, H_0\right)$$

$$= P_r\left(\frac{\sum\sum\left[(y_{ij}-\bar{y}_{\cdot\cdot})^2-(y_{ij}-\bar{y}_{\cdot\cdot}-\bar{y}_{i\cdot})^2\right]}{\sum\sum(y_{ij}-\bar{y}_{i\cdot})^2} > c''_\alpha \,\middle|\, H_0\right)$$

$$= P_r\left(\frac{\sum\sum\left[(y_{ij}\boxed{-\bar{y}_{i\cdot}+\bar{y}_{i\cdot}}-\bar{y}_{\cdot\cdot})^2-(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2\right]}{\sum\sum(y_{ij}-\bar{y}_{i\cdot})^2} > c''_\alpha \,\middle|\, H_0\right)$$

let's take a careful look at the numerator:

$$\sum\sum\left[\left(y_{ij}\boxed{-\bar{y}_{i.}+\bar{y}_{i.}}-\bar{y}_{..}\right)^2-\left(y_{ij}-\bar{y}_{i.}-\bar{y}_{..}\right)^2\right]$$

$$=\sum\sum\Big[(y_{ij}-\bar{y}_{i.})^2+2(y_{ij}-\bar{y}_{i.})(\bar{y}_{i.}-\bar{y}_{..})+(\bar{y}_{i.}-\bar{y}_{..})^2 -\left[(y_{ij}-\bar{y}_{i.})^2-2\bar{y}_{..}(y_{ij}-\bar{y}_{i.})+\bar{y}_{..}^2\right]\Big]$$

$$=\sum\sum\Big[\boxed{(y_{ij}-\bar{y}_{i.})^2}+2(\bar{y}_{i.}-\bar{y}_{..})(y_{ij}-\bar{y}_{i.})+(\bar{y}_{i.}-\bar{y}_{..})^2 \boxed{-(y_{ij}-\bar{y}_{i.})^2}+2\bar{y}_{..}(y_{ij}-\bar{y}_{i.})-\bar{y}_{..}^2\Big]$$

$$=\sum\sum\left[2(\bar{y}_{i.}-\bar{y}_{..})(y_{ij}-\bar{y}_{i.})+(\bar{y}_{i.}-\bar{y}_{..})^2+2\bar{y}_{..}y_{ij}-2\bar{y}_{..}\bar{y}_{i.}-\bar{y}_{..}^2\right]$$

$$=2\sum_i\left[(\bar{y}_{i.}-\bar{y}_{..})\sum_j(y_{ij}-\bar{y}_{i.})\right]+\sum_i\sum_j(\bar{y}_{i.}-\bar{y}_{..})^2 +2\bar{y}_{..}\underbrace{\boxed{\sum_i\sum_j y_{ij}}}_{IJ\bar{y}_{..}}-2J\bar{y}_{..}\underbrace{\boxed{\sum_i\bar{y}_{i.}}}_{\sum_i\frac{1}{J}\sum_j y_{ij}}-IJ\bar{y}_{..}^2$$

$$=2\sum_i\left[(\bar{y}_{i.}-\bar{y}_{..})\sum_j(y_{ij}-\bar{y}_{i.})\right]+\sum\sum(\bar{y}_{i.}-\bar{y}_{..})^2-IJ\bar{y}_{..}^2$$

$$=2\sum_i\left[(\bar{y}_{i.}-\bar{y}_{..})\sum_j(y_{ij}-\bar{y}_{i.})\right]-IJ\bar{y}_{..}^2+\sum\sum(\bar{y}_{i.}-\bar{y}_{..})^2$$

$$=2\sum_i\left[\bar{y}_{i.}\sum_j(y_{ij}-\bar{y}_{i.})-\bar{y}_{..}\sum_j(y_{ij}-\bar{y}_{i.})-\frac{J}{2}\bar{y}_{..}^2\right]+\sum\sum(\bar{y}_{i.}-\bar{y}_{..})^2$$

$$=2\sum_i\Big[\bar{y}_{i.}\big(\underbrace{\sum_j y_{ij}-\sum_j y_{ij}}_{0}\big)-\bar{y}_{..}\big(\underbrace{\sum_j y_{ij}-\sum_j y_{ij}}_{0}\big)-\frac{J}{2}\bar{y}_{..}^2\Big]+\sum_i\sum_j(\bar{y}_{i.}-\bar{y}_{..})^2$$

$$=IJ\bar{y}_{..}^2+\sum_i\sum_j(\bar{y}_{i.}-\bar{y}_{..})^2$$

Putting this altogether:

$$\alpha = P_r\left(\frac{IJ\bar{y}_{\cdot\cdot}^2 + \sum_i \sum_j (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\sum\sum (y_{ij} - \bar{y}_{i\cdot})^2} > C_\alpha'' \,\middle|\, H_0 \right)$$

$$= P_r\left(\frac{IJ\bar{y}_{\cdot\cdot}^2 + J\sum_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\sum\sum (y_{ij} - \bar{y}_{i\cdot})^2} > C_\alpha'' \,\middle|\, H_0 \right)$$

$$\vdots$$

Please check my work!

$$= P_r\left(\frac{J\,I\,(J-1)}{I-1}\left[\sum_{i=1}^{I} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \Big/ \sum_{I=1}^{I}\sum_{j=1}^{J} (y_{ij} - \bar{y}_{i\cdot})^2 \right] > K_\alpha \,\middle|\, H_0 \right)$$

which is the rejection rule for the F-test!

∎

11-14-22

In Class discussion on HW 19

Choice/Decision:

Stakeholder	**Potential results**	
	Harm	**Benefit**
You		
Your boss/client		
Colleagues/ peers		

- Example harms: cost of money, time, effort; negative impact to reputations; can be tangible or intangible with immediate or delayed effects
- Example benefits: earning or gaining money; removal of a harm; saved time or effort; improved reputation; demonstration of expertise.

