STATISTICS 401 **FINAL EXAM** Spring 2000

SOLUTIONS

Problem 1. **[8 points]** An insurance company believes that people can be divided into two classes—those who are accident-prone and those who are not. Their statistics show that an accident-prone person will have an accident at some time within a fixed 1-year period with probability .4, whereas this probability decreases to .2 for a non-accident-prone person. Assume that 30 percent of the population is accident prone.

(a) [4] What is the probability that a new policy holder will have an accident within a 1-year period?

Define events

$$\begin{aligned} A &= \{\text{Person has an accident within a 1-year period}\} \\ B &= \{\text{Person is accident-prone}\}. \end{aligned}$$

Then

$$\begin{aligned} P(A) &= P(A \text{ and } B) + P(A \text{ and } B') \\ &= P(A|B)P(B) + P(A|B')P(B') \\ &= .4(.3) + .2(.7) \\ &= .26 \end{aligned}$$

(b) [4] Given that a new policy holder has an accident within a year of purchasing a policy, what is the probability that he or she is accident-prone?

$$\begin{aligned} P(B|A) &= \frac{P(A \text{ and } B)}{P(A)} \\ &= \frac{P(A|B)P(B)}{P(A)} \\ &= .4(.3)/.26 \\ &= .46 \end{aligned}$$

Problem 2. **[16 points]** A number from $\{1, 2, 3\}$ will be selected at random but with unequal probabilities: The probability that i will be selected is $i/6$.

(a) [4] Show that these probabilities give a legitimate probability mass function.

The probabilities must sum to 1:

$$\frac{1}{6} + \frac{2}{6} + \frac{3}{6} = 1$$

(b) [4] Calculate the expected value of the selected number.

Let X be the number selected. Then

$$\begin{aligned} \mathrm{E}(X) &= \sum_x x \cdot P(X = x) \\ &= 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + 3 \cdot P(X = 3) \\ &= \frac{1 + 4 + 9}{6} \\ &= 7/3 \end{aligned}$$

(c) [4] Complete the table below with values of the cumulative distribution function, $F(x)$:

x	0	1	1.5	2	2.5	3	3.5
$F(x)$	0	1/6	1/6	1/2	1/2	1	1

By definition, $F(x)$ equals $P(X \le x)$. For example,

$$F(2.5) = P(X \le 2.5) = P(X = 1 \text{ or } X = 2) = \frac{1}{6} + \frac{2}{6}$$

(d) [4] If number i is selected, you win $5 + 3i^2$ dollars. Find the expected amount you will win.

Let Y be the amount won. (In other words, $Y = 5 + 3X^2$.) Then Y takes values 8, 17, and 32 with probabilities 1/6, 2/6, and 3/6, respectively. Therefore,

$$\begin{aligned} \mathrm{E}(Y) &= \sum_y y \cdot P(Y = y) \\ &= 8(1/6) + 17(2/6) + 32(3/6) \\ &= 23 \end{aligned}$$

Problem 3. **[12 points]** Suppose that Y is a continuous random variable on the interval $(0, 4)$ such that its density is proportional to y. In other words, the density on $(0, 4)$ is ky.

(a) [4] Prove that $k = 1/8$.

The density must integrate to one. Therefore,

$$\int_0^4 ky\, dy = k\left(\frac{y^2}{2}\Big|_0^4\right) = 8k = 1.$$

Solving for k gives $k = 1/8$.

(b) [4] Find $P(Y < 1)$.

Integrate the density over the desired region to find the probability:

$$P(Y < 1) = \int_0^1 ky\, dy = \frac{1}{8}\left(\frac{y^2}{2}\Big|_0^1\right) = \frac{1}{8}\left(\frac{1}{2}\right) = \frac{1}{16}$$

(c) [4] Find $\mathrm{E}(Y)$.

If $f(y)$ denotes the density, then

$$\mathrm{E}(Y) = \int y \cdot f(y)\, dy = \frac{1}{8}\int_0^4 y^2\, dy = \frac{1}{8}\left(\frac{4^3}{3}\right) = \frac{8}{3}.$$

Problem 4. [12 points] Suppose Xtreme brand batteries have a mean lifetime of 50 hours (with variance 144) and Yowza brand batteries have a mean lifetime of 60 hours (with variance 121). Suppose that it is known that Yowza brand batteries' lifetimes are normally distributed, whereas the distribution of Xtreme brand batteries is unknown.

Suppose $X_1, X_2, \ldots, X_{100}$ is a random sample of Xtreme batteries and $Y_1, Y_2, \ldots, Y_{64}$ is a random sample of Yowza batteries.

(a) [4] Find the mean and standard deviation of $\overline{X}$.

