Statistics 2000, Section 001, Final (300 Points)

Wednesday, December 15, 2010

Part I: Text Answers

Your Name: ________________

Question 1: Statistical Inference (**50 Points**)

A lab scientist is interested in whether lab rats that grow up alone grow as large as lab rats that grow up with other rats around them to play with. He randomly selects ten young rats with approximately the same age and size. Five of these will spend the next 4 months by themselves and the other five rats will each have three other rats to play with during that same time. After 4 months, the scientist measures the abdomen circumference of all the rats (in mm). The results are shown below:

Alone group (#1)	110	123	113	103	120
Play group (#2)	119	125	131	128	136

The sample means and sample standard deviations are: $\overline{x}_1 = 113.8$ and $s_1 = 7.98$, $\overline{x}_2 = 127.8$ and $s_2 = 6.38$. It is reasonable to believe that rat weights in each group roughly follow a normal distribution. **Show your work!**

1. (**8 Points**) What are the hypotheses to test whether the rats grow up to be equally large or whether the playing rats grow up larger? Use the proper mathematical notation and symbols.

   $H_0: \mu_1 = \mu_2$ 4

   $H_a: \mu_1 < \mu_2$ 4

   −4 if swapped
   −2 if 2-sided
   −2 if wrong side

2. (**4 Points**) Briefly justify your choice of a one-sided or two-sided alternative in the previous part.

   one-sided (two-sample t-test):

   we have an assumption (before conducting the experiment) that rats that are playing with other rats grow up larger 4

−2 for each calculation error

3. (**8 Points**) Calculate the test statistic.

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{113.8 - 127.8}{\sqrt{\frac{7.98^2}{5} + \frac{6.38^2}{5}}} = \frac{-14}{\sqrt{20.88}} = \underline{-3.06}$$

8

4. (**6 Points**) What are the appropriate degrees of freedom for this test?

df = smaller of (5−1) and (5−1) = 4   6

[so we have a t-distribution with 4 df]

5. (**6 Points**) Determine the P-value.

$$2.999 < 3.06 < 3.747$$ 3

$$\downarrow \qquad \downarrow$$

$$0.02 > p > 0.01$$ 3

6. (**6 Points**) If you use the usual 5% significance level, should you reject the null hypothesis? **Yes** / **No** ? Circle your answer and briefly explain why/why not.

4

Yes, we reject the null hypotheses at the 5% significance level because the p-value is <u>less than 0.05</u>.   2

7. (**6 Points**) If you use the 1% significance level instead, should you reject the null hypothesis? **Yes** / **No** ? Circle your answer and briefly explain why/why not.

4

No, we fail to reject the null hypothesis at the 1% significance level because the p-value is <u>greater than 0.01</u>.   2

8. (**6 Points**) Give a short summary of your conclusion for the usual 5% significance level, i.e., how would you explain the result to a person who does not know much about statistics?

As we reject the null hypothesis at the 5% significance level, we have some reason to believe that the population mean μ_2 (for all playing rats) is 2 greater than the population mean μ_1 (for all alone rats).   6

Question 2: Regression (**40 Points**)

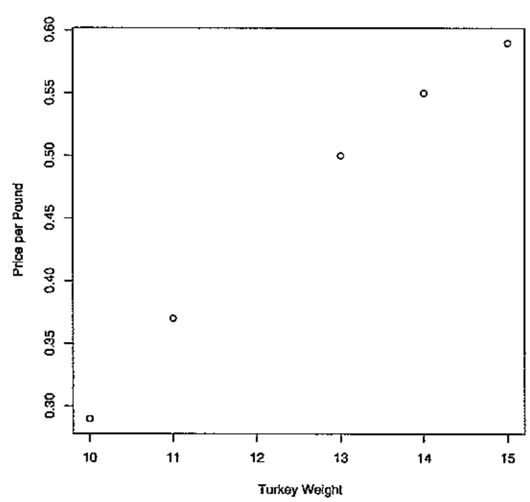
Before Thanksgiving, a supermarket ran the following ad regarding their turkey sale:

Frozen Hen Turkeys - $9 Off

10-15 lbs. average. Price per lb. after $9 discount.

