UNIVERSITY OF VICTORIA
EXAMINATIONS APRIL 2015
STATISTICS 359/563, BIOLOGY 563
DATA ANALYSIS

Name:_____ Student No.:_____

Section:_____

TEST QUESTIONS TO BE ANSWERED ON EXAM PAPER
Duration: 3 hours

Instructor:
F. Nathoo

STUDENTS MUST COUNT THE NUMBER OF PAGES IN THIS EXAMINATION PAPER BEFORE BEGINNING TO WRITE, AND REPORT ANY DISCREPANCY IMMEDIATELY TO THE INVIGILATOR.

THIS QUESTION PAPER HAS 12 PAGES plus COVER PAGE and an APPENDIX.

Instructions:

1. Write your NAME and STUDENT NUMBER on your exam paper.

2. You will be required to show ID and sign the list during the exam.

Good luck!

Question	**Marks**	**Value**
1		10
2		10
3		5
4		8
5		9
6		14
7		9
8		7
9		4
Total		76

Question 1. Using R, I simulate a random sample of size $n = 1000$ from four different distributions:

- $Uniform(-2, 2)$
- $N(0, 1)$
- t_4
- $Poisson(\mu = 5)$

The histograms (in no particular order) corresponding to these samples are displayed in Figure 1, panels (a) to (d). Based on these histograms, indicate which of the four distributions listed above corresponds to each of the histograms (a) to (d).

Histogram (a): $N(0,1)$

Histogram (b): $U(-2,2)$

Histogram (c): $Poisson(5)$

Histogram (d): t_4

Figure 2 panels (e) and (f) display normal Q-Q plots for two of the samples generated. Based on these plots, indicate which of the four distributions listed above corresponds to each of the Q-Q plots in (e) and (f).

(e) t_4

(f) $U(-2,2)$

(g) I calculate the *skew* of the $Poisson(\mu = 5)$ sample and get a value of $skew = 0.48$. If I simulate a *new* sample from a Poisson distribution but change the mean μ to $\mu = 25$, do we expect the *skew* of the new sample to be larger or smaller than the *skew* of the old sample? Do we expect the variance of the new sample to be larger or smaller than that of the old sample? Explain the rational for each of your answers?

- skew will be smaller (Poisson looks more normal as $\mu \uparrow$)
- variance will be larger since variance = mean and $\mu \uparrow$

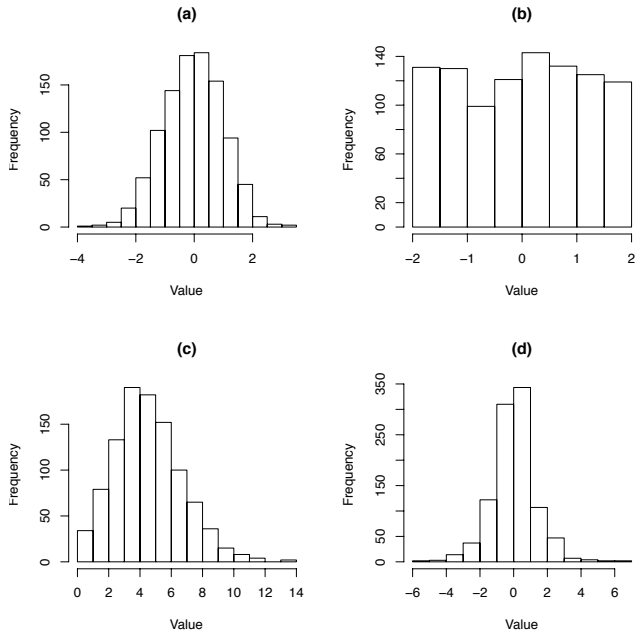


Figure 1: Histograms of the random samples of size 1000 drawn from four different distributions.