The expected value of the sample mean from a simple random sample is the same as the population mean of the population. The variance of the sample mean equals the population variance divided by the sample size. Therefore,

$$\begin{aligned} \mu_{\overline{X}} &= 50 \\ \sigma_{\overline{X}} &= \sqrt{144/100} = 1.2 \end{aligned}$$

(b) [4] Find the mean and standard deviation of $\overline{Y} - \overline{X}$.

Expected value is linear, so

$$\mathrm{E}\left(\overline{Y} - \overline{X}\right) = \mathrm{E}\left(\overline{Y}\right) - \mathrm{E}\left(\overline{X}\right) = 60 - 50 = 10.$$

Since the two samples may be assumed independent of one another, the random variables $\overline{X}$ and $\overline{Y}$ are independent. Thus,

$$\begin{aligned} \mathrm{Var}\left(\overline{Y} - \overline{X}\right) &= \mathrm{Var}\left(\overline{Y}\right) + \mathrm{Var}\left(-\overline{X}\right) = \mathrm{Var}\left(\overline{Y}\right) + \mathrm{Var}\left(\overline{X}\right) \\ &= \frac{144}{100} + \frac{121}{64} = 3.331 \end{aligned}$$

Therefore, the standard deviation of $\overline{Y} - \overline{X}$ is $\sqrt{3.331} = 1.825$.

(c) [4] State the distributions of $\overline{X}$ and $\overline{Y}$ and whether each is exact or approximate.

By the central limit theorem, $\overline{X}$ is approximately normal (with mean 50 and standard deviation 6/5).

Because the Yowza batteries' lifetimes are normally distributed, $\overline{Y}$ is exactly normal (with mean 60 and standard deviation 11/8).

Problem 5. [12 points] Manatees are large, gentle sea creatures that live along the Florida coast. Many manatees are killed or injured by powerboats. Below are data on powerboat registrations (in thousands) and the number of manatees killed by boats in Florida for four different years between 1977 and 1990:

Boats (thousands)	498	526	559	614
Manatee deaths	16	15	34	33

Some Minitab output for this dataset is given below:

```
Predictor        Coef       StDev         T        P
Constant       -71.24       44.77     -1.59    0.253
ThouBoat      0.17432     0.08127      2.14    0.165

S = 7.017        R-Sq = 69.7%      R-Sq(adj) = 54.6%
```

(a) [4] Determine the equation of the estimated regression line.

Reading directly from the output, the equation is

$$\text{Deaths} = -71.24 + 0.174 \times (\text{Boats in thousands})$$

(b) [2] Use the estimated regression line to predict the number of manatee deaths in a year with 550,000 powerboat registrations.

Plug in to obtain

$$\text{prediction} = -71.24 + 0.174(550) = 24.5$$

Note that it is not necessary to round to a whole number because this is the predicted average number of deaths using the model.

(c) [4] Find a 99% confidence interval for the slope parameter.

The formula (for this and many other confidence intervals) is

$$\text{estimate} \pm t_{\alpha/2,df} \times (\text{standard error of estimate})$$

In this case, the estimate of β_1 is 0.174, the standard error of this estimate is 0.0813, $\alpha = .01$, and df $= n - 2 = 2$. From Table A.5, we find $t_{.005,2} = 9.925$. Thus the confidence interval is

$$0.174 \pm 9.925(.0813) \qquad \text{or} \qquad (-0.633, 0.981)$$

(d) [2] What proportion of the variability of the response is explained by the regression model?

The answer is provided by R^2, which is 69.7% in this case (this number comes directly from the output).

Problem 6. [12 points] In the manufacture of integrated circuit chips, an important variable is the width of the microscopic lines that are etched into the chip to serve as electrical conductors. Each wafer of chips contains a line-width test pattern used for quality control measurements. For one type of chip, the target width is $\mu = 3.00 \times 10^{-6}$ meters and the historical value of σ is 0.25×10^{-6}. Below are the average line widths for the last 25 samples of three wafers each, in time order reading across rows.

Avg. line widths, in 10^{-6} meters (in order, reading across)								
2.75	2.88	3.01	3.09	2.98	2.88	2.97	2.98	3.27
2.98	3.08	3.17	3.06	3.39	3.12	3.17	3.36	3.21
3.37	2.93	3.32	3.39	3.18	3.36	3.22		

(a) [4] Specify the upper and lower control limits (UCL and LCL) for an $\overline{X}$ chart (you do not have to create the chart).

The UCL is $\mu + 3\sigma/\sqrt{n} = 3.00 + 3(.25)/\sqrt{3} = 3.43$.
The LCL is $\mu - 3\sigma/\sqrt{n} = 3.00 - 3(.25)/\sqrt{3} = 2.57$.
Note that n is the size of each sample (3 in this case), not the number of samples considered. Units here are 10^{-6} meters.

