Stat 3022: Midterm Exam 1

March 5 (Tuesday), 2013

- **Name**:
- **ID number**:
- This exam must be your own work entirely. You can not talk to or share information with anybody. You are not allowed to share materials, and calculators.
- Cell phone must be turned off.
- You have 50 minutes to complete the exam.

Problem 1 (21 points total, 3 points each)

Choose one of the listed choices for each question (no explanation is needed), put your answers in the table on page 4.

1. Suppose you have a sample $x_1, x_2, \ldots, x_n$ from a population. Which of the following is **NOT a random variable**?

(A). population mean

(B). sample variance

(C). sample mean

(D). t-statistic

2. In paired t-test with sample size $n_1 = n_2 = 20$, you have $H_0 : \mu = 0$ vs. $H_a : \mu < 0$, the t-statistic is 3.4. what is the p-value?

(A). `1 - pt(3.4, 20)`

(B). `1 - pt(3.4, 19)`

(C). `pt(3.4, 20)`

(D). `pt(3.4, 19)`

3. Suppose y is response, x_1 is numerical predictor, and x_2 is categorical predictor with 2 levels. Which of the following R code will generate parallel line models?

(A). `lm(y ~ 1)`

(B). `lm(y ~ x1)`

(C). `lm(y ~ x1 + x2)`

(D). `lm(y ~ x1 * x2)`

4. In paired t-test with $H_0 : \mu = 0$, a 95% confidence interval for the mean of difference is (-0.035, 0.057). The corresponding t-test (two-sided H_a) would:

(A). reject H_0 at the α=0.05 significance level.

(B). fail to reject H_0 at the α=0.05 significance level.

(C). can't tell without more information.

5. For what type of experiment you can make causal inference, according to chapter 1 in the textbook?

(A). Randomized experiment

(B). Observational experiment

(C). Neither

(D). Both

6. When given a data set for regression analysis, which one of the following is usually done first?

(A). ANOVA F-test

(B). Log transformation

(C). Graphical analysis

(D). Coefficients estimation

7. In the following table, which cell corresponds to the type I Error?

	H_0 is true	H_a is true
Reject H_0	(1)	(2)
Do not reject H_0	(3)	(4)

(A). (1)

(B). (2)

(C). (3)

(D). (4)

8. Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma)$. $\bar{X}$ is the sample mean. s is the sample standard deviation. Then $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ follows a:

(A). t-distribution with $n-2$ degrees of freedom

(B). standard Normal distribution

(C). t-distribution with $n-1$ degrees of freedom

(D). $N(0, \sigma/\sqrt{n})$

9. In the simple linear regression, a Q-Q plot is used to check:

(A). normality assumption

(B). unequal variances

(C). significance of the explanatory variable

(D). all three above

10. Which of the following statements (taken individually) is always true?

(A). The significance level α cannot be larger than 0.05.

(B). If we reject H_0 at the $\alpha = 0.05$ significance level, then we would reject H_0 at the $\alpha = 0.01$ significance level.

(C). If we fail to reject H_0 at the $\alpha = 0.05$ significance level, then we would also fail to reject H_0 at the $\alpha = 0.01$ significance level.

(D). A p-value of 0.0988 always indicates that there is no evidence to reject H_0.

Put your answers for multiple choice problems here (use capital letters):

1	2	3	4	5	6	7	8	9	10
A	D	C	B	A	C	A	C	A	C

Problem 2 (30 points)

(Fish Oil and Blood Pressure data.) Researchers used 7 red and 7 black playing cards to randomly assign 14 volunteer males with high blood pressure to one of two diets for four weeks: a fish oil diet and a standard oil diet. The reductions in diastolic blood pressure are recorded. A summary of the data is shown on the next page. Some useful R output is also shown.

Diet	Sample size	Average	Sample Standard Deviation (SD)
Fish oil	$n_1 = 7$	$\bar{Y}_1 = 6.57$	$s_1 = 5.86$
Regular oil	$n_2 = 7$	$\bar{Y}_2 = -1.14$	$s_1 = 5.02$

```
> pt(0.975, df = 6)
[1] 0.8163927
> qt(0.975, df = 6)
[1] 2.446912
> pt(2.97, df = 6)
[1] 0.9875212
> pt(1.12, df = 6)
[1] 0.8472333
> pt(6.787, df = 12)
[1] 0.9999903
> pt(2.644, df = 12)
[1] 0.9892925
```

(a). [6 points] Compute the standard error $SE(\bar{Y}_1)$ for the average reduction of the

fish oil group.

Answer:

$$SE(\bar{Y_1}) = \frac{s_1}{\sqrt{n_1}} = \frac{5.86}{\sqrt{7}} = 2.21$$

(b). [2 points] What is the degree of freedom associated with the above standard error $SE(\bar{Y_1})$?

Answer:

$$n_1 - 1 = 6$$

(c). [8 points] Construct a 95% confidence interval for μ_1, where μ_1 is the population mean of the fish oil group.

