PROBABILITY AND STATISTICS FOR ENGINEERS MIDTERM 3		
Code : *CVE 303* Acad.Year : *2018-19* Semester : *Spring* Date : *29.05.2019* Time : *16:00* Duration : *120 min*	Last Name : Name : Solution Student ID : Signature :	#:
	8 QUESTIONS ON 5 PAGES TOTAL 100 POINTS	

P1. (20)	P2. (20)	P3. (20)	P4. (20)	P5. (20)		Total. (100)

1. *(4×2=8pts)* Short answer questions about general hypothesis testing.

(A) What p-values are possible in hypothesis testing? Why?

$0 \leq p \leq 1$

p-values are probabilities

(B) Is it better to have big or small p-value? Why?

small p-value is better

to reject H_0

(C) What are the two possible results of a hypothesis test?

- Reject H_0
- Fail to reject H_0

(D) In t-tests how does degrees of freedom and sample variance affect p-value?

high deg. of freedom $\Rightarrow$ low p-value

low sample variance $\Rightarrow$ low p-value

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

(A) What is a two sample t-test used for?

Test difference between means

$\mu_X - \mu_Y$

(B) What is a two sample F-test used for?

Test equality of variance

$\sigma_X^2 = \sigma_Y^2$

(C) What is "pooled sample variance" in two-sample testing?

Combined (weighted average) variance of X & Y

(D) Why would you use a pooled variance two-sample tests instead of a regular two-sample test?

Pooled variance tests have higher degrees of freedom

(E) When is paired sample testing used?

When X & Y are not independent.

→ Testing mean of $X_i - Y_i$

(F) Why use paired sample testing instead of two-sample?

It yields lower variance.

3. *(10× 2=20pts)* Short answer questions about ANOVA.

(A) What do the letters in ANOVA stand for?

ANalysis Of VAriance

(B) When do you use ANOVA?

Want to test equality of means for random variable split into factors $X^{(f)}$

(C) What is the difference between ANOVA and two sample t-tests?

Two sample t-tests have two populations. X & Y

ANOVA has many populations $X^{(f)}$

(D) What are the "residuals" in ANOVA?

Residuals measure variance within factors

$\varepsilon_i = x_i^{(f)} - \bar{x}^{(f)}$ ← Difference from factor mean.

(E) What does "SSF", "SSE", and "SST" stand for?

- SSF: Sum of Squares (Factor)
- SSE: Sum of Squares (Error)
- SST: Sum of Squares (Total)

(F) Identify which are "MSF", "MSE", "MST"

- Pooled sample variance (Variance within factors): MSE
- Over-all variance: MST
- Variance between factors: MSF

(G) What is the test statistic F used in ANOVA?

$F = \frac{\text{Variance Between Factors}}{\text{Variance Within Factors}} = \frac{MSF}{MSE}$

(k-1 df) (N-k df)

(H) What is the distribution of the test statistic F in ANOVA?

$F \sim F(\#\text{factors} - 1, \#\text{samples} - \#\text{factors})$

$= F(k-1, N-k)$

(I) What is the "effect size" η^2 in ANOVA?

$\eta^2 = \frac{SSF}{SST}$ ← "Proportion of total variance caused by the factors."

(J) What is Tukey's HSD Test used for?

It tests which factor means are significantly different.

4. *(10×2=20pts)* Short answer questions about regression analysis.

(A) When do you use regression analysis?

When you expect Y to (approximately) linearly depend on X. (Y & X not independent)

(B) What is the "regression line"?

$\hat{y}(x)$ is the estimated line of means $\mu_Y^{(x)}$ of $Y(x)$

→ The best fit line through data.

(C) What is the relationship between Y and X assumed by regression analysis?

$$Y = (\beta_0 + \beta_1 X) + \varepsilon$$

(D) Write a formula for the regression line in terms of $\bar{x}$ and $\bar{y}$ and $\hat{\beta}_1$.

$$\hat{y}(x) = \bar{y} + \hat{\beta}_1 (x - \bar{x})$$

(E) What point does the regression line always go through?

The means $(\bar{x}, \bar{y})$

(F) What does R^2 measure?

R^2 measures how close sample points are to regression line

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

Short answer questions about confidence intervals of regression lines.

(G) How many degrees of freedom do $\hat{\beta}_0$ and $\hat{\beta}_1$ have.