(b) [4] Are there any points out of control? If so, which ones? If not, why not?

No. There are no points less than 2.57 nor greater than 3.43.

(c) [4] Based on Western Electric supplemental rule 2, shown below, is there any evidence that the process is out of control? If so, explain.

2. Four out of five successive points fall outside 1-sigma limits on the same side of the center line.

The 1-sigma limit above the center line is $3 + .25/\sqrt{3} = 3.14$. There are several examples of strings of five observations in a row for which four of the five are greater than 3.14 (for example, consider the last five observations, all of which are greater than 3.14). Thus, there is evidence that this process is out of control.

Problem 7. **[16 points]** In a poll conducted by Newstime Magazine, a random sample of 1500 Americans who will be eligible to vote in the next presidential election is asked whether they support campaign finance reform legislation.

(a) [8] Suppose John McCain is interested in running as George W. Bush's vice presidential candidate unless there is evidence that less than 60% of eligible voters support campaign finance reform legislation.

Suppose George W. Bush wants to make campaign finance reform legislation an issue only if there is evidence that more than 50% of eligible voters support it.

(i) State the null and alternative hypotheses that John McCain wants to test.

$$H_0 : p = .6 \qquad H_a : p < .6$$

(ii) State the null and alternative hypotheses that George W. Bush wants to test.

$$H_0 : p = .5 \qquad H_a : p > .5$$

(iii) What will John McCain do if his null hypothesis is rejected?

He will be uninterested in running as VP.

(iv) What will George W. Bush do if his null hypothesis is rejected?

He will make campaign finance reform an issue.

(b) [4] Suppose that in the poll above, 780 say they support campaign finance reform legislation. Give a p-value for the test of $H_0 : p = .5$ versus $H_a : p > .5$

This test is based on a z statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{(780/1500) - .5}{\sqrt{.5(.5)/1500}} = 1.55$$

With this z statistic, Table A.3 gives a 1-sided p-value to be $1 - .9394 = .0606$.

(c) [4] Next month, Newstime Magazine wants to estimate how many eligible voters intend to vote for Al Gore for president. They wish to give a 95% confidence interval for this proportion which is no wider than 4 percentage points (i.e., the interval should be plus or minus 2%). Assuming the Newstime editors have no idea about the approximate true proportion, how large a sample should they take?

If the true proportion is unknown, even approximately, we assume the worst case scenario, which is $p = .5$. This gives a value of n equal to

$$\frac{z^2}{w^2} = \frac{1.96^2}{.04^2} = 2401$$

Problem 8. [12 points] A large research project studied the physical properties of wood materials constructed by bonding together small flakes of wood. Three different species of trees (aspen, birch, and maple) were used, and the flakes were made in two different sizes (.015 inches by 2 inches and .025 inches by 2 inches). One of the physical properties measured was the tension modulus of elasticity in the direction perpendicular to the alignment of the flakes in pounds per square inch. There were three observations per cell.

The minitab output from an appropriate analysis is given below.

Analysis of Variance for Elastic

Source	DF	SS	MS	F	P
Species	2	146011	73006	5.95	0.016
Size	1	3308	3308	???	???
Interaction	2	41707	20854	1.70	0.224
Error	12	147138	12261		
Total	17	338164			

(a) [**4**] Calculate the value of the missing F statistic and test whether it is significant at the .05 level (show the critical value you use for the test).

The F statistic equals the ratio of the MS for size to the MS for error, which is $3308/12261 = .27$. We compare this to the critical F on 1 and 12 degrees of freedom (from Table A.9): $F_{.05,1,12} = 4.75$ Since the statistic is smaller than the critical value, it is not significant (i.e., we do not reject H_0, which means we have no evidence that Size impacts Elasticity).

(b) [**2**] Test at $\alpha = .05$ whether an additive model appears appropriate for these data. Explain. (No credit will be given for an answer with no explanation.)

The p-value for the interaction, .224, is not small enough to reject the null hypothesis that no interaction term is necessary. Thus, an additive model (i.e., with no interaction) appears appropriate.

(c) [**6**] Tukey's procedure with experimentwise error rate .05 is performed on the species factor. The three group means are given below.

Aspen	Birch	Maple
211.5	224.0	408.5

Using a Tukey procedure, tell which pairs of means are significantly different.

Following the steps on p. 452, we first find $Q_{\alpha,I,IJ(K-1)} = Q_{.05,3,12} = 3.77$. (Because $I = 3$, $J = 2$, and $K = 3$.) Multiplying this by $\sqrt{\text{MSE}/(JK)} = 45.2$ gives $w = 170.42$. Thus, any sample means that differ by more than 170.42 are significantly different. We conclude that Aspen and Birch are both significantly smaller than Maple (but Aspen and Birch are not significantly different from each other).