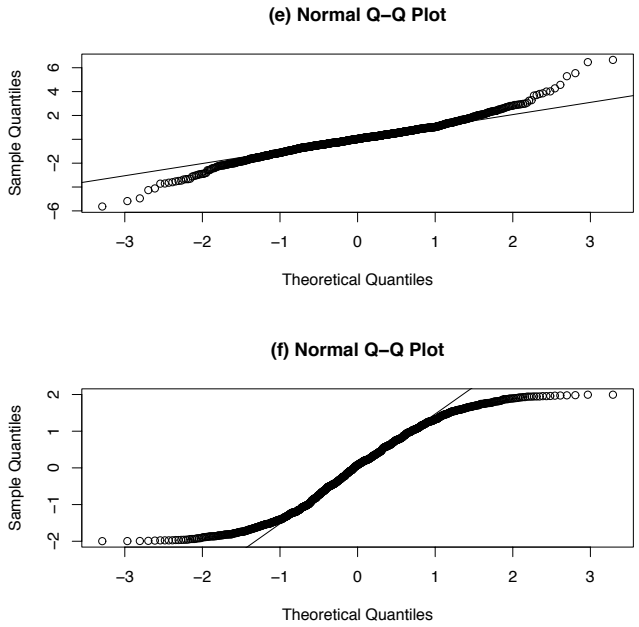


Figure 2: Normal Q-Q plot of the random samples of size 1000 drawn from two of the distributions.

Question 2.

Assume samples $Y_{11}, \ldots, Y_{1m}$ and $Y_{21}, \ldots, Y_{2n}$ are collected from two populations, and it is of interest to compare the population means through the hypothesis test

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

using a two sample pooled t-test.

1. State the four primary assumptions underlying this test procedure.

- equal variance in both pops
- independent observations
- normal distributed observations

2. Give the form of the test statistic T, and state its distribution under the assumption that $\mu_1 = \mu_2$.

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\left[\left(\frac{1}{M} + \frac{1}{N}\right)\left(\frac{M-1}{M+N-2} S_1^2 + \frac{N-1}{M+N-2} S_2^2\right)\right]^{1/2}}$$

$\cdot\ t_{M+N-2}$

3. Suppose $m = n = 9$, and from the data we observe $T^{(obs)} = 3$, which on applying the t.test() function in R yields a p-value = 0.0085. Give the R command that will compute this p-value from the observed value of the test statistic $T^{(obs)} = 3$.

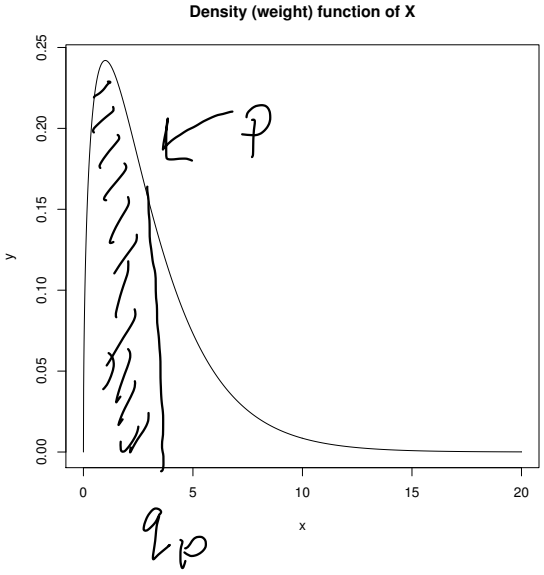
```
2(1-pt(abs(3), df=16))
```

4. The p-value< 0.05 so we reject H_0 at a $\alpha = 0.05$ level. In words, how does one interpret a p-value? For example, what does it mean in this example where p-value = 0.0085?

- under repeated samplings and assuming H_0 is true, the probability of observing something at least as extreme of $T^{(obs)} = 3$ is 0.0085.

Question 3.

Suppose X is a continuous random variable (measurement) whose density (weight) function is depicted in Figure 1. For some number p with $0 < p < 1$, give the definition of q_p, the p^{th} quantile of X, and draw a diagram illustrating q_p on Figure 1.



$P_r(X \le q_p) = p$

Figure 3: Provide an illustration of q_p on this figure.

Suppose the random variable $Z \sim N(0,1)$ and $X \sim \chi^2_4$ with Z and X independent. What is the distribution of Z^2? What is the distribution of $4Z^2/X$?

$\chi^2_{(1)}$

$F_{1,4}$

Question 4.
Given two samples $Y_{11}, \ldots, Y_{1m}$ and $Y_{21}, \ldots, Y_{2n}$ collected from two populations, prior to conducting a two-sample t-test to compare the means of the two populations, it is often useful to conduct an F-test to compare the population variances

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

1. Give the form of the appropriate test statistic T, state two primary assumptions underlying this test, and state the distribution of the test statistic under the assumption that $\sigma_1^2 = \sigma_2^2$.

