PROBABILITY AND STATISTICS FOR ENGINEERS MIDTERM 3		
Code : *CVE 303* Acad.Year: *2019-20* Semester : *Fall* Date : *10.01.2020* Time : *9:00* Duration : *110 min*	Last Name: Name: Solution Student ID: Signature:	#:
	6 QUESTIONS ON 4 PAGES TOTAL 94 POINTS	

Q1 (12)	Q2 (20)	Q3 (16)	Q4 (13)	Q5 (14)	Q6 (19)	Total. (94)
12	20	16	13	14	19	

1. *(6×2=12pts)* Short answer questions about two sample hypothesis testing.

(A) When do you use two sample hypothesis testing?

To check if samples from two different independent populations have different means.

(B) When do you use pooled variance two sample hypothesis testing?

When you expect the two populations to have (approximately) the same variance.

(C) Why would you use pooled variance?

Pooling the variance calculation lowers the standard error of estimates. ⇒ Lower p-value.

(D) How can you check if pooled variance is appropriate?

Use an F-test on the two sample variances to see if they are close enough.

(E) What is the distribution of $\overline{X} - \overline{Y}$ if $X \sim \text{Normal}(\mu_X, \sigma_X)$ sampled n times, $Y \sim \text{Normal}(\mu_Y, \sigma_Y)$ sampled m times?

$$(\overline{X} - \overline{Y}) \sim \text{Normal}\left(\mu_X - \mu_Y, \sqrt{\sigma_X^2/n + \sigma_Y^2/m}\right)$$

(F) Are X and Y independent in two-sample t-testing?

Yes. [Otherwise you would use a paired-data t-test.]

2. *(4×2=8pts)* Short answer questions about ANOVA.

(A) What does "ANOVA" stand for?

ANalysis Of VAriance

(B) What is ANOVA used for?

To test whether there is a difference (in means) among many populations (factor values)

(C) Why use ANOVA instead of many two sample t-tests?

α-Inflation causes a loss of significance when performing many t-tests.

(Also it is a lot of computation...)

(D) When is ANOVA equivalent to two sample t-testing?

ANOVA gives the same result as two-sample t if there are only two factor values.

2. *(6×2=12pts)* More short answer questions about ANOVA.

(E) What are "Residuals" in ANOVA?

"Residuals" = "Errors"

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

(Observed value) − (Factor mean)

(F) What does "SSF" and "SSE" stand for?

SSF = "Sum of Squares, Factor"

SSE = "Sum of Squares, Error"

(G) What is the relationship between SSF, SSE, SST?

$SSF + SSE = SST$

"Factor + Error = Total"

(H) What is the relationship between MSF, MSE, MST?

Trick Question!!!

$MSF + MSE \neq MST$

No simple relationship....

$(k-1)MSF + (N-k)MSE = (N-1)MST$

(I) In ANOVA is it good or bad to have large SSF?

Large SSF is good.

(Want SSE to be small
– most of SST should be SSF)

(J) What does "Tukey's HSD Test" do?

It tests for factor values whose means are "significally different"

3. *(8×2=16pts)* Short answer questions about regression analysis.

(A) When is Regression used?

When you expect a linear relationship between X & Y

(X & Y not independent.)

(B) Are X and Y independent in Regression Analysis?

No.

(C) What is the Linear Regression Model?

$Y \sim \text{Normal}(\beta_1 x + \beta_0, \sigma_\varepsilon)$

– or –

$Y = \beta_1 X + \beta_0 + \varepsilon$ ← Normal$(0, \sigma_\varepsilon)$

(D) Give the formula for the regression line in terms of sample means $\bar{x}$ and $\bar{y}$.

$\hat{y}(x) = \beta_1 (x - \bar{x}) + \bar{y}$

(E) Give a formula for $\hat{\beta}_1$ in terms of Variance and Covariance.

"Y" / "X"

$\hat{\beta}_1 = \frac{Cov[X,Y]}{Var[X]} = \frac{Cov[X,Y]}{Cov[X,X]}$

(F) What are "observed" and "fitted" values in Regression analysis?

Observed Values: y_i

Fitted Values: $\hat{y}_i = \hat{y}(x_i)$ ← From regression line

(G) What does "SSR" correspond to from ANOVA?

Regression		ANOVA
SSR	⟷	SSF
"Regression"		"Factor"

(H) What does "R^2" correspond to from ANOVA?

Regression		ANOVA
$R^2 = \frac{SSR}{SST}$	⟷	$\eta^2 = \frac{SSF}{SST}$
		"Effect Size"

4. *(5+4×2=13pts)* Complete the ANOVA table below and answer questions about its values.

