Statistics 2000, Section 001, Final (300 Points)

Wednesday, May 4, 2011

Part I: Text Answers

Your Name: ____________________

Question 1: Statistical Inference (**68 Points**)

Eight people volunteered to be part of an experiment. All 8 people were Caucasian, between the ages of 25 and 35, and were supplied with nice clothes. 4 of the people were male and 4 were female. The question of interest in this experiment was whether females receive faster service at restaurants than males. Each of the 4 male participants was randomly assigned to a restaurant, and each of the 4 females was randomly assigned to one of these **same** (!) 4 restaurants. One Friday night, all 8 people went out to eat, each one alone. The male and female assigned to the same restaurant would arrive within 5 minutes of each other, with the order determined by flipping a coin (male first or female first). Each person then ordered a similar drink and a similar meal. The time (in minutes) until the food arrived at the table was recorded. They are shown below:

Restaurant i	1	2	3	4
Male t_{m_i}	22	14	16	26
Female t_{f_i}	25	12	13	21
difference d_i	−3	2	3	5

It is reasonable to believe that the differences in serving times at the same restaurant roughly follow a normal distribution.

−4 if H_0, H_a swapped

−6 if not about differences

Show your work!

1. (**8 Points**) We want to determine whether women are served **faster** than men. State the two hypotheses. Be careful what goes into the null hypothesis and what goes into the alternative. Use the proper mathematical notation and symbols.

Let d_i be the difference between times for the male and female at restaurant i, i.e., $d_i = t_{m_i} - t_{f_i}$, $i = 1, 2, 3, 4$.

(4) $H_0: \mu_d = 0$ (i.e., there is no difference in the average serving times)

(4) $H_a: \mu_d > 0$ (i.e., average difference is > 0, i.e., average serving time for men is longer than average serving time for women)

−2 for each calculation error

(4)

2. **(6 Points)** Do we have to conduct a **matched pairs t–test** or a **2–sample t–test**? Circle the correct test and briefly justify your answer.

We are keeping all possible confounding factors such as day of week, arrival time, food ordered, appearance of customer, etc. constant; only gender differs. This is the ideal scenario for a matched pairs t-test. (2)

3. **(9 Points)** Calculate the proper mean(s) and standard deviation(s). No need to show your work for this part — just write down the final results.

differences (comp. 1): (2)

$$\bar{x}_d \left[= \frac{1}{4}\sum_{i=1}^{4} d_i = \frac{-3+2+3+5}{4} = \frac{7}{4}\right] = 1.75$$ (3)

$$s_d^2 \left[= \frac{1}{3}\sum_{i=1}^{4}(d_i-\bar{x}_d)^2 = \frac{1}{3}\left((-3-1.75)^2+(2-1.75)^2+(3-1.75)^2+(5-1.75)\right)^2 = \frac{34.75}{3}\right] = 11.58$$ (2)

$$s_d \left[= \sqrt{s_d^2} = \sqrt{11.58}\right] = 3.40$$ (2)

4. **(10 Points)** Calculate the test statistic.

(10)

$$t = \frac{\bar{x}_d - \mu_d}{s_d/\sqrt{n}} = \frac{1.75-0}{3.40/\sqrt{4}} = 1.029$$

5. **(6 Points)** What are the appropriate degrees of freedom for this test?

(6)

$$df = n-1 = 4-1 = 3$$

6. **(8 Points)** Determine the P–value.

$$0.978 < \underset{t}{1.029} < 1.250$$ (4)

$$0.20 > p > 0.15$$ (4)

[i.e., p-value between 0.15 and 0.20]

7. **(7 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)

p-value > 0.05 (i.e., 5%); (3)

therefore, do not reject the null hypothesis at the 5% sig. level

8. **(7 Points)** If you use the 20% significance level instead, should you reject the null hypothesis? **Yes** / **No ?** Circle your answer and briefly explain why/why not.

(4)

p-value < 0.20 (i.e., 20%); (3)

therefore, reject the null hypothesis at the 20% sig. level

[a 20% sig. level is unusual, but we are aware that the sample size is very small; one may want to get a basic idea whether something unusual is happening w.r.t. service times and then conduct a bigger experiment.]

9. **(7 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?

