Math 171 Final Exam (Spring 2004) – Solutions

1. (a) Let X denote the player's net winnings. Then, $X = 7 - 3$ or $X = 0 - 3$ with corresponding probabilities

$$P(X = 4) = 0.4 \text{ and } P(X = -3) = 0.6.$$

Thus, by definition,

$$E(X) = 4 \cdot P(X = 4) + -3 \cdot P(X = -3) = 4 \cdot 0.4 - 3 \cdot 0.6 = -0.2.$$

Also, since $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{E(X^2) - [E(X)]^2}$, and

$$E(X^2) = 4^2 \cdot P(X = 4) + (-3)^2 \cdot P(X = -3) = 16 \cdot 0.4 + 9 \cdot 0.6 = 11.8,$$

we conclude that

$$\mathrm{SD}(X) = \sqrt{11.8 - (-0.2)^2} = \sqrt{11.76} \approx 3.43.$$

(b) If $\overline{X} = \dfrac{X_1 + \cdots + X_{100}}{100}$ denotes the player's average net winnings on 100 plays, where the distribution of each X_i is the same as X, then

$$E(\overline{X}) = E\left(\frac{X_1 + \cdots + X_{100}}{100}\right) = \frac{E(X_1) + \cdots + E(X_{100})}{100} = \frac{-0.2 \cdot 100}{100} = -0.2.$$

Since the 100 plays are independent,

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{X_1 + \cdots + X_{100}}{100}\right) = \frac{\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_{100})}{100^2} = \frac{11.76 \cdot 100}{100^2} = 0.1176.$$

Thus,

$$\mathrm{SD}(\overline{X}) = \sqrt{0.1176} \approx 0.343.$$

(c) By the central limit theorem, the distribution of $\overline{X}$ is *approximately* normal with mean -0.2 and standard deviation 0.343. Thus,

$$P(\overline{X} < 0) = P\left(\frac{\overline{X} - -0.2}{0.343} < \frac{0 - -0.2}{0.343}\right) = P(Z < 0.583) \approx 0.7190$$

where $Z \sim \mathcal{N}(0,1)$ and the probability is calculated using Table A.

(d) If the person spins the wheel over and over again, then by the law of large numbers, the limiting value of the average net winnings is simply the expected value of the average net winnings, namely -0.2. That is,

$$\lim_{n\to\infty} \overline{X} = \lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = -0.2.$$

2. (a) **(i)**The population of interest is *all customers of this national restaurant chain.* **(ii)** The sample consists of the 140 customers of the Ithaca branch who were selected in the national restaurant chain's simple random sample. **(iii)** The variable of interest is *quality of service.* It is *categorical* as there are 5 possibilites for it: poor, below average, average, above, average, outstanding.

(b) Let $\hat{p}$ be the proportion of customers who rated the quality of service as above average or outstanding. Then,

$$\hat{p} = \frac{67+19}{140} = \frac{86}{140}.$$

Thus, a 99% confidence interval for p is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.614 \pm 2.576 \cdot 0.0411 \approx (0.508, 0.720).$$

(c) The results from the given data *cannot* be extended to all other branches in the restaurant chain. This is simply because the data do not form a representative sample of the population of interest here. They *are* a simple random sample of the Ithaca branch's population base, but not of the restaurant chain as a whole. In order to extend results to all branches in the chain, it would be necessary to produce a simple random sample of customers from all the branches collectively. The results given can only be extended to the Ithaca branch; among other things, regional differences in personnel and customer expectations are likely. Thus, we are only 99% confident in concluding that over 50% of the Ithaca branch's customers rate the quality of service as above average or better.

3. (a) Suppose that p_a is the true, but unknown, proportion of African miners who died on the Gold Coast in 1936, and suppose that p_e is the true, but unknown, proportion of European miners who died on the Gold Coast in 1936. If we are interested in knowing whether the proportion of African miners who died on the Gold Coast in 1936 was higher than the proportion of European miners who died there that year, then the appropriate hypotheses are:

$$H_0 : p_a - p_e = 0 \text{ and } H_a : p_a - p_e > 0.$$

(b) We find that our estimators of p_a and p_e are

$$\hat{p}_a = \frac{223}{33809} \approx 0.006596 \text{ and } \hat{p}_e = \frac{7}{1541} \approx 0.004543, \text{ respectively.}$$

If we assume that H_0 is true, then we find that the *pooled* proportion is

$$\tilde{p} = \frac{223+7}{33809+1541} = \frac{230}{35350} \approx 0.006506.$$

Thus, the required test statistic is

$$\begin{aligned} z &= \frac{p_a - p_e}{\sqrt{\tilde{p}(1-\tilde{p})\left(\frac{1}{n_a}+\frac{1}{n_e}\right)}} \\ &\approx \frac{0.006596 - 0.004543}{\sqrt{0.006506(1-0.006506)\left(\frac{1}{33809}+\frac{1}{1541}\right)}} \\ &\approx 0.98 \end{aligned}$$

(c) From Table A, we find that the critical value of $z = 0.98$ corresponds to an upper tail probability of

$$p \approx 0.1635.$$

(d) Since $p > 0.005$, we are not able to reject H_0 at the $\alpha = 5\%$ significance level. Thus, there is insufficient *evidence* at the 5% significance level to conclude that the proportion of *African miners* who died on the *Gold Coast* in 1936 was higher than the proportion of *European miners* who died there that year.