	10 lbs.	11 lbs.	12 lbs.	13 lbs.	14 lbs.	15 lbs.
Meijer	$0.29	$0.37	$0.44	$0.50	$0.55	$0.59

Scatterplot of the data:



-2 for each calculation error

From a simple random sample of 5 customers who bought a frozen hen turkey, the data in the following table were obtained. A scatterplot of the data is shown on the previous page.

x Turkey Weight	y Price per Pound (in \$)	$\hat{y}$ Predicted Price per Pound (in \$)	e Residual
10	0.29	0.3027	−0.0127
11	0.37	0.3632	+0.0068
13	0.50	0.4842	+0.0158
14	0.55	0.5447	+0.0053
15	0.59	0.6052 5	−0.0152 5

The average turkey weight (in this sample of 5) was 12.6000 pounds with a standard deviation of 2.0736 pounds. The average price per pound was \$0.4600 with a standard deviation of \$0.1261. The correlation between weight and price per pound was 0.9944. **Work with 4 decimal digits (as above) and show your work!**

1. **(8 Points)** Calculate the regression equation that predicts price per pound from turkey weight.

slope: $b_1 = r \cdot \frac{s_y}{s_x} = 0.9944 \cdot \frac{0.1261}{2.0736} = 0.0605$ 3

intercept: $b_0 = \bar{y} - b_1 \bar{x} = 0.4600 - 0.0605 \cdot 12.6000 = -0.3023$ 3

regression equation: $\hat{y} = b_0 + b_1 x = -0.3023 + 0.0605 \cdot x$ 2

2. **(8 Points)** Predict the price per pound, using your regression equation, for turkey weights of 10, 11, 13, 14, and 15 pounds. Just show one sample calculation below and fill in all results in the table above.

3

$\hat{y}$ (for 10) $= -0.3023 + 0.0605 \cdot 10 = \underline{\underline{0.3027}}$

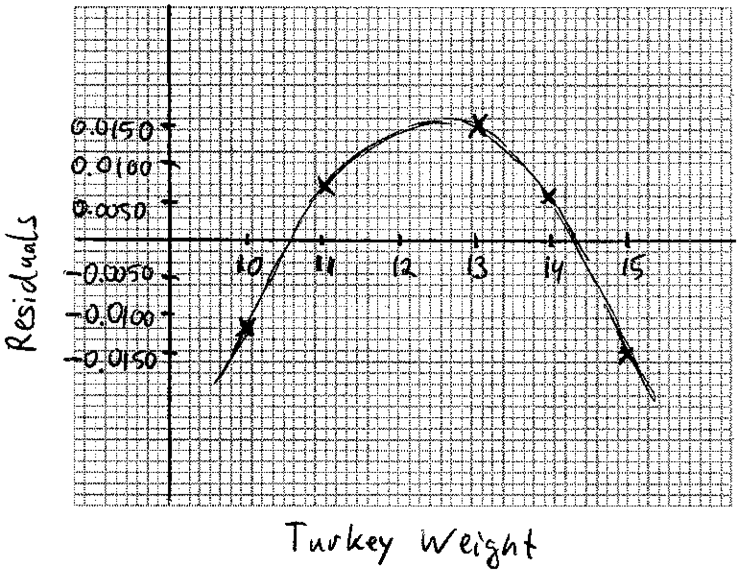
3. **(8 Points)** Calculate the 5 residuals. Just show one sample calculation below and fill in all results in the table above.

3

$\underset{(\text{for } 10)}{\text{residual}} = \underset{(\text{for } 10)}{y} - \underset{(\text{for } 10)}{\hat{y}} = 0.29 - 0.3027 = \underline{\underline{-0.0127}}$

4. (**8 Points**) Sketch a residual plot that shows turkey weight on the horizontal axis and the residuals on the vertical axis.