Source: Tractenberg, R. E. (2019). Teaching and Learning about ethical practice: The case analysis. https://doi.org/10.31235/OSF.IO/58UMW

11-16-22

Recap - One-Way ANOVA

(balanced) $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$; $i = 1, \dots, I$, $j = 1, \dots, J$ $\Rightarrow n = IJ$

where we assume

1) $\varepsilon_{ij} \sim$ Normal
2) $Var(\varepsilon_{ij}) = \sigma^2$ is an unknown constant
3) Each ε_{ij} is independent of the others

Overall F test for treatment effects $\equiv$ Generalized LHR test

where $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_J = 0$

and H_1: At least one $\alpha_j \neq 0$ for $j = 1, \dots, J$

Multiple Comparisons:

Necessary if you want to control the overall Type I error rate when conducting many tests (or finding many CIs) on the same set of data.

- Tukey's method (AKA Tukey's honest significance difference)
- Bonferroni's method

Non-parametric version of one-way ANOVA F-test:

Kruskal-Wallis test

Topic: 2-Way ANOVA Models

Setting + Notation:

Data
- two categorical factors
  - 1st factor ("main" effect) $\alpha_i, \; i=1,\dots,I$
  - 2nd factor $\beta_j, \; j=1,\dots,J$
- numeric, continuous response/measurement $Y_{ijk}, \; k=1,\dots,K$ where K is the number of observations per combination of factor levels

Model:

$$Y_{ijk} = \mu + \alpha_i + \boxed{\beta_j} + \boxed{\delta_{ij}} + \varepsilon_{ijk}$$

"effect" of 2nd factor varb.

"effect" of an interaction btwn level i and level j

where $$\sum_{i=1}^{I} \alpha_i = 0, \; \sum_{j=1}^{J} \beta_j = 0, \; \sum_{i=1}^{I} \delta_{ij} = \sum_{j=1}^{J} \delta_{ij} = 0$$

and $$\varepsilon_{ijk} \overset{IID}{\sim} N(0, \sigma^2)$$

Thus: $E(Y_{ijk}) = \mu + \alpha_i + \beta_j + \delta_{ij}$

$\mathrm{Var}(Y_{ijk}) = \sigma^2$

Analysis of Variance : $SS_{Tot} = SS_A + SS_B + SS_{AB} + SS_E$

$$SS_{Tot} = \sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{...})^2$$

$$SS_E = \sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{ij.})^2$$

$$SS_A = JK \sum_i (\overline{Y}_{i..} - \overline{Y}_{...})^2$$

$$SS_B = IK \sum_j (\overline{Y}_{.j.} - \overline{Y}_{...})^2$$

$$SS_{AB} = K \sum_i \sum_j (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2$$

And we estimate $Var(\varepsilon_{ij}) = \sigma^2$ w/ SS_E b/c:

$$E(SS_E) = IJ(K-1)\sigma^2$$

if ε_{ijk} are indep. w/ $E(\varepsilon_{ijk}) = 0$ and $Var(\varepsilon_{ijk}) = \sigma^2$.

[Proof see Thm A of Ch. 12.3, pg. 494]

Q) What is the likelihood?

$$Lik(\mu, \underset{\sim}{\alpha}, \underset{\sim}{\beta}, \underset{\sim}{\delta}, \sigma^2) =$$

$$\prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \left(\frac{1}{\sqrt{2\pi\sigma}}\right) exp\left\{\frac{-1}{2\sigma^2}\left(y_{ijk} - [\mu + \alpha_i + \beta_j + \delta_{ij}]\right)^2\right\}$$

Maximum Likelihood Estimates

$$\hat{\mu}_{MLE} = \overline{Y}_{\cdots}$$

$$\hat{\alpha}_i = \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots} \quad , \; i=1,\dots,I$$

$$\hat{\beta}_j = \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots} \quad , \; j=1,\dots,J$$

$$\hat{\delta}_{ij} = \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}$$

2-Way ANOVA Table

Source	df	SS	MS	F
Main facter	$I-1$	$SS_A = JK\sum_i (\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2$	$MS_A = \frac{SS_A}{df_A}$	$\frac{MS_A}{MSE}$
2nd facter	$J-1$	$SS_B = IK\sum_j (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdot\cdot\cdot})^2$	$MS_B = \frac{SS_B}{df_B}$	$\frac{MS_B}{MSE}$
Interaction	$(I-1)(J-1)$	$SS_{AB} = K\sum_i\sum_j (\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdot\cdot\cdot})^2$	$MS_{AB} = \frac{SS_{AB}}{df_{AB}}$	$\frac{MS_{AB}}{MS_E}$
Error	$IJ(K-1)$	$SS_E = \sum\sum\sum (Y_{ijk} - \overline{Y}_{ij\cdot})^2$	$MSE = \frac{SS_E}{df_E}$	
Total	$KIJ-1$	$SS_{Tot} = \sum\sum\sum (Y_{ijk} - \overline{Y}_{\cdot\cdot\cdot})^2$		

Theorem: Model tests for 2-Way, balanced ANOVA

If $Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}$

where $\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$, $\sum_{i=1}^{I} \delta_{ij} = \sum_{j=1}^{J} \delta_{ij} = 0$

and $\varepsilon_{ijk} \overset{\text{IID}}{\sim} N(0, \sigma^2)$

(1) then $\frac{SS_E}{\sigma^2} \sim \chi^2_{(IJ(K-1))}$

and $SS_E \perp\!\!\!\perp SS_A \perp\!\!\!\perp SS_B \perp\!\!\!\perp SS_{AB}$

(2) and if $\alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$ then

$\frac{SS_A}{\sigma^2} \sim \chi^2_{(I-1)}$

(3) and If $\beta_1 = \beta_2 = \cdots = \beta_J = 0$ then

$\frac{SS_B}{\sigma^2} \sim \chi^2_{(J-1)}$

(4) and if $\delta_{11} = \delta_{12} = \cdots = \delta_{1J} = \delta_{21} = \cdots = \delta_{2J} = \cdots \delta_{IJ} = 0$

then

$\frac{SS_{AB}}{\sigma^2} \sim \chi^2_{((I-1)(J-1))}$

Additive Factor Effects

Every mean response for any $i=1,\dots,I$ and $j=1,\dots,J$ can be obtained by adding (or subtracting) the levels' main effects (say, α_I and β_J) to the grand mean.

The relationship the 1st factor has w/ the response varb is independent of the relationship the 2nd factor has w/ the response.

this language is only justified in randomized experimental settings

Ie. The "effect" of either factor does not depend on the level of the other factor.