Answer:

$$\begin{aligned} &\bar{Y_1} \pm t_{0.975,6} \times SE(\bar{Y_1}) \\ &= 6.57 \pm 2.447 \times 2.21 \\ &= (1.16, 11.98) \end{aligned}$$

(d). [8 points] Compute the t-statistic for testing the hypothesis (two sample)

$$H_0 : \mu_1 - \mu_2 = 0 \textbf{ vs. } H_a : \mu_1 - \mu_2 \neq 0$$

where μ_1 is the population mean of the fish oil group and μ_2 is the population mean of the regular oil group. Answer:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{6 \times 5.86^2 + 6 \times 5.02^2}{12}} = 5.456$$

$$\bar{Y}_1 - \bar{Y}_2 = 7.71$$

$$SE(\bar{Y}_1 - \bar{Y}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 5.456 \times \sqrt{\frac{2}{7}} = 2.92$$

$$t = \frac{7.71}{2.92} = 2.64$$

(f). [7 points] Find the p-value for the above test. What is your statistical conclusion? Answer:

$$p - value = 2(1 - pt(2.644, df = 12) = 2 \times (1 - 0.98929) = 0.02142$$

Since this p-value is less than 0.05, there is strong evidence that there is a difference between the means of two groups. Or there is strong evidence to reject H_0.

Problem 4 (30 points)

Black wheatears, Oenanthe leucura, are small birds of Spain and Morocco. Males of the species demonstrate an exaggerated sexual display by carrying many heavy stones to nesting cavities. This 35-gram bird transports, on average, 3.1kg of stones per nesting season. Different males carry somewhat different sized stones, prompting a study of whether larger stones may be a signal of higher health status. M. Soler et al. calculated the average stone mass (g) carried by each of 21 male black wheatears, along with T-cell response measurements reflecting their immune systems' strengths.

```
> summary(data)
      Mass             Tcell
 Min.   :3.330   Min.   :0.183
```

```
 1st Qu.:6.290   1st Qu.:0.251
 Median :6.810   Median :0.312
 Mean   :7.204   Mean   :0.324
 3rd Qu.:8.180   3rd Qu.:0.411
 Max.   :9.950   Max.   :0.508
> m <- lm(Mass ~ Tcell, data)
> summary(m)

Call:
lm(formula = Mass ~ Tcell, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1429 -0.7327  0.3448  0.7472  3.2736

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.911      1.112   3.517  0.00230 **
Tcell         10.165      3.296   3.084  0.00611 **
---

Residual standard error: 1.426 on 19 degrees of freedom
Multiple R-squared: 0.3336,Adjusted R-squared: 0.2986
F-statistic: * on 1 and 19 DF,  p-value: 0.006105
> qt(0.975, 19)
[1] 2.093024
> qt(0.975, 18)
[1] 2.100922
> qt(0.95, 19)
[1] 1.729133
> qt(0.95, 18)
[1] 1.734064
```

(a). [3 points] Write down the regression line.

Answer:

$$\mu\{Mass|Tcell\} = 3.911 + 10.165 \times Tcell$$

(b). [3 points] Construct a 95% confidence interval for β_1, the slope of the regression.
Answer: $10.165 \pm 2.093 \times 3.296 = [3.266, 17.064]$

(c). [3 points] Suppose you want to perform two-sided t-test with $H_0 : \beta_1 = 10$. Calculate the t-statistic.
Answer: $\frac{10.165-10}{3.296} = 0.05$

(d). [3 points] Calculate RSS (Residual Sum of Squares).
Answer: $\text{RSS} = \hat{\sigma}^2 \times \text{d.f.} = 1.426^2 \times 19 = 38.64$

(e). [3 points] If someone want to estimate the value of "Mass" when "Tcell" is 1, is the estimation reliable? Why? (you don't need to calculate the estimated value)
Answer: No. Because of extrapolation. $1 > \max(\text{Tcell})$

(f). [3 points] Calculate the (1) and (2) in the following output.

```
> predict(m, data.frame(Tcell=0.5), interval="predict")
       fit      lwr      upr
1      (1) 5.706761     (2)
```

Answer:

$$\text{fit} = 3.991 + 10.165 \times 0.5 = 8.99$$

$$\text{upr} = 8.99 + (8.99 - 5.706) = 12.28$$

(g). [4 points] Consider equal-mean model $\mu(\text{Mass}|\text{Tcell}) = \mu$. From the R output of simple linear regression, do you think the equal-mean model is good enough? Why?
Answer: No. The p-value for testing $\beta_1 \neq 0$ is less than 0.05, or equivalently the p-value for F-test < 0.05.

(h). [4 points] R-squared is only 0.3336. What do you suggest we do next?
Answer: Make the residual plot and other graphs, to find out if the small R^2 is due to non-linearity or non-constant variance or some other reasons.

(i). [4 points] There are two possible regression models:

```
m1 <- lm(Mass ~ Tcell, data)
m2 <- lm(Tcell ~ Mass, data)
```

Re-read the description of the problem. Which model is more appropriate for the purpose of the study, $m1$ or $m2$? Why?
Answer: $m2$. It is a study of whether larger stones may be a signal of higher health status, so you want to use 'Mass' to predict 'Tcell'.