8



5. (**8 Points**) Based on your residual plot above, is your regression equation a good way to predict the price per pound based on the turkey weight? **Yes** / **No**? Circle your answer and provide some explanation. 4

No, the residual plot above shows a clear <u>curvature</u>. The regression equation is <u>not</u> a good way to predict the price per pound, based on turkey weight. 4

Note: Revisit the ad. This suggests

$$\text{Price per Pound} = \frac{\text{Original Price per Pound} \cdot \text{Weight} - 9.00}{\text{Weight}}$$

$$= \text{Original Price per Pound} - \frac{9.00}{\text{Weight}}$$

where

$$\text{Original Price per Pound} = \frac{0.29 \cdot 10 + 9.00}{10} = 1.19.$$

This is clearly not a linear relationship with respect to Weight; so our regression model is totally off!

Grading: 0 incorrect result (no calculation, no P-notation)
+1 some P-notation
+1 some calculations
+2 some P-notation & some calculations
+4 correct result, no calculation, no P-notation
+7 correct result + calculation
+8 correct result + calculation + P-notation

Question 3: Probability (**40 Points**)

For his next road trip, a student places the following 10 CDs into the glove compartment of his car:

Most direct solution below; alternative solution via tree diagram (see Midterm 2)!

- 4 *modern rock* CDs (Neon Trees, Black Keys, Civil Twilight, Phoenix),
- 3 *pop* CDs (Katy Perry, Lady Gaga, Ke$ha),
- 3 *American Idol* CDs (Lee DeWyze, Kris Allen, David Cook).

On his trip, the student blindly grabs a CD from the glove compartment, listens to it, and places it on the back seat when finished. Then he blindly grabs a second CD from the glove compartment. You should NOT comment on the musical taste of this student, but **answer each of the following questions separately. Show your work!** As a part of your answer, translate the everyday language into probability statements, using the proper notation, e.g., $P(A), P(A \text{ and } B), P(A \text{ or } B), P(A|B)$, etc., where A (and B) are the events of interest.

* Note that (2) & (5) ask similar questions; one for pop (2) and one for Idol (5). Both calculations are equivalent and yield exactly the same numerical result!

1. (**8 Points**) What is the chance that he listens to the *Katy Perry* CD, followed by the *Lady Gaga* CD? The chance is 1.11 %

$P(\text{1st Katy Perry and 2nd Lady Gaga}) = P(\text{1st Katy Perry}) \cdot P(\text{2nd Lady Gaga} \mid \text{1st Katy Perry})$

$= \frac{1}{10} \cdot \frac{1}{9} = \frac{1}{90} = 0.0111 = \underline{\underline{1.11\%}}$

2. (*) (**8 Points**) What is the chance that he will listen to a *pop* CD as one of his two selections? The chance is 53.33 %

$P(\text{1st pop or 2nd pop}) = P(\text{1st pop}) + P(\text{2nd pop}) - P(\text{1st pop and 2nd pop})$

$= \frac{3}{10} + \frac{3}{10} - \frac{3}{10} \cdot \frac{2}{9} = \frac{27}{90} + \frac{27}{90} - \frac{6}{9} = \frac{48}{90} = 0.5333 = \underline{\underline{53.33\%}}$

3. (**8 Points**) What is the chance that he will listen to **none** of the *modern rock* CDs? The chance is 33.33 %

$P(\text{1st not rock and 2nd not rock}) = P(\text{1st not rock}) \cdot P(\text{2nd not rock} \mid \text{1st not rock})$

$= \frac{6}{10} \cdot \frac{5}{9} = \frac{30}{90} = 0.3333 = \underline{\underline{33.33\%}}$

4. (**8 Points**) What is the chance that the **second** CD will be a *pop* CD or an *American Idol* CD? The chance is 60 %

$P(\text{2nd pop or 2nd Idol}) = P(\text{2nd pop}) + P(\text{2nd Idol})$

$= \frac{3}{10} + \frac{3}{10} = \frac{6}{10} = 0.6 = \underline{\underline{60\%}}$