Interacting (multiplicative) Factor Effects

There is a differential influence of one factor that depends on the levels of the other factor. Some ways to assess the appropriateness of an interaction model include:

- compare the mean difference for any two levels of the 1st factor to see if this is roughly the same for all levels of the 2nd factor (or vice versa)
- plot the treatment means for different factor levels to determine if the "curves" are roughly parallel.

Q) Derive a test statistic & rejection rule to test each of the following:

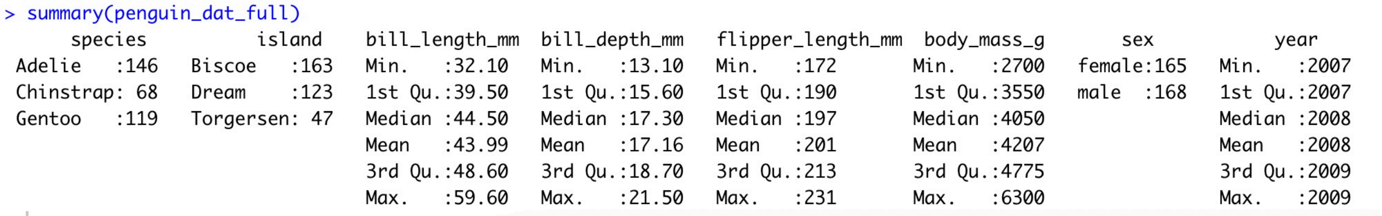
(a) Are the averages of the response significantly different according to the levels of the main factor?

(b) Are the avereages of the response significantly different according to the possible combinations of the main facter levels & the secondary facter levels?

The idea behind these tests is to consider a test statistic that is a ratio of one mean square term divided by the mean square error term.

If this ratio is much larger than 1 then this indicates the presense of a signal (ie. facter "effect") that is discernable from the noise (ie. the unexplained variability due to error).

11-18-22



```
> summary(penguin_dat_full)
      species          island    bill_length_mm  bill_depth_mm   flipper_length_mm  body_mass_g       sex           year     
 Adelie   :146   Biscoe   :163   Min.   :32.10   Min.   :13.10   Min.   :172       Min.   :2700   female:165   Min.   :2007  
 Chinstrap: 68   Dream    :123   1st Qu.:39.50   1st Qu.:15.60   1st Qu.:190       1st Qu.:3550   male  :168   1st Qu.:2007  
 Gentoo   :119   Torgersen: 47   Median :44.50   Median :17.30   Median :197       Median :4050                Median :2008  
                                 Mean   :43.99   Mean   :17.16   Mean   :201       Mean   :4207                Mean   :2008  
                                 3rd Qu.:48.60   3rd Qu.:18.70   3rd Qu.:213       3rd Qu.:4775                3rd Qu.:2009  
                                 Max.   :59.60   Max.   :21.50   Max.   :231       Max.   :6300                Max.   :2009  
```

Ex) Come up w/ research questions about these penguins that can be answered w/:

☐ a paired t-test
☐ a two sample (independent) t-test
☐ a one-way ANOVA F-test
☐ a two-way ANOVA (partial) F-test
+ interaction vs additive models?

R-code for comparing means

Q) Consider a hypothetical arguement:

"A treatment is a treatment, whether the study involves a single factor or multiple factors. The number of factors has little effect on the interpretation of the results (of an ANOVA model)."

Evaluate this arguement and form a response.'

1-Way $$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\left\{\begin{array}{l} i = 1, \ldots, I \\ j = 1, \ldots, J \\ k = 1, \ldots, K \end{array}\right.$$

2-Way

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}$$

Topic : Comparing Count Data (ch.13)

In situations where our data does not represent a measurement of a numeric variable, but rather represents counts of distinct qualitative features, the previous methods we've discussed are no longer relevant.

We will now shift our attention to a couple of error controlled statistical tests to aid in the analysis of categorical data.

Ex) In the penguin data, consider a setting in which the only data we have are

- the Island for each observational unit.
- the species
- the sex

Method 1 : Fisher's exact test

Is exact because the testing theory does not rely on any assumptions of $n \to \infty$.

The test statistic follows a hypergeometric distb'n under the assumption of H_0.

Method 2: Chi-Square test of homogeneity

method 3: Chi-Square test of independence

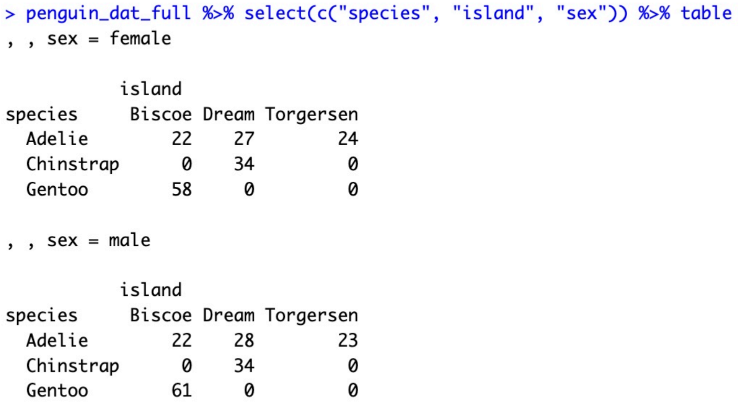
These methods require a large sample size (typically that each "cell" count is ≥ 5) b/c the testing theory relies upon the assumption that $n \to \infty$.

In each method, the test statistic asymptotically follows a Chi-Square distbn under the assumption of H_0.

Note: There are many modern exact methods under development that are made possible by the computational power available today.