4. (a) If we wish to test for some association between Math and Verbal scores, then the hypotheses are

$$H_0 : \beta = 0 \text{ and } H_a : \beta \neq 0$$

where β is the slope of the mean response line.

(b) There are several assumptions necessary for inference. We must assume that the true relationship is linear so that

$$\mu_y = \alpha + \beta x.$$

The observations must be made independently. Note that this is NOT the same as conducting a simple random sample from the *same* distribution because we are assuming that for each x, the observation y is normal with mean $\alpha + \beta x$ and standard deviation σ. Hence, another assumption is that for each x, the standard deviation of y is constant σ. Ideally, the residuals should be normally distributed. The histogram in **(iii)** provides strong evidence that this is a reasonable assumption. From **(ii)** we see that there is no apparent 'trend' to the residuals when plotted against Verbal score. That is, the regression line for Verbal vs. Residuals appears to have a slope of 0. Finally, the sample size was sufficiently large ($n = 162$), and there are only a few outliers, so the regression model seems reasonable.

(c) The test statistic is given by

$$t = \frac{b}{SE_b} = \frac{0.675075}{0.0568} \approx 11.885$$

with degrees of freedom $df = 162 - 2 = 160$. We can use Table C to find the corresponding critical value for $\alpha = 5\%$: for $df = 100$, we have $t^* = 1.984$, and for

$df = 1000$, we have $t^* = 1.962$. Thus, we see that our test statistic corresponds to a *very* small p-value. In fact, $p \ll 0.001 < 0.05$. Thus, we reject H_0 and conclude that there is sufficient *evidence* to assert that some *association* exists between scores on the *Math section* and the *Verbal section* of the SAT.

(d) For the data given, we compute that the equation of the linear regression line is

$$\hat{y} = 209.554 + 0.675075x$$

where the variable x represents *Verbal score* and the variable y represents *Math score.* Thus, if the verbal score is $x^* = 500$, then $\hat{y} = 209.554 + (0.675075)(500) = 547.0915$. Note that since

$$SE_b = \frac{s}{\sqrt{\sum(x-\overline{x})^2}}$$

we conclude that

$$\sum(x-\overline{x})^2 = \frac{s^2}{SE_b^2} = \left(\frac{71.75}{0.0568}\right)^2 = 1595684.915.$$

Hence, a 90% confidence interval for the mean Math score for all students with a verbal score of 500 is

$$\begin{aligned}
&\hat{y} \pm t^* s\sqrt{\frac{1}{n} + \frac{(x^*-\overline{x})^2}{\sum(x-\overline{x})^2}} \\
&= 547.0915 \pm 1.646 \cdot 71.75\sqrt{\frac{1}{162} + \frac{(500-596.296)^2}{1595684.915}} \\
&\approx 547.0915 \pm 12.9287.
\end{aligned}$$

(e) Using the linear regression line as above, if $x^* = 730$, then

$$\hat{y} = 209.554 + (0.675075)(730) = 702.35875.$$

Hence, a 90% prediction interval for Jennifer's Math score is

$$\begin{aligned}
&\hat{y} \pm t^* s\sqrt{1 + \frac{1}{n} + \frac{(x^*-\overline{x})^2}{\sum(x-\overline{x})^2}} \\
&= 702.35875 \pm 1.646 \cdot 71.75\sqrt{1 + \frac{1}{162} + \frac{(730-596.296)^2}{1595684.915}} \\
&\approx 702.35875 \pm 119.1221.
\end{aligned}$$

5. (a) The appropriate null hypothesis of interest here is:

H_0 : there is no relationship between the two categorical variables *vegetable plant type* and *sulfur dioxide leaf damage.*

Equivalently,

$$H_0 : \text{Leaf damage by sulfur dioxide is not associated with plant type.}$$

(b) To calculate the expected number of lettuce with severe leaf damage, for example, we have

$$\frac{40 \times 81}{120} = 27.$$

The rest of the table follows similarly:

	Severe leaf damage	Not severe or no leaf damage
Lettuce	27	13
Spinach	27	13
Tomato	27	13

(c) The Chi-square test statistic is

$$\frac{(32-27)^2}{27} + \frac{(8-13)^2}{13} + \frac{(28-27)^2}{27} + \frac{(12-13)^2}{13} + \frac{(21-27)^2}{27} + \frac{(19-13)^2}{13}$$
$$= \frac{62}{27} + \frac{62}{13} \approx 7.0655$$

and the degrees of freedom are $df = (r-1)(c-1) = (3-1)(2-1) = 2$. Thus, from Table E we find that the associated p-value satisfies

$$0.025 < p < 0.05.$$

(d) Since $p < 0.05$ we are able to reject the null hypothesis H_0 at the $\alpha = 10\%$ significance level. Thus, there is sufficient *evidence* to conclude that there is some relationship between vegetable *plant type* and *leaf damage* caused by sulfur dioxide. Equivalently, the Chi-square statistic with $df = 2$ and $\alpha = 0.10$ is 4.61 which is *less than* 7.0655, so again we conclude that at the 10% level there is strong *evidence* to reject the hypothesis that *leaf damage* is not associated with *plant type.* That is, there is evidence that different *plant type* do sustain varying degrees of *leaf damage.*