Based on this experiment, there is no statistically significant difference (at the 5% sig. level) in the average serving times for men and women. Both genders are served equally fast (at the population level). (7)

Question 2: Regression (**37 Points**)

Data were obtained from the A&W Web site for the Total Fat in grams and the Protein content in grams for various items on their menu. Some summary statistics are also provided:

Item	Total Fat (grams)	Protein (grams)
Kid's Cheeseburger	24	23
Kid's Hamburger	22	21
Original Bacon Cheeseburger	33	27
Original Bacon Double Cheeseburger	48	45
Original Double Cheeseburger	42	40
Papa Burger	42	41

	$x=$ Total Fat		$y=$ Protein
Mean	35.167		32.833
Standard Deviation	10.591		10.362
Correlation		r = 0.983	

−2 for each calculation error (or no final result)

−2 if x, y flipped

The scatterplot (not reproduced here) shows that there is indeed a linear relationship between the two variables.

Work with 3 decimal digits (as above) and show your work!

−1 if <3 decimal digits (each time)

1. **(16 Points)** Find the regression equation for predicting the Protein content from the Total Fat content.

slope: $b_1 = r \cdot \frac{s_y}{s_x} = 0.983 \cdot \frac{10.362}{10.591} = \underline{0.962}$ (6)

y-intercept: $b_0 = \bar{y} - b_1 \bar{x} = 32.833 - 0.962 \cdot 35.167 = \underline{-0.998}$ (6)

regression equation: $\boxed{\hat{y} = -0.998 + 0.962 \cdot x}$ (4)

2. **(5 Points)** Using your regression equation, estimate the Protein content for a menu item with a Total Fat content of 40 grams.

−3 for "implausible" result

$\hat{y}$ (for 40) $= -0.998 + 0.962 \cdot 40 = \underline{37.482}$ (5)

Note: Statistics is not about punching numbers in a calculator, but about interpreting data. A result that makes no sense, but is left without interpretation, results in a "−3" point deduction. When you get an implausible result (such as −20 or 150), you must comment that something went wrong!

3. **(8 Points)** The "Kid's Cheeseburger" contains 24 grams Total Fat and 23 grams Protein. Using your regression equation, calculate the corresponding residual (i.e., error). Be careful with the sign!

(3)

$\hat{y}$ (for 24) = $-0.998 + 0.962 \cdot 24 = 22.090$

residual (for 24) = observed y − predicted y

= 23 − 22.090 (5)

= 0.910

4. **(8 Points)** The recently introduced "Triple Meat Mega Fat Burger" (not yet listed on the menu above) contains 80 grams Total Fat. Is it safe to use your regression equation from 1. above to predict the Protein content for this burger? Yes / No. **Circle your answer and provide an explanation.**

(4)

Note that 80 grams of fat is more than $\frac{80 - 35.167}{10.591} = 4.233$ standard deviations above the mean.

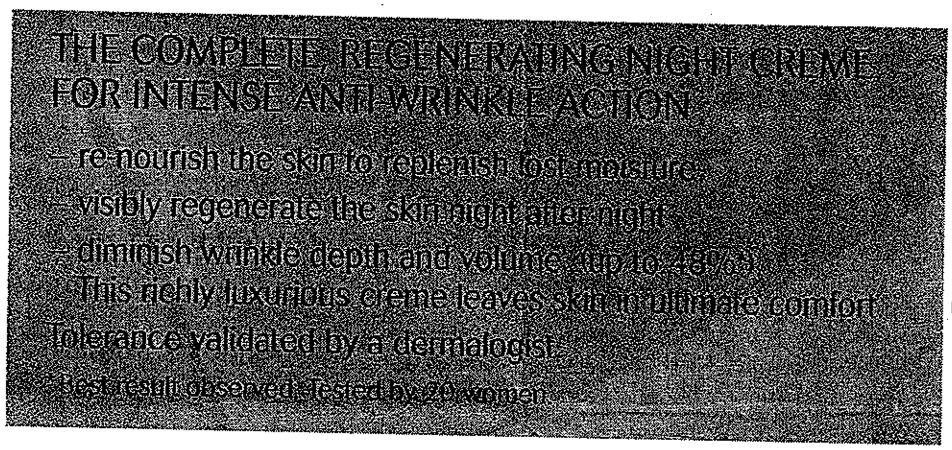
Also, 80 grams of fat is $\frac{80 - 48}{10.591} = 3.021$ standard deviations above the largest observed fat content for any of the other menu items.

Thus, 80 grams of fat represents a clear outlier. (4)

We have no reason to believe that the regression equation is still valid for such an extreme fat value. Using the regression equation would be extrapolation!