5. (*) (**8 Points**) What is the chance that he will listen to **at least one** of the *American Idol* CDs? The chance is 53.33 %

$P(\text{at least one Idol}) = 1 - P(\text{both not Idol})$

$= 1 - P(\text{1st not Idol}) \cdot {}^{6}P(\text{2nd not Idol} \mid \text{1st not Idol})$

$= 1 - \frac{7}{10} \cdot \frac{6}{9} = \frac{90}{90} - \frac{42}{90} = \frac{48}{90} = 0.5333 = \underline{\underline{53.33\%}}$

−2 for each calculation error
−2 if no final result

Question 4: Probability and Distributions (**50 Points**)

A few months later, without changing the CDs, the student takes off for a Spring Break trip to Florida (i.e., he still has 4 *modern rock*, 3 *pop*, and 3 *American Idol* CDs). He plans to listen to 60 CDs on this trip. Therefore, he changes his CD replacement strategy. Instead of placing a finished CD on the back seat, he replaces the CD into the glove compartment, shuffles all CDs in the compartment, and blindly grabs the next CD. He follows this strategy for each CD he listens to. **Show your work!**

1. (**8 Points**) What is the chance that he listens to the *Katy Perry* CD, followed by the *Lady Gaga* CD as his first 2 selections? The chance is ____1___%

$$P(\text{1st Katy Perry and 2nd Lady Gaga}) = P(\text{1st Katy Perry}) \cdot P(\text{2nd Lady Gaga})$$
$$= \frac{1}{10} \cdot \frac{1}{10} = \frac{1}{100} = 0.01 = \underline{1\%}$$
[points → Question 3]

2. (**6 Points**) For the remaining parts of this question, the focus will be on the *modern rock* CDs. We will call listening to a *modern rock* CD a "success". Using the proper statistical notation, write down the underlying statistical distribution for a random variable X that counts the number of *modern rock* CDs the student listens to on this road trip.

$$X \sim Bin(\underset{n}{60}, \underset{p}{4/10})$$
6

3. (**6 Points**) What is the expected number of *modern rock* CDs he expects to listen to on his trip to Florida.

$$\mu = n \cdot p$$
$$= 60 \cdot \frac{4}{10} = \underline{24}$$
(modern rock CDs)
6

4. (**6 Points**) What is the corresponding standard deviation?

$$\sigma = \sqrt{n \cdot p(1-p)}$$
$$= \sqrt{60 \cdot \frac{4}{10} \cdot \frac{6}{10}}$$
$$= \sqrt{14.4}$$
$$= \underline{3.79}$$
(modern rock CDs)
6

5. (**6 Points**) Can we use a Normal approximation to determine the probability that 20 to 30 of the CDs he will listen to are *modern rock* CDs? **Yes** / **No.** Circle the correct answer and explain why you chose this answer.

2

$$n \cdot p = 60 \cdot \frac{4}{10} = 24 \geq 10 \checkmark \quad 2$$

$$n \cdot (1-p) = 60 \cdot \frac{6}{10} = 36 \geq 10 \checkmark \quad 2$$

both conditions are fulfilled to use a Normal approximation

6. (**10 Points**) Independently from your previous answer (i.e., no matter whether you answered yes or no), use a Normal approximation to determine the probability that 20 to 30 of the CDs he will listen to are *modern rock* CDs. For the purpose of this question part, we believe that the necessary requirements hold.

$$X \text{ approx } \mathcal{N}\left(n \cdot p, \sqrt{n \cdot p (1-p)}\right) = \mathcal{N}(24, 3{,}79) \quad 2$$

$$P(20 \leq X \leq 30) \; 2 = P\left(\frac{20-24}{3.79} \leq \frac{X-24}{3{,}79} \leq \frac{30-24}{3{,}79}\right)$$

$$= P(-1.06 \leq Z \leq 1.58) \quad 2$$

$$= 0.9429 \; 2 - 0.1446 \; 2$$

$$= \underline{\underline{0.7983}}$$

7. (**8 Points**) Now, do the same calculation as in the previous part, but use a Normal approximation with **continuity correction**.