11-21-22

Ex) Suppose we are studying the palmer penguins data but the only info we have for each penguin is their island of residence, their species, & their sex.



```
> penguin_dat_full %>% select(c("species", "island", "sex")) %>% table
, , sex = female

           island
species     Biscoe Dream Torgersen
  Adelie        22    27        24
  Chinstrap      0    34         0
  Gentoo        58     0         0

, , sex = male

           island
species     Biscoe Dream Torgersen
  Adelie        22    28        23
  Chinstrap      0    34         0
  Gentoo        61     0         0
```

Some Terminology:

Odds – prob. of success to failure for a given (fixed) row

Ex) odds a penguin is female $= \dfrac{\text{Pr(penguin is F)}}{\text{Pr(penguin is M)}}$

- odds = 1 ⇔ success & failure are equally likely
- odds < 1 ⇔ success less likely than failure

Odds ratio – a ratio of related odds

Ex) $\dfrac{\text{odss a Biscoe Island Penguin is F}}{\text{odds a Dream Island Penguin is F}}$

Fishers Exact Test

		Female	Male	
		pos	neg	
Biscoe	A	N_{11}	N_{12}	$n_{1\cdot}$
Dream	B	N_{21}	N_{21}	$n_{2\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

$$\begin{cases} H_0: \theta = 1 \\ H_A: \theta \neq 1 \end{cases}$$

$$\hat{\theta}_{MLE} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

$$Pr(N_{11} = n_{11}) \overset{H_0}{=} \frac{\binom{n_{1\cdot}}{n_{11}} \binom{n_{2\cdot}}{n_{21}}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}}$$

probabilities from a hypergeometric distb'n

Note: For 2×2 tables $H_0: \theta = 1$ is equivalent to testing the independence of the row & col variables.

Note: P-values from exact tests can be conservative (ie. measured larger than they really are).

Assumptions

- Row totals and column totals are fixed by design
- At least 3 cells have expected counts < 5 but no expected cell count is < 1.

Ex) Lady tasting tea experiment

8 cups of tea poured & presented in random order
4 had milk poured first, 4 had tea poured first

		Lady says poured first		
		tea	milk	total
Actually poured first	tea	3	1	4
	milk	1	3	4
	total	4	4	8

fixed b/c experimenter decides to pour 4 of each

fixed b/c lady knows half are tea first

H_0: The lady has no discerning ability (independence)

ie. $N_{11} \sim$ Hypergeometric

Note: If data are not a random sample & no random assignment is performed, then the question is not necessarily appropriate for (frequentist) probabilistic inference.

Chi-Square Test of Independence

Contingency Table w/ $\begin{cases} I \text{ rows} \\ J \text{ columns} \end{cases}$ & n = total sample size

$I \times J$ Contingency Table of Cell Counts:

Factor w/ J levels

Factor w/ I levels

	1	2	3	$\dots$	J	
1	n_{11}	n_{12}	n_{13}	$\dots$	n_{1J}	$n_{1\cdot}$
2	n_{21}	n_{22}			$\vdots$	$n_{2\cdot}$
3	n_{31}		$\ddots$			
$\vdots$	$\vdots$					
I	n_{I1}	n_{I2}	n_{I3}	$\dots$	n_{IJ}	
	$n_{\cdot 1}$	$n_{\cdot 2}$				n

statistical inference considers the joint distribution of the cell counts n_{ij}, for $i=1,\dots,I$ and $j=1,\dots,J$

$I \times J$ contingency table of probabilities:

Factor w/ J levels

Factor w/ I levels

	1	2	3	$\dots$	J	
1	π_{11}	π_{12}	π_{13}	$\dots$	π_{1J}	$\pi_{1\cdot}$
2	π_{21}	π_{22}			$\vdots$	$\pi_{2\cdot}$
3	π_{31}		$\ddots$			$\vdots$
$\vdots$	$\vdots$					
I	π_{I1}	π_{I2}	π_{I3}	$\dots$	π_{IJ}	$\pi_{I\cdot}$
	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	$\dots$		$\pi_{\cdot J}$	1

Marginal probabilities

i^{th} row: $\pi_{i\cdot} = \sum_{j=1}^{J} \pi_{ij}$

j^{th} column: $\pi_{\cdot j} = \sum_{i=1}^{I} \pi_{ij}$

H_0: The factor w/ J levels is independent of the factor w/ I levels

ie. $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ for every i,j

H_A: The factor w/ J levels is not independent of the factor w/ I levels

If the row levels are independent of the column levels

then $n_{ij} \sim \text{multinomial}(\pi_{ij})$

Assumption: The row and column totals can vary (are not(*) fixed by design) and the sample total n is large enough (expected cell counts all > 5)

$$\boxed{\hat{\pi}_{ij}^{MLE} \overset{H_0}{=}} \frac{n_{i\cdot}}{n} \times \frac{n_{\cdot j}}{n}$$

This is the MLE for π_{ij} under the assumption that the J cols & I rows are indep.

Pearson's Chi-Sq. Test Statistic:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \underset{n\to\infty}{\overset{H_0}{\sim}} \chi^2_{((I-1)(J-1))}$$

Observed: $O_{ij} = n_{ij}$

Expected: $E_{ij} = n \cdot \hat{\pi}_{ij}^{MLE} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$

Note: Independence of factor varbs can be understood as homogeneity of conditional distb'ns

*Edited on 11/23/22

Chi-Square Test for Homogeneity

Tests the homogeneity of a multinomial distb'n

J = the number of multinomial distributions we are testing for homogeneity (# of populations)

I = the number of levels of the factor that follows a multinomial distb'n

Assumptions: Data consists of independent samples from a multinomial distb'n (ie. either the column or row totals are fixed).
Sample is large enough (expected cell counts all > 5).

Note: This is a special case of the Chi-Squared goodness of fit test.

$$H_0: \pi_{i1} = \pi_{i2} = \cdots = \pi_{iJ}, \quad \text{for } i=1,\dots,I$$

$$\hat{\pi}_i^{\,MLE} \overset{H_0}{=} \frac{n_{i\cdot}}{n_{\cdot\cdot}}$$

[proof involves Lagrange multiplier]

So for the j^{th} multinomial, the expected count in the i^{th} category is

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

Pearson's Chi-Squared Test Statistic:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \underset{as\, n \to \infty}{\overset{H_0}{\sim}} \chi^2_{((I-1)(J-1))}$$

Next class:

Chi-Square Test for Goodness of Fit (Ch. 9.5)

R-code

"Office hours"

11-23-22

(*) Notes on Chi-Square sampling assumptions

The grand total, n, is fixed for any of these tests. What differs is whether or not row and/or column totals are also fixed, or allowed to vary.

Eg) You randomly select 100 individuals
> 54 turn out to be registered Democrats
> 46 turn out to be registered Republicans

and survey these individuals on whether or not reproductive rights are a top issue this election year.