Question 3: General Statistical Concepts (**30 Points**)

Yves Rocher, a leading producer of cosmetics, includes the information below in the package of its brand new anti–wrinkle night creme:



(6) for each meaningful question

1. (**18 Points**) Assume that you are invited to a press conference where this new product is presented. You are allowed to ask **THREE** different **statistical** questions that apparently have not been answered on the package information above. You are not allowed to ask non–statistical questions such as the retail price or stores where this new product will be sold. Be precise in formulating your questions, i.e., use proper statistical terms in your questions! You can assume that the head of the Yves Rocher Statistics Department will be at the press conference and can answer your questions.

1) What is the mean (and median) reduction in wrinkle depth and volume?

2) What is the corresponding standard deviation?

3) How does the corresponding distribution look like? Is it skewed? If so, in which direction?

4) Are there any outliers? What are the values for these outliers?

5) The product was tested by 20 women. How were these women selected?

6) Was the product testing done via a (a) controlled experiment or via a (b) observational study?

7) If (a), which other products were part of the experiment? Was there a placebo? What was it?

8) If (b), how did you control for possible confounding factors, such as age, length of nightly sleep, etc.?

9) Did you conduct any statistical test? What were the hypotheses, p-value?...

... and much more one could ask...

⑥ for each meaningful reason / explanation

2. **(12 Points)** Actually, the two numbers from the package information above (i.e., 48% and 20 women) represent the only statistical information revealed by Yves Rocher for this new product. Can you speculate why none of your questions has been answered in the full package information? We can assume that a well–trained statistician such as the head of the Yves Rocher Statistics Department should have anticipated most of your questions before they were even asked... Provide **TWO** likely reasons why Yves Rocher may have chosen to not reveal additional statistical information. Briefly explain each reason in 1 (or at most 2) sentences.

1) The 48% reduction likely is a clear outlier that makes the new product look good. The mean (and median) reduction is much smaller.

2) Yves Rocher may have conducted an experiment. But the reduction in wrinkle depth and volume may not have been statistically significant when compared with a placebo. So, it may be better to not reveal any details...

3) The sample may be biased, e.g., women with very large wrinkle depths and volume, that are not representative for the entire population of targeted women.

etc. ...

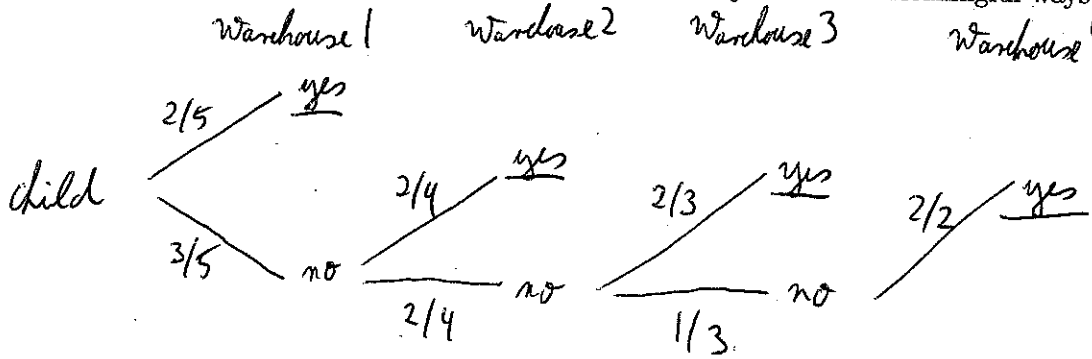
−6 for answers such as "customers don't understand much about statistics anyway" or "customers aren't interested in additional statistical data"

−2 for each calculation error (or no final result)

Question 4: Probability and Expectations (**45 Points**)

A child is looking for the preferred kind of pralines as a Mothers' Day gift for its mother. There are 5 warehouses in town. Unknown to the child, only 2 of these 5 warehouses carry this particular kind of pralines. The child visits warehouse after warehouse until it finds the preferred kind of pralines. The child **does not** go to the same warehouse twice! Once the child has found the preferred kind of pralines, it does not go to any additional warehouse.
Let X be the random variable that represents the number of warehouses the child has to visit until it finds the preferred kind of pralines. X can only take values 1, 2, 3, and 4.
Show your work!

1. (**8 Points**) Sketch a (probability) tree diagram that summarizes the information from the text above. Label the paths and nodes of your tree in meaningful ways.



(8)

2. (**12 Points**) Use your probability tree above to read off (or calculate) the following probabilities:

$P(X = 1) = \frac{2}{5} = 4/10 = 40\%$ (3)

$P(X = 2) = \frac{3}{5} \