Both have same #deg. of free. as MSE ↝ #samples − 2

$N - 2$

(H) Give the formula for the standard error of $\hat{\beta}_0$ in terms of MSF, MSE, MST, etc.

$$\sigma_{\hat{\beta}_0} = \sqrt{MSE/N}$$

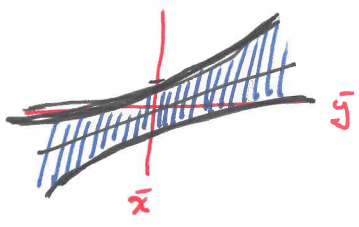
N = total # samples

(I) Which values affect the size of the confidence interval for a regression line at x?

- Distance of x from $\bar{x}$
- Value of MSE
- Value of N (#samples)
- Value of S_{xx} ↝ or Var[X]

(J) Describe the shape of the confidence interval for a regression line.

Narrowest at $\bar{x}$, getting wider as x moves away from $\bar{x}$



5. *(5+5=10pts)* Independent normal populations X and Y have $\sigma_X = 7$ and $\sigma_Y = 9$.

- X is sampled 50 times yielding sample mean $\bar{x} = 20$
- Y is sampled 40 times yielding sample mean $\bar{y} = 10$

You do a one-tailed two sample z-test against

$H_0: \quad \mu_X - \mu_Y = 5$

(A) Under the null hypothesis, what is the distribution of $(\bar{X} - \bar{Y})$?

$\sqrt{7^2/50 + 9^2/40}$

$\bar{X} \sim \text{Normal}(\mu_X, 7/\sqrt{50})$

$\bar{Y} \sim \text{Normal}(\mu_Y, 9/\sqrt{40})$ $\Rightarrow (\bar{X} - \bar{Y}) \sim \text{Normal}(5, \sqrt{49/50 + 81/40})$

(B) What is the z-score of the test statistic?

$$z = \frac{(\bar{x} - \bar{y}) - 5}{\sqrt{49/50 + 81/40}} = \frac{5}{\sqrt{49/50 + 81/40}}$$

6. *(15pts)* Complete the ANOVA table below and answer questions about its values.

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

(A) Total number of factors = $10 + 1 = 11$

(B) Total number of samples = $50 + 1 = 51$

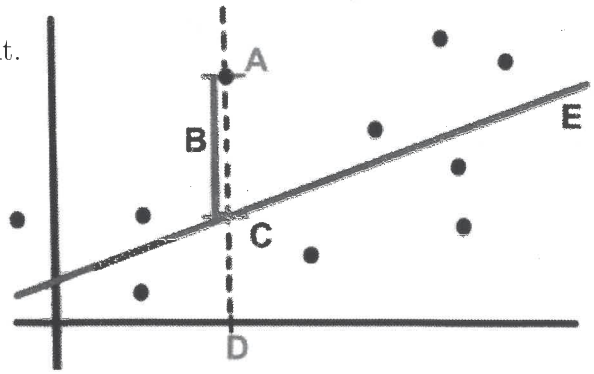
(C) Pooled sample variance = $MSE = 2$

(D) Total sample variance = $MST = 3$

(E) What hypothesis has p-value $p = 0.00219$?

H_0: All factor means are identical.

7. *(5pts)* Give labels for the diagram to the right.



A. y_i "observed value"

B. ε_i "residual value"

C. $\hat{y}_i$ "fitted value"

D. x_i "sample point"

E. $\hat{y}(x)$ "regression line"

6. *(10pts)* Complete the regression tables below and answer questions about its values.

Predictor	Coeff	Std.Error	t	p
Const	1	0.310	3.225	0.00111
x	$\frac{1}{2}$	0.224	2.236	0.0298

Note: These are equal. Both test $H_0: \beta_1 = 0$.

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

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

Note: This is $\sqrt{MSE}$ "pooled sample std. dev"

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

$\hat{\beta}_0 = 1$, $\hat{\beta}_1 = 1/2$ $\Rightarrow$ $\hat{y}(x) = 1 + \frac{1}{2}x$

(B) What hypothesis has p-value $p = 0.00111$?

$H_0: \beta_0 = 0$

(C) What hypothesis has p-value $p = 0.0298$?

$H_0: \beta_1 = 0$