This is a Chi-Square test of independence since the total individuals sampled is fixed but the number of each type of voter can vary.

Eg) In your county, 54% of registered voters are registered Democrats and 46% are registered Republicans. You randomly select 54 registered Democrats and 46 registered Republicans and survey these individuals on whether or not reproductive rights are a top issue this election year.

This is a Chi-Square test of homogeneity to determine if $Pr(\text{repro rights} \mid D) = Pr(\text{repro rights} \mid R)$.
The marginal totals of registered Dems & Reps is fixed by design in addition to the grad total of individuals being surveyed.

The asymptotic part relates to whether or not the entire grand total is increasing to infinity or the row (or column) totals are both increasing to infinity.

Chi-Square Test for Goodness of Fit (Ch. 9.5)

AKA : Likelihood ratio test (LHR) for the multinomial distb'n

A multinomial distb'n is an extension of the binomial but here there are more than 2 possible outcomes w/ associated probabilities. If

$$(N_1, \dots, N_k) \sim \text{Multinomial}(n_1, n_2, \dots, n_k ; p_1, p_2, \dots, p_k)$$

where $\sum_{i=1}^{k} p_i = 1$ are the probabilities associated w/ each of the k outcomes, and $\sum_{i=1}^{k} n_i = n$ are the corresponding counts.

then $Pr(N_i = n_i) = \binom{n}{n_i} p_i^{n_i} (1-p_i)^{n-n_i}$

and jointly

$$Pr(N_1 = n_1, N_2 = n_2, \dots, N_k = n_k) = \binom{n}{n_1 \cdots n_k} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

Note, a Chi-Sq test of homogeneity is testing whether or not $p_1 = p_2 = \dots = p_k$.

The Chi-Square goodness of fit test determines if the data support a particular value for each p_i, $i=1,\dots,k$ probability.

ie. $H_0: \underset{\sim}{p} = \underset{\sim}{p}(\theta)$ where $\theta \in \omega_0 \subseteq \mathbb{R}^k$

The entire joint parameter space is
$\Omega = \{x_i : x_i \geq 0 \text{ and } \sum_{i=1}^{k} x_i = 1\}. \subseteq \mathbb{R}^{k-1}$

By the invariance property of MLEs, if
$\hat{\theta}_{MLE}$ is the MLE for θ restricted to ω_0,
then $\underset{\sim}{p}(\hat{\theta}_{MLE}) = (p_1(\hat{\theta}_{MLE}), p_2(\hat{\theta}_{MLE}), \dots, p_k(\hat{\theta}_{MLE}))$
is the MLE for $\underset{\sim}{p}$ under H_0.

un-restricted, the MLE for $\underset{\sim}{p}$ over Ω is
$\hat{p}_{MLE} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \dots, \frac{n_k}{n}\right)$.

The GLHR test statistic, Λ, is asymptotically equivalent to the Pearson's Chi-Sq. test stat, χ^2:

$$\Lambda = \frac{\frac{n!}{n_1! n_2! \cdots n_k!} p_1(\hat{\theta})^{n_1} \cdots p_k(\hat{\theta})^{n_k}}{\frac{n!}{n_1! n_2! \cdots n_k!} \hat{p}_1^{n_1} \cdots \hat{p}_k^{n_k}} \overset{n \to \infty}{\approx} \sum_{i=1}^{k} \frac{\overset{\text{observed cell counts}}{n_i} - n p_i(\hat{\theta})}{\underset{\text{expected cell counts}}{n p_i(\hat{\theta})}} = \chi^2$$

as can be seen on pg. 342 of your textbook (using a Taylor series expansion).

R-code cheat sheet

11-28-22

Announcements + updates

- HW 21
- Power + t-tests for comparing means
- Categorical data quick review/highlights

Notes on comparing means: (from blackboard)

Common among all methods below is the assumption of constant, common variance.

CI's for the true difference in group means takes form:

$$\bar{D} \pm \underbrace{[t^* \cdot SE(\bar{D})]}_{\text{margin of error small} \iff \text{larger power}}$$

- Unpaired t-test

  $t^*_{(m+n-2;\, \alpha/2)}$, $SE(\bar{D}) = \hat{\sigma}\sqrt{\frac{1}{n}+\frac{1}{m}}$, $\hat{\sigma} = S_{pool}$

  > balanced $n=m$, total sample size is $2n$

  > unbalanced (less powerful than balanced), total samp size is $n+m$

- Paired t-test (implicitly balanced $m=n$)

  total sample size is $2n$

  > is more powerful than balanced un-paired t-test only if the paired data are strongly linearly correlated.

  $t^*_{(n-1,\, \alpha/2)}$, $SE(\bar{D}) = \hat{\sigma}\sqrt{1/n}$, $\hat{\sigma} = \frac{1}{n-1}\sum_i^n (D_i - \bar{D})^2$ $(= s_D)$

11-30-22

Group Work - Simulation Studies to understand "chance"

Specifically, let's investigate what "random chance" can look like in the context of a Chi-Square procedure to test for homogeneity.

1) What is your true model that you will use to generate many observed data from?

2) How many times will you generate new data sets and how will you summarize these data?

3) Do you see (in any of your simulated data sets) patterns that look like they came from a non-homogenous population model? How often does this happen?

Topic: Linear Least Squares
(ch. 14)

Setting:

Data → quantitative, continuous response Y_i — observed w/ some measurement error

Data → group of p-1 categorical and/or quantitative predictors $(x_{1i}, x_{2i}, \ldots, x_{p-1,i})$ — observed w/out any measurement error

The idea is to use a (p-1)-dimensional hyperplane to model a predictive relationship btwn the p-1 predictors and the response.

This is a very powerful modeling technique that provides the foundations for more advanced extensions including

- multivariate response $Y_{1i}, Y_{2i}, \ldots, Y_{mi}$
- binary response (logistic reg) $Y_i = \begin{cases} 0, \text{ with prob } 1-p \\ 1, \text{ with prob } p \end{cases}$
- dependent response (eg. time series) $Y_i = W_{t+m} - W_t$
  - t = point in time
  - m = lag
- random predictors (allowing for measurement errors in the predictors) (eg. random effects or mixed effects)
- and more!