$$P(19.5 \leq X \leq 30.5) \; 2 = P\left(\frac{19.5-24}{3.79} \leq \frac{X-24}{3.79} \leq \frac{30.5-24}{3.79}\right)$$

$$= P(-1.19 \leq Z \leq 1.72) \quad 2$$

$$= 0.9573 \; 2 - 0.1170 \; 2$$

$$= \underline{\underline{0.8403}}$$

[i.e., a difference of about 4% between the 2 calculations!]

Statistics 2000, Section 001, Final (300 Points)

Wednesday, December 15, 2010

Part II: Multiple Choice Questions

Your Name: ____________________

Question 5: Multiple Choice Questions (**120 Points**)

Mark your answer for each multiple choice question in the table below. There is only one correct answer for each question. Each correct answer is worth 4 points.

Question	(a)	(b)	(c)	(d)	Question	(a)	(b)	(c)	(d)
1	○	○	○	●	16	○	○	○	●
2	●	○	○	○	17	○	○	●	○
3	○	●	○	○	18	○	○	●	○
4	○	○	○	○	19	○	○	○	●
5	○	●	○	○	20	○	●	○	○
6	○	○	○	●	21	○	○	○	●
7	○	○	●	○	22	○	○	●	○
8	○	○	●	○	23	○	○	○	●
9	○	○	○	●	24	○	○	●	○
10	○	○	●	○	25	●	○	○	○
11	○	●	○	○	26	○	●	○	○
12	○	○	●	○	27	○	○	○	●
13	○	●	○	○	28	○	○	●	○
14	●	○	○	○	29	○	○	○	●
15	○	○	○	●	30	○	●	○	○

Stat 2000, Final, Question 5 — Solutions

1. (d) In a SRS of a fixed size n, each set of size n has exactly the same probability of being selected. Despite the impression that set (a) appears to be more likely, in fact all of these three sets of 5 subjects are equally likely.

2. (a) For a fair coin, it is $P(H) = P(T) = 1/2$. Thus, it is

$$P(H, H, H, H) = (1/2)^4 = 1/16 = 0.0625 = 6.25\%$$

and

$$P(H, T, T, H, T) = P(T, T, T, T, T) = (1/2)^5 = 1/32 = 0.03125 = 3.125\%.$$

So H, H, H, H is more likely than the other two outcome sequences.

3. (b) A stratified random sample because we first divide the population into two strata (men and women).

4. (??) This is volunteer sampling as the *USA Today* readers decided themselves to participate by calling-in. [4 points for everyone as the actual question got mislabeled as an answer here.]

5. (b) An observational study, not an experiment, as the tomatoes were simply allowed to grow in the same part of the garden in each year. There was no experimentation being done by applying different types or amounts of fertilizer, nor using different watering techniques, etc.

6. (d) There was no blinding as students could see which computer they were working on. The current design of the laptop served as control, students were randomly assigned to the computer type, $75/3 = 25$ students were assigned to each computer type, thus we have repetition.

7. (c) Common response. As we are looking at the number of people who failed to complete high school and the number of infant deaths, states with large populations naturally will have higher numbers for each of the variables, even if there is no causal relationship.

8. (c) Confounding. Both variables "length of time contacts are worn each day" and "length of time spent at the office per day" have an effect on irritation in the eyes, but it is indistinguishable which of the two variables contributes how much to this observed outcome.

9. (d) Correlation measures the strength of the linear relationship between two quantitative variables. Correlation clearly is meaningless if the relationship is not linear.

10. (c) It is slope $= r \cdot \frac{SD_y}{SD_x}$. Thus, $r =$ slope $\cdot \frac{SD_x}{SD_y} = 2.12 \cdot \frac{5970}{15300} = 0.827$.

