STAB22F:1998 Final Solutions

Name:

Student Number:

Tutorial:

Time: 3 hours
Aids: The exam is open book. Students may use any notes, books and calculators in writing this exam.

1. (10 marks)
Suppose a random variable X has the following distribution.

x	-2	-1	1	2
$P(X = x)$	.2	.3	.1	.4

Indicate how you would calculate the mean and variance of the random variable $Y = X^2$.

The mean of X is given by $\mu = E[X] = .2(-2)^2 + .3(-1)^2 + .1(1)^2 + .4(2)^2$ $= .6(4) + .4(1)$ and the variance is given by $\sigma^2 = Var[X] = .4(1-\mu)^2 + .6(4-\mu)^2$.

It also acceptable to give the Minitab commands that provide these answers.

2. Suppose that $X_1 \sim N(5, 3)$ and $X_2 \sim N(2, 1)$ are statistically independent.
(a) (5 marks) What is the distribution of $Y = X_1 + 2X_2$?

$Y \sim N(5 + 2(2), \sqrt{9 + 4})$

(b) (5 marks) What is the conditional distribution of X_1 given X_2?

The conditional distribution of X_1 given X_2 is $N(5, 3)$.

3. (10 marks) Suppose you wanted to make inference about the proportion of students at the University of Toronto who have ever attended a university football game. Based on a sample of 500 students from this population indicate how you would construct an interval that contained the true proportion with probability approximately equal to .99 in repeated sampling. Justify your answer.

First we construct the sample proportion

$$\hat{p} = \frac{\text{\# in sample who have attended a university football game}}{500}$$

as an estimate of the true proportion p. Because $\hat{p}$ is an average with $E[\,\hat{p}] = p, Var[\,\hat{p}] = p(1-p)/500$ the Central Limit Theorem says that

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/500}} \approx N(0,1).$$

An extended version of this theorem says that

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/500}} \approx N(0,1).$$

Then letting $z_{.995}$ be the point that has .995 of the probability to its left for the $N(0,1)$ distribution we have that

$$\begin{aligned} .99 &\approx P\left(-z_{.995} \le \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/500}} \le z_{.995}\right) \\ &= P\left(\hat{p} - \sqrt{\hat{p}(1-\hat{p})/500}\, z_{.995} \le p \le \hat{p} + \sqrt{\hat{p}(1-\hat{p})/500}\, z_{.995}\right) \end{aligned}$$

and so $\left(\hat{p} - \sqrt{\hat{p}(1-\hat{p})/500}\, z_{.995}, \hat{p} + \sqrt{\hat{p}(1-\hat{p})/500}\, z_{.995}\right)$ is an approximate .99 confidence interval for p.

4. (10 marks) Suppose that $X_1 \sim$ Student(3) statistically independent of $X_2 \sim$ Student(6). Indicate how you would carry out a simulation in Minitab to approximate $E\left[3X_1^2 - 4X_2^3\right]$. Also show how you would assess the accuracy of this approximation.

The quantities k1 and k2 printed out using the Minitab commands below respectively give the estimate of this quantity and the standard error of the estimate.

```
MTB >random 10000 c1;
SUBC>student 3.
MTB >random 10000 c2;
SUBC>student 6.
MTB >let c3=3*c1**2-4*c2**3
MTB >let k1=mean(c3)
MTB >let k2=stdev(c3)
MTB >let k2=k2/sqrt(10000)
MTB >print k1 k2
Data Display
K1 8.32493
K2 0.735533
```

Since the interval given by estimate ± 3(standard errors) contains the true value of the quantity with virtual certainty the standard error gives a measure of the uncertainty in the approximation.

5. A sample was taken of size $n = 100$ from the population of students at a university and each student was given a test measuring their ability at basic mathematics. The mean score (out of 100) of the sample was 75 and the standard deviation of the sample was 5.
(a) (5 marks) What is the standard error of the estimate of the population mean μ? How do we interpret the standard error?

The standard error is given by $5/10 = .5$. The interval given by the estimate plus or minys three standard errors contains the true value of the mean with virtual certainty.

(b) (5 marks) Suppose you want to test the null hypothesis $H_0 : \mu = 72$ versus $H_a : \mu \neq 72$. Indicate how you would do this using Minitab and discuss how you justify this test.

The Minitab commands are

```
MTB >let k1=absolute((75-72)/.5)
MTB >cdf k1;
SUBC>normal 0 1.
Cumulative Distribution Function
Normal with mean = 0 and standard deviation = 1.00000
 x       P( X <= x)
 6.0000     1.0000
MTB >let k2=2*(1-1)
MTB >print k2
Data Display
K2 0
```

where k1 gives the value of the test statistic and k2 gives the P-value. This test is justified by the CLT since for large n we have that

$$\frac{\bar{X} - 72}{S/\sqrt{n}}$$

is distributed approximately $N(0, 1)$ when the true mean is 72. The P-value is a measure of how reasonable the observed value of this quantity is with respect to the standard normal distribution.

(c) (5 marks) Suppose you wanted to estimate the power of the test you performed in (b) at the $N(75, 5)$ distribution. Show how you would do this using Minitab. Also discuss why it is important to evaluate the power of a test..

The estimate of the power of the test, when we decide to reject at the .05 level, is given in k2 below.

```
MTB >let k1=5/10
MTB >random 10000 c1;
SUBC>normal 75 k1.
MTB >let c2=absolute((c1-72)/.5)
MTB >invcdf .975;
SUBC>normal 0 1.
Inverse Cumulative Distribution Function
Normal with mean = 0 and standard deviation = 1.00000
 P( X <= x)       x
 0.9750      1.9600
MTB >let c3=c2>1.96
MTB >let k2=mean(c3)
MTB >print k2
Data Display
K2 1.00000
```

The power of the test at different values in the alternative indicates how sensitive the test is to detecting deviations from the null hypothesis. The fact that we do not reject a null hypothesis is not necessarily of much value if the probability of rejecting is quite low at practically meaningful values in the alternative.