MLR Model : Multiple Linear Regression

In this model, there are p-1 > 1 predictor varbs which can be categorical or numeric.

$$(*)\quad \underset{n\times 1}{\underset{\sim}{Y}} = \underset{n\times p}{X}\,\underset{p\times 1}{\underset{\sim}{\beta}} + \underset{n\times 1}{\underset{\sim}{\varepsilon}}$$

where $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & & x_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & & x_{np-1} \end{bmatrix}$ is called the design matrix

$\underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ is the vector of unknown regression coefficients (including the intercept)

and $\underset{\sim}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ is the random vector of (mutually) independent and identically distb'ted noise (or measurement error)

The fitted (or estimated) model is thus

$$\underset{n\times 1}{\hat{\underset{\sim}{Y}}} = \underset{n\times p}{X}\,\underset{p\times 1}{\hat{\underset{\sim}{\beta}}}$$

where $\hat{\underset{\sim}{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix}$ is a $p\times 1$ vector of the least squares estimates for each β_j parameter in the model (*) above.

SLR: Simple Linear Regression

In this model, there is only one quantitative predictor, x.

$$\underset{\sim}{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

To solve for $\underset{\sim}{\hat{\beta}}$ we want to minimize

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

w/ respect to $\underset{\sim}{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$.

Q) How can we express $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ in terms of matricies or vectors?

$$\underset{\sim}{Y} - \underset{\sim}{\hat{Y}} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$$

So $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \| \underset{\sim}{Y} - \underset{\sim}{\hat{Y}} \|^2 = \| \underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}} \|^2$

Recall the definition of the Euclidean norm:

for vector $\underset{\sim}{u} = [u_1 \ u_2 \ \cdots \ u_m]^T$,

$$\|u\| = \left[\sum_{i=1}^{m} u_i^2\right]^{1/2}$$

MLR Model : Least Squares Estimates

To solve for $\hat{\beta}$, we want to minimize

$$\|\underset{\sim}{Y} - \underset{\sim}{\hat{Y}}\|^2 = \|\underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}}\|^2$$

w/ respect to $\hat{\beta}$.

Note $\underset{\sim}{Y} - X\hat{\beta} = \begin{bmatrix} y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_{11} + \hat{\beta}_2 x_{12} + \cdots + \hat{\beta}_{p-1} x_{1,p-1} \\ \vdots \\ y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_{n1} + \hat{\beta}_2 x_{n2} + \cdots + \hat{\beta}_{p-1} x_{n,p-1}) \end{bmatrix}$

so the minimizing estimator, $\hat{\beta}$, solves

$$\sum_{i=1}^{n} y_i - \left[n\hat{\beta}_0 + \left(\sum_{i=1}^{n} x_{i1}\right)\hat{\beta}_1 + \ldots + \left(\sum_{i=1}^{n} x_{ip-1}\right)\hat{\beta}_{p-1}\right] = 0$$

and

$$\sum_{i=1}^{n} y_i x_{ik} - \left[\left(\sum_{i=1}^{n} x_{ik}\right)\hat{\beta}_0 + \left(\sum_{i=1}^{n} x_{i1} x_{ik}\right)\hat{\beta}_1 + \ldots + \left(\sum_{i=1}^{n} x_{ik} x_{ip-1}\right)\hat{\beta}_{p-1}\right] = 0$$

for all $k = 1, \ldots, p-1$.

In matrix notation this means that $\hat{\beta}$ solves

$$(X^T Y - X^T X\hat{\beta}) = 0$$

ie. $X^T X \hat{\beta} = X^T Y$.

These are called the "normal equations". and they imply that $\hat{\beta}_{LSE} = (X^T X)^{-1} X^T \underset{\sim}{Y}$ (provided $X^T X$ is invertible).

12-2-22

The normal equations solve the problem of finding a $\hat{\underset{\sim}{\beta}}$ that minimizes $\|\underset{\sim}{Y} - \hat{\underset{\sim}{Y}}\|^2$, ie.

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

Computational Concerns

When both n and p are large, the design matrix, X, becomes unweildy. making inverting $X^T X$ very costly (in computing time).

A few common numerical techniques can help make this inversion possible. These include

- QR Method
  factors $X = QR$ so that $Q^T Q = I_{p\times p}$ and R is upper triangular
- Cholesky Decomposition
  factors $X^T X = R^T R$, so that R is upper triangular

Other Issues

If p is large, the design matrix is large and, depending on the sample size, n, X may be quite sparse if we are using categorical predictors). Another potential issue occurs when we are using many numeric/quantitative predictors that are closely related. In particular, if one numeric predictor, say $\underset{\sim}{X}_1$, is approximately (or exactly) linearly associated to another, say $\underset{\sim}{X}_2$ then the association $\underset{\sim}{X}_1 \approx a + b\underset{\sim}{X}_2$ reduces the rank of the design matrix and makes $X^T X$ singular, ie. non-invertible.

MLR Model

$$\underset{n\times 1}{\underset{\sim}{Y}} = \underset{n\times p}{X}\,\underset{p\times 1}{\underset{\sim}{\beta}} + \underset{n\times 1}{\underset{\sim}{\varepsilon}}$$ where

$$E(\underset{\sim}{\varepsilon}) = \begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ \vdots \\ E(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \underset{\sim}{0}$$ (zero mean) and

$$\Sigma \overset{def}{=} Var(\underset{\sim}{\varepsilon}) = \begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1,\varepsilon_2) & \cdots & Cov(\varepsilon_1,\varepsilon_n) \\ Cov(\varepsilon_2,\varepsilon_1) & Var(\varepsilon_2) & & \vdots \\ Cov(\varepsilon_3,\varepsilon_1) & \cdots & \ddots & Cov(\varepsilon_{n-1},\varepsilon_n) \\ \vdots & & & \\ Cov(\varepsilon_n,\varepsilon_1) & \cdots & \cdots & Var(\varepsilon_n) \end{bmatrix}_{n\times n}$$

"covariance matrix of $\underset{\sim}{\varepsilon}$"

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \sigma^2 \end{bmatrix} = \sigma^2 I_{n\times n}$$

(constant variance & uncorrelated errors)

Q) What is $Var(\underset{\sim}{Y})$? $Var(\underset{\sim}{Y}) = \sigma^2 I_{n\times n}$

From these model assumptions we have that

$$\begin{aligned} E(\hat{\underset{\sim}{\beta}}) &= E[(X^TX)^{-1}X^T\underset{\sim}{Y}] \\ &= E[(X^TX)^{-1}X^T(X\underset{\sim}{\beta}+\underset{\sim}{\varepsilon})] \\ &= E[(X^TX)^{-1}X^TX\underset{\sim}{\beta} + (X^TX)^{-1}X^T\underset{\sim}{\varepsilon}] \\ &= E[\underset{\sim}{\beta} + (X^TX)^{-1}X^T\underset{\sim}{\varepsilon}] \\ &= \underset{\sim}{\beta} + (X^TX)^{-1}X^TE(\underset{\sim}{\varepsilon}) = \underset{\sim}{\beta} \end{aligned}$$

ie. $\hat{\underset{\sim}{\beta}}$ is unbiased for $\underset{\sim}{\beta}$

and we also have a way to derive the (co)variance (matrix) for the sample estimate $\hat{\underset{\sim}{\beta}}$:

$$\mathrm{Var}(\hat{\underset{\sim}{\beta}}) = \begin{bmatrix} \mathrm{Var}(\hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \cdots & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_{p-1}) \\ \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \mathrm{Var}(\hat{\beta}_1) & & \vdots \\ \vdots & & \ddots & \mathrm{Cov}(\hat{\beta}_{p-2}, \hat{\beta}_{p-1}) \\ \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_{p-1}) & \cdots & \mathrm{Cov}(\hat{\beta}_{p-2}, \hat{\beta}_{p-1}) & \mathrm{Var}(\hat{\beta}_{p-1}) \end{bmatrix}$$

Since $\mathrm{Var}(\underset{\sim}{\varepsilon}) = \sigma^2 I_{n\times n}$ and

$$\begin{aligned} \hat{\underset{\sim}{\beta}} &= (X^TX)^{-1}X^T\underset{\sim}{Y} \\ &= (X^TX)^{-1}X^T(X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}) \\ &= \underbrace{(X^TX)^{-1}X^TX\underset{\sim}{\beta}}_{\text{const}} + \underbrace{(X^TX)^{-1}X^T\underset{\sim}{\varepsilon}}_{\text{random}} \end{aligned}$$

, we have that

$$\begin{aligned} \mathrm{Var}(\hat{\underset{\sim}{\beta}}) &= \mathrm{Var}\left((X^TX)^{-1}X^T\underset{\sim}{\varepsilon}\right) \\ &= (X^TX)^{-1}X^T\,\mathrm{Var}(\underset{\sim}{\varepsilon})\,X(X^TX)^{-1} \\ &= (X^TX)^{-1}X^T(\sigma^2 I_{n\times n})\,X\,(X^TX)^{-1} \\ &= \sigma^2\left[(X^TX)^{-1}X^TX(X^TX)^{-1}\right] \\ &= \sigma^2(X^TX)^{-1} \end{aligned}$$

here we use that $(X^TX)^{-1}$ is symmetric

Now that we have $E(\hat{\underset{\sim}{\beta}})$ and $\mathrm{Var}(\hat{\underset{\sim}{\beta}})$, we can describe the sampling variability of our estimators.

If we assume each ε_i follow a particular distribution (eg. $N(0,\sigma^2)$), we can also describe the sampling distribution of $\hat{\underset{\sim}{\beta}}$ and construct tests or CIs for $\underset{\sim}{\beta}$.

Estimating σ^2

The residuals of a MLR model are

$$\underset{\sim}{e} = \underset{\sim}{y} - \underset{\sim}{\hat{y}} = \underset{\sim}{y} - \underbrace{X(X^TX)^{-1}X^T}\underset{\sim}{y}$$

It can be shown that for $P \overset{def}{=} X(X^TX)^{-1}X^T$

- $P = P^T = P^2$
- $I - P = (I-P)^T = (I-P)^2$

Hence, the matrix product represented by P is called the projection matrix (or the "hat" matrix) because it projects the observed $\underset{\sim}{y}$ onto the subspace spanned by the design matrix (thereby producing the fitted $\hat{\underset{\sim}{y}}$ values).

So now we have that the sum of the squared residuals (RSS) is:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \|\underset{\sim}{y} - \underset{\sim}{\hat{y}}\|^2 = \|\underset{\sim}{y} - P\underset{\sim}{y}\|^2$$

$$= \ldots$$

$$= \underset{\sim}{y}^T (I-P)\underset{\sim}{y}$$

[Recall: $MSE = \frac{RSS}{n-p}$]

with mean $E[\underset{\sim}{y}^T(I-P)\underset{\sim}{y}] = \ldots = (n-p)\,\sigma^2$.

Therefore, an unbiased estimate for the error ($\underset{\sim}{\varepsilon}$) variance is:

$$\hat{\sigma}^2 = \frac{\|\underset{\sim}{y} - \underset{\sim}{\hat{y}}\|^2}{n-p}$$

HW Review reg to mean +

(root) MSE

pg 561

error here means modeling error (i.e. resids)

Key: Covariance + Variance operators

Uncorrelated relationships such as

$\bar{Y}$ uncorrelated w/ each $Y_i - \bar{Y}$,

$\bar{Y}$ uncorrelated w/ $\hat{\beta}_1$,

sample variance uncorrelated w/ sample mean.

Check Moodle for complete solns.

12-5-22

Note on HW 23 #1c

The key to this problem is knowing the definition of the coefficent of determination.

$$r^2 = 1 - \frac{SSE}{S_{YY}} \Rightarrow 1 - r^2 = \frac{SSE}{S_{YY}} \quad \text{where } S_{YY} = \Sigma(y_i - \bar{y})^2$$
$$= (n-1)\, S_y^2 .$$

Now MSE (mean square error) is

$$MSE = \frac{SSE}{n-2} = \frac{(1-r^2)\, S_{YY}}{n-2}$$
$$= \frac{(1-r^2)(n-1) S_y^2}{n-2}$$

and RMSE (root MSE) is

$$RMSE = \sqrt{(1-r^2)}\, S_y \left(\frac{n-1}{n-2}\right) .$$

Since you are given r, S_y, and n you can solve for RMSE.