Source	df	SS	MS	F	p
Factor	10	70	7	7/2	0.00219
Error	40	80	$\frac{80}{40} = 2$		
Total	50	150	$\frac{150}{50} = 3$		

(A) How many total samples were there?

$50 + 1 = \boxed{51}$

(B) How many factors were there?

$10 + 1 = \boxed{11}$

(C) What hypothesis has p-value 0.00219?

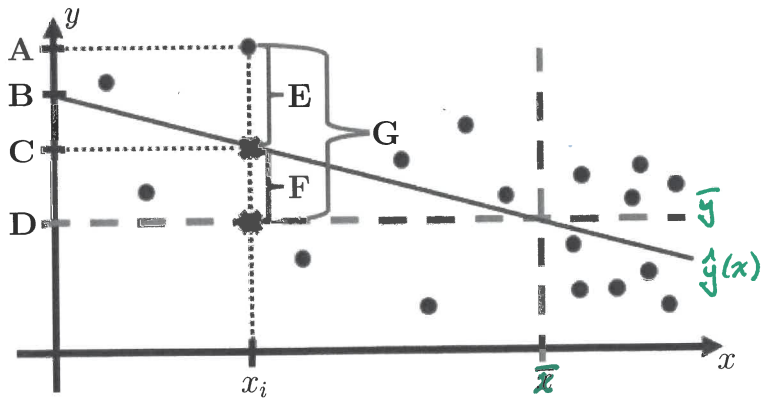
H_0: All factor values have same mean

(D) What is the effect size η^2?

$\eta^2 = \frac{SSF}{SST} = \frac{70}{150} = \boxed{\frac{7}{15}}$

5. *(7×2=14pts)* Give labels for the following parts of the regression plot.

Label the y-coordinates:



A y_i (Observed)

B $\hat{\beta}_0$ (Intercept)

C $\hat{y}_i = \hat{y}(x_i)$ (Fitted)

D $\bar{y}$ (Sample Mean)

Label the sum of squares (e.g. "SSR") measured by the distances:

E = $(y_i - \hat{y}_i)$ (Observed − Fitted)

F = $(\hat{y}_i - \bar{y})$ (Fitted − GrandMean)

G = $(y_i - \bar{y})$ (Observed − Grand Mean)

SSE ("Error") + SSR ("Regression") = SST ("Total")

6. *(11+4×2=19pts)* Complete the regression tables below and answer questions about their values.

Predictor	Coeff	Std.Error	t	p
Const	1	0.310	$\frac{1}{0.310}$	0.00111
x	$\frac{1}{2}$	0.224	$\frac{1/2}{0.224}$	0.0298

$\beta_0 \rightarrow$ Const; $\beta_1 \rightarrow x$; $H_0: \beta_0 = 0$; $H_0: \beta_1 = 0$ — Equal!

Source	df	SS	MS	F	p
Regression	1	10	$\frac{10}{1} = 10$	$\frac{10}{2} = 5$	0.0298
Residual	50	100	$\frac{100}{50} = 2$		
Total	51	110	2.157		

"Error" (ε) → Residual; $H_0: \beta_1 = 0$

Goodness of Fit	
R^2	$\frac{10}{110} = \frac{1}{11}$
S	$\sqrt{2}$

$\sqrt{MSE} = \sigma_\varepsilon$

(A) What is the formula for the regression line?

$$\hat{y}(x) = \tfrac{1}{2}x + 1$$

(B) What is the 95% confidence interval for β_0?
(use `qt(..)` notation)

$$\beta_0 \approx \underbrace{1}_{\hat{\beta}_0} \pm \underbrace{0.310}_{\sigma_{\beta_0}} \cdot qt(\overbrace{0.025}^{\alpha/2}, \overbrace{50}^{df \text{ of Error } \varepsilon})$$

$\sigma_{\beta_0} = \sqrt{MSE/n} = \sqrt{2/52}$ (oops...)

(C) What is the 95% confidence interval for β_1?
(use `qt(..)` notation)

$$\beta_1 \approx \underbrace{1/2}_{\hat{\beta}_1} \pm \underbrace{0.224}_{\sigma_{\beta_1}} \cdot qt(\overbrace{0.025}^{\alpha/2}, \overbrace{50}^{df \text{ of Error } \varepsilon})$$

(D) What is the 95% confidence interval for y?
(use `qt(..)` notation)

$$y(x) \approx \underbrace{(\tfrac{1}{2}x + 1)}_{\hat{y}(x)} \pm \sqrt{\underbrace{0.224^2(x-\bar{x})^2}_{\sigma^2_{\beta_1}} + \underbrace{0.310^2}_{\sigma^2_{\beta_0}}} \cdot qt(0.025, 50)$$