6. Suppose we have samples of size $n_1 = n_2 = n_3 = 20$ respectively from each of three populations and we measure the quantitative variable Y for each sampled element. Further suppose the investigator wants to compare the population distributions of Y.
(a) (5 marks) If the investigator decides to use the one-way ANOVA for this then discuss the assumptions that are involved in using this method.

The followinfg assumptions are used

(1) The response from the i-th population is assumed to take the form $Y = \beta_i + Z$ where Z is a random variable with mean 0 and unknown standard deviation σ.

(2) Z is assumed to be normally distributed.

(3) the responses are assumed to be statistically independent.

(b) (5 marks) How would you check that the assumptions discussed in (a) are not being violated?

After fitting the model we compute the standardized residuals. The standardized residuals are plotted against the predictor variable; i.e. we examine the scatter plots of the standardized residuals at each level of the predictor. These plots should look like samples from the $N(0, 1)$ distribution. Also we do a normal probability plot of the standardized residuals. This plot should give approximately a straight line. If these plots look flawed then there is likely an assumption that is being violated.

(c) (10 marks) For this data suppose the following (partial) output is obtained from Minitab where the responses are in c4 and the indicator for population are in c5. Indicate what hypothesis is being tested in the ANOVA table, what the outcome of this test is and what you subsequently conclude.

```
Analysis of Variance for C4
Source DF      SS      MS     F     P
C5      2  54.245 27.123 32.98 0.000
Error  57  46.879  0.822
Total  59 101.124
 Individual 95% CIs For Mean
 Based on Pooled StDev
Level N     Mean StDev
1    20  1.7821 0.7170
2    20  2.1561 0.9599
3    20 -0.0218 1.0157

Pooled StDev = 0.9069
Fisher's pairwise comparisons
 Family error rate = 0.0515
Individual error rate = 0.0200
Critical value = 2.394
Intervals for (column level mean) - (row level mean)
        1      2
 2 -1.0606
    0.3125
 3  1.1173  1.4913
    2.4904  2.8644
```

The hypothesis being tested is $H_0 : \beta_1 = \beta_2 = \beta_3$ versus $H_a : \beta_i \neq \beta_j$ for some i and j. In this case the ANOVA test has a P-value of 0 so we definitel reject the null hypothesis. We then look at the 3 pairwise comparisons using the Fisher test with an individual error rate of .02 so that the family error rate is .05. In this case we reject $H_0 : \beta_1 = \beta_3$ and $H_0 : \beta_2 = \beta_3$ and accept $H_0 : \beta_1 = \beta_2$.

(d) (5 marks) Explain the meaning of Individual error rate and Family error rate in (c).

By individual error rate we mean that an individual null hypothesis is rejected with that probability when the null hypothesis is true. By Family error rate we mean the probability of rejecting at least one of the null hypotheses when all the null hypotheses are true.

7. Suppose an investigator is interested in determining whether or not a cause-effect relationship exists between a response variable Y and two quantitative predictor variables W_1 and W_2 in a particular population. Further the investigator decides that the model

$$Y = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \epsilon$$

is appropriate where $\epsilon \sim N(0, \sigma)$ and $\beta_0, \beta_1, \beta_2$ and σ are unknown.
(a) (5 marks) Discuss what conditions the investigator must ensure hold in the study if he/she wants to be able to assert that a cause-effect relationship exists.

To assert that a cause-effect relationship exists between the response and the predictor variables after the analysis is concluded the investigator must ensure that each value of W_1 and W_2 is assigned to an experimental units via randomization so that each possible assignment can occur with the same probability.

(b) (5 marks) Suppose the following output is obtained from fitting the model to $n = 25$ data points with the response in c4, W_1 in c1 and W_2 in c2. Further suppose that the conditions necessary for asserting a cause-effect relationship do indeed hold. What are your conclusions about the existence of the relationship and why?

```
MTB >regress c4 2 c1 c2
Regression Analysis
The regression equation is
C4 = 3.10 + 2.07 C1 + 4.69 C2
Predictor     Coef    StDev       T     P
Constant     3.102    1.007    3.08 0.005
C1          2.0708   0.3036    6.82 0.000
C2          4.6931   0.3036   15.46 0.000
S = 2.147 R-Sq = 92.8% R-Sq(adj) = 92.2%
Analysis of Variance
Source          DF       SS       MS      F       P
Regression       2  1315.67   657.83 142.75 0.000
Residual Error  22   101.38     4.61
Total           24  1417.05
```

The ANOVA table test the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ versus $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$. This is equivalent to testing H_0 : no relationship between the response and the predictors exists versus H_a : a relationsip exists. In this case the P-value is 0 so we reject H_0 and so conclude that a cause-effect relationship does indeed exist.

(c) (5 marks) How good is this model at predicting the value of the response from the predictor variables? Justify your answer.

The value of the coefficient of determination is given by $R^2 = 92.8$ which indicates that approximately 935 of the observed variation in the response is explained by the predictors. So this indicates that the model will perform excellently for predicting future values of the response when the predictors take values in the range of those used in this study.

(d) (5 marks) What checks would you do to ensure that your inferences with this model make sense?

The standardized residuals should be plotted against each of the predictors to ensure that this looks approximately like what you would get when plotting a sample of 25 from a $N(0, 1)$ distribution. Also we should do a normal probability plot to see if we get something that is approximately a straight line.