Page	Case Title
	PLANNING/DESIGN
232	**Case 1.** You recognize during the planning stage that there is/you have/the team has an incomplete understanding of the problem to be addressed
238	**Case 2.** You are asked to create one computational step in a multi-step process, and *no one will tell you* what will happen with your results
244	**Case 3.** You seek to incorporate sensitivity checks along the planning/development process but meet with resistance
250	**Case 4.** You recognize a better way to achieve a computational result than the proprietary way you were told to follow. Your way takes longer, so there is resistance to trying your method; but you can show it uses less data and results are less biased

Page	Case Title
256	**Case 5.** You are asked to use a specific analysis or system design that is methodologically inappropriate given the research question or objective
263	**Case 6.** You are asked to design a study or system that will collect either implausible/unreasonably low amounts of data (small sample size) or unnecessarily high amounts of data
	COLLECT/MUNGE/WRANGLE DATA
275	**Case 7.** A plan is created to collect data that cannot possibly be housed securely
284	**Case 8.** Data collection is carried out by scraping the Internet; you notice that at least some of the time, the results of confidentiality and privacy breeches get swept up in the scraping
292	**Case 9.** Your supervisor directs you to assume that if *any* of the data in your collection was obtained with any level of consent (whether none *or* opt-out), then treat *all* of the data as if it was obtained "with consent"
299	**Case 10.** Standard Operating Procedures (SOP) manuals direct you to ignore data provenance

Page	Case Title
307	**Case 11.** You discover that there has been no consent obtained for any of the data you are asked to collect/wrangle/munge
314	**Case 12.** You have collected/wrangled data from multiple sources and provenance information about the data is inconsistent – different people at work describe it differently and there's no real evidence about the provenance of *any* of the data
	ANALYSIS
327	**Case 13.** You are told to implement an analysis plan that you suspect was written by someone else (who does not know it is being used) and for another problem/project
336	**Case 14.** Your supervisor ignores your requests for reviews of your work and tells you that no one else can review it either
343	**Case 15.** You are asked to carry out an analysis you are confident that you do *not know how to do or interpret* (or troubleshoot)

Page	Case Title
	INTERPRETATION
401	**Case 22.** You discover that prior (expected) results cannot be reproduced. Sensitivity analyses strongly suggest that earlier results were spurious; reading the team's report of that analysis confirms this: the results were improperly interpreted to favour the team's objective
410	**Case 23.** At the end of a long project, you realize you made an error early on. The results cannot be interpreted in a valid way. Everything has to be redone
418	**Case 24.** At the end of a long project, you realize your supervisor made an error early on. The results cannot be interpreted in a valid way. Everything has to be redone
426	**Case 25.** You complete a very large set of analyses; one result happens to be "significant". A senior team member highlights this result, interpreting it without considering the context
436	**Case 26.** Your supervisor singles out one "meaningful" result to demonstrate that whatever you've been doing "is working", even after you carry out multiple simulations that show their single, "favourite," result is totally spurious

Page	Case Title
DOCUMENTING YOUR WORK	
452	**Case 27.** It takes as long to fully and transparently document your work as it does to do the work itself. Since this is just *your* job, not documenting it will only affect you (for the foreseeable future) –and is faster
458	**Case 28.** You failed to fully document your work a few months ago and now your supervisor is requesting your comprehensive documentation so that another person can replicate your work. You really only have time for minimal documentation
464	**Case 29.** You receive documentation of an ongoing program/analysis that lacks all information about data provenance
472	**Case 30.** Prior documentation of an organization-wide method is complete and correct. The method development did not include sensitivity analyses. You do a few and identify two important errors in the method

Page	Case Title
479	**Case 31.** You are given documentation that is not complete: it lacks details about exactly what methods and in what order were used
486	**Case 32.** You provide complete and correct documentation, and this gets "edited" by a supervisor so that it is now no longer complete or correct
499	**Case 33.** The documentation you receive specifies an analysis method that is not appropriate for the specific question that must be addressed
	REPORTING
512	**Case 34.** You discover that incorrect results (yours and/or your team's) are going to be featured in a high-profile publication
522	**Case 35.** You follow SOP and the GLs/CE, and report your methods and results fully, but the final report has incorrect methods and results that were "edited" to suit a senior member of the team without your knowledge or agreement

Page	Case Title
533	**Case 36.** Stakeholders (donors, funders, employers) are given a misleading summary of your methods and results
545	**Case 37.** Your sensitivity analyses that pinpoint the next logical step in your work are omitted from a final report to funders because "we could get a grant to support the team for another 5 years to figure that out!"
553	**Case 38.** If you report your method fully and transparently, then you will lose the opportunity to patent it
560	**Case 39.** If you report your method fully and transparently, then a reviewer might notice that you are not the original developer of this method – although the same method was published over 30 years ago and in *another* field
568	**Case 40.** You prepare a report identifying difficulties you encountered in your evaluation of a system your organization wants to deploy or an analysis that was done. The organization does not have a mechanism for submitting or sharing this report (or peer review of any type) either internally or with stakeholders

Page	Case Title
	TEAM WORK/TEAM SCIENCE
577	**Case 41.** You notice a pattern of bullying by a senior team member
587	**Case 42.** You are asked to do some coding/analysis by someone who is prevented from acknowledging that you helped. Your contribution cannot be recognized
596	**Case 43.** Your supervisor tells you that you "only need to read/review your own work" and you are not allowed to see the final/full document or work product
604	**Case 44.** You complete the analysis plan/system design, oversee its operation, and draft the report. You suddenly receive a "new draft" of the report that excludes all of the work you did, nor does any of the documentation of the system or work from your original report appear. You can tell without carefully reading it that the "new draft" has obvious errors in the design/analysis, results, and other reported elements, but you are asked to "approve" the new draft – and agree to be/remain a co-author on the report – within the next two days. You also have another project deadline in two days

Page	Case Title
613	**Case 45.** Someone on your team suggests a technical method to overcome a lack of consent from data contributors and collect their data even if they do not consent
622	**Case 46.** You recognize the potential for "dual use" of your team's code, data, and/or results
630	**Case 47.** You inadvertently discover that a proprietary "new method" that you were told to prepare for publication/patent application was actually published decades ago and was apparently unnoticed until you found it