$T = \text{Max}\{S_1^2, S_2^2\} / \text{MIN}\{S_1^2, S_2^2\}$

- normal data, indepl observations

- $F_{M-1, N-1}$

2. Suppose $m = n = 9$, and from the data we observe $T^{(obs)} = 2$, which on applying the var.test() function in R yields a p-value = 0.173. Give the R command that will compute this p-value from the observed value of the test statistic $T^{(obs)} = 2$.

1 - pf(2, df1 = 8, df2 = 8)

3. If we reject H_0 in this initial step, what kind of t-test should be applied? Why is it better than the alternative?

- Welch t-test
- Pooled t-test would be best on an incorrect assumption

Question 5. A credit card company collects data on the number of card holders who purchase protection insurance. In an effort to develop a better marketing strategy, the number of new card holders who purchase or do no purchase insurance is cross-classified by age, giving the data below. Here, it is of interest to examine the data for an association between age and the purchase of protection insurance.

Age	**Purchased insurance**	**Did not purchase insurance**
Age <30	900	985
$30 \leq$ Age < 50	500	521
$50 \leq$ Age	187	124

1. Assuming that there is no association between age and the purchase of insurance, compute a table of 'expected' counts.

Age	Purchased	Did not
< 30	929.9	955.1
30 ≤ Age < 50	503.67	517
50 ≤	153.4	157.6

2. By hand, compute the observed value of the test statistic for testing association.

$$\sum\sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 16.47$$

3. Assuming there is no association, what is the distribution of the test statistic?

χ^2_2

4. Give the R command that will compute the p-value from the observed value of the test statistic.

```
1-pchisq(16.47,df=2)
```

Question 6.

Consider a simple linear regression relating Y and X applied to $n = 50$ observations.

$$Y_i = a + bX_i + \epsilon_i, \;\; \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \;\; i = 1, \ldots, n$$

Suppose we are interest in testing $H_0 : b = 0$ Vs $H_1 : b \neq 0$ and we have the following incomplete ANOVA table:

Source	Sum of Squares	Degrees of Freedom	Mean Squares	F-ratio
Regression	?517.48	? 1	?517.48	136.68
Error	?181.74	? 48	?3.79	
Total	699.22	?49		

1. Fill in the seven missing values in the table above.

(1) $F = \frac{MSR}{MSE} = \frac{SSR}{SSE/48} = \frac{48\,SSR}{SSE} = 136.68$

(2) $SST = 699.22 = SSE + SSR \rightarrow SSE = (699.22 - SSR)$

(1) $48\,SSR = (699.22 - SSR)\,136.68$

$\rightarrow SSR = 517.48$

2. Compute the value of R^2, and give an interpretation of this number.

$R^2 = \frac{SSR}{SST} = 0.74$

- proportion of variation in Y explained by regression line

3. Describe two situations involving a response Y and a covariate X where the value of R^2 will not be particularly meaningful.

- outlier
- non-linear

4. Consider adding a second covariate Z_i ($= 0$ or $= 1$) to the model

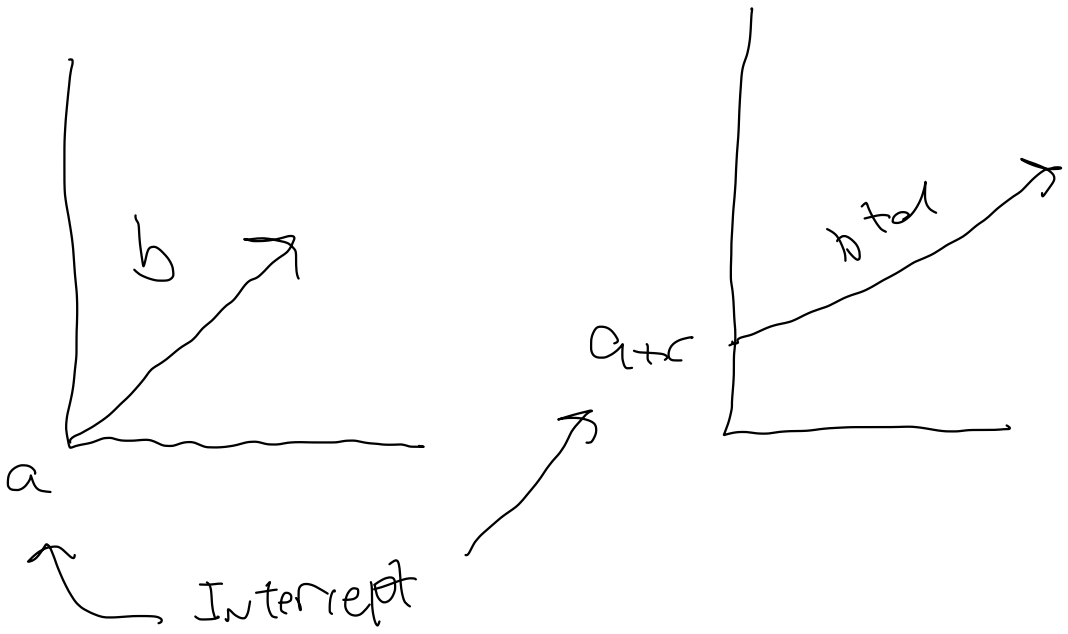
$$Y_i = a + bX_i + cZ_i + \epsilon_i, \;\; \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \;\; i = 1, \ldots, n.$$