11. (b) The formula for the expected cell count of a given cell is

$$\text{expected cell count} = \frac{\text{row total} \times \text{column total}}{\text{table total}}.$$

In this case, row total = 100, column total = 20, and table total = 150. Thus,

$$\text{expected cell count} = \frac{100 \cdot 20}{150} = 13.33.$$

12. (c) The degrees of freedom for a χ^2 test are $(r-1)(c-1)$ which is 12 here.

13. (b) The data can be considered as either two proportions or a 2×2 table. The two tests are equivalent.

14. (a) The degrees of freedom is $(r-1)(c-1) = 3$. The value 21.236 exceeds the largest table entry of 12.92 (for 3 df), so the P-value must be below 0.0005.

15. (d) When considering the size of a table, we speak of rows by columns. This table has three rows and five columns, ignoring the column and row of marginal totals and also ignoring the column and row with the identifiers of the classes.

16. (d) The degrees of freedom are $I-1$ and $N-I$ where there are I groups and a total of N observations. We have $I = 3$ groups with 10 observations per group, so $N = 30$.

17. (c) The null hypothesis for the test is $H_0 : \mu_1 = \mu_2 = \ldots = \mu_I$. With a p-value of 0.001 we reject H_0. The correct alternate hypothesis is that not all the group means are the same, which implies that some of the group means differ from the others.

18. (c) No — because the largest standard deviation is more than twice the smallest.

19. (d) Recall that a level C confidence interval for the mean μ of a normal population is $\overline{x} \pm t^* \frac{s}{\sqrt{n}}$ where n is the sample size, $\overline{x}$ is the sample mean, s is the sample standard deviation, and t^* is the upper $(1-C)/2$ critical value of the t distribution with $n-1$ degrees of freedom. Here, $n = 25$, $\overline{x} = 69.72$, $s = 4.15$, and $t^* = 1.711$ (the upper 0.95 critical value of the $t(24)$ distribution). Our 90% confidence interval is therefore $69.72 \pm 1.711 \frac{4.15}{\sqrt{25}} = 69.72 \pm 1.42$.

20. (b) Husbands and their wives come naturally in pairs. We have two observations on each couple. Therefore, a paired samples t-test would be appropriate here.

21. (d) Calculate $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$.

22. (c) Calculate $\binom{6}{4} = \frac{6!}{4! \cdot (6-4)!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(4 \cdot 3 \cdot 2 \cdot 1) \cdot (2 \cdot 1)} = 15$.

23. (d) We need to calculate the proportion of students whose time to complete the exam is under 60 minutes. The z–score corresponding to 60 is $(60-45)/10 = 1.5$. The area to the left is 0.9332, or about 0.93, which corresponds to the proportion who complete the exam in under an hour.

24. (c) The z–score corresponding to the 92nd percentile is 1.41. We need to multiply the standard deviation by 1.41 and add it to the mean, which gives $65 + 1.41 \cdot 11 = 80.5$.

25. (a) X takes the value \$0 with probability 6/10 (the probability of a blue ball), the value \$2 with probability 2/10, and the value \$4 with probability 2/10. The mean of X is given by $\mu_X = \$0 \cdot 6/10 + \$2 \cdot 2/10 + \$4 \cdot 2/10 = \$1.20$.

26. (b) X, the number correct if you were guessing, has a $B(80, 0.25)$ distribution. The standard deviation of X is $\sqrt{n \cdot p \cdot (1-p)} = \sqrt{80 \cdot 0.25 \cdot 0.75} = \sqrt{15} = 3.87$.

27. (d) We cannot say the Central Limit Theorem applies in this case because a sample size of 5 is not "large", so we cannot determine the probability with the provided information.

28. (c) Calculate $-7 + (-5) + (-2) + 5 = -9$.

29. (d) Resort the data from smallest to largest:

    -7 -5 -2 3 5 7

    Then calculate $-5 + (-2) + 3 + 5 = 1$.

30. (b) We are summing up 6 times the same value (-2), so

$$-2 + (-2) + (-2) + (-2) + (-2) + (-2) = 6 \cdot (-2) = -12.$$