Explain how the interpretation of b differs in this model, when compared to the simpler model that excludes Z_i.

- association between Y and X has now been adjusted for Z

5. Carefully describe (using a diagram if you wish) the relationship between Y and X in the following model

$$Y_i = a + bX_i + cZ_i + d\,Z_i * X_i + \epsilon_i, \;\; \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \;\; i = 1, \ldots, n.$$



Question 7. Consider a multiple regression relating a response Y to three explanatory variables X_1, X_2, X_3 applied to $n = 100$ observations.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{3i}^2 + \epsilon_i, \;\; \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \;\; i = 1, \ldots, n$$

R output obtained from fitting such a model is displayed on page 1 of the Appendix.

1. Based on the output, give the R command to compute the p-value (=0.0251) for testing $H_0 : \beta_4 = 0$ Vs $H_1 : \beta_4 \neq 0$.

   2*(1-pt(abs(2.276), df=95))

2. Based on the output, give the R command to compute a 90% confidence interval (represented by a vector with two elements) for β_1.

   c(-1.049 - qt(0.95, df=95)*0.058, -1.049 + qt(...))

3. Based on the output, give a clear and careful interpretation of the estimated relationship between Y and X_3.

   - Y increases with X_3 in a quadratic manner

4. For the data being considered, the explanatory variables for the first subject are given by $X_{11} = -1.5$, $X_{21} = -0.49$ and $X_{31} = 0.14$, and we have $Y_1 = 0$. Based on the output, compute the value of the corresponding residual $\hat{\epsilon}_1$ for this subject.

   $\hat{\beta}_0 + \hat{\beta}_1 X_{11} + \hat{\beta}_2 X_{21} + \hat{\beta}_3 X_{31} = 3.77$

   So assuming $Y_1 = 0$

   $\hat{\epsilon}_1 = -3.77$

Question 8.

Consider the three-parameter nonlinear regression model

$$Y_i = \beta X_{1i}^{\alpha} X_{2i}^{\gamma} + \epsilon_i, \;\; \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \;\; i = 1, \ldots, n$$

where Y_i is the response variable; $X_{1i} > 0$ and $X_{2i} > 0$ are positive valued explanatory variables; and β, α, γ are unknown parameters.

1. Write down an expression for the residual sum of squares.

$$RSS = \sum_{i=1}^{n} (Y_i - \beta X_{1i}^{\alpha} X_{2i}^{\gamma})^2$$

2. Carefully write down a system of nonlinear equations whose solution defines the least squares estimator.

$$\frac{\partial RSS}{\partial \beta} = 0, \quad \frac{\partial RSS}{\partial \alpha} = 0, \quad \frac{\partial RSS}{\partial \gamma} = 0$$

Question 9.

1. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \;\; \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \;\; i = 1, \ldots, n$$

where Y_i is a continuous response, and X_i is an explanatory variable. Under this model, what is the mean of Y_i? What is the variance of Y_i?

$E[Y_i] = \beta_0 + \beta_1 X_i$

$Var[Y_i] = \sigma^2$

2. Consider the Poisson log-linear regression model

$$Y_i \overset{ind}{\sim} \text{Poisson}(\lambda_i), \;\; i = 1, \ldots, n$$

$$log(\lambda_i) = \beta_0 + \beta_1 X_i$$

where Y_i is a count response, and X_i is an explanatory variable. Under this model, what is the mean of Y_i? What is the variance of Y_i?

$E[Y_i] = \lambda_i = \exp\{\beta_0 + \beta_1 X_i\}$

$Var[Y_i] = \lambda_i = \exp\{\beta_0 + \beta_1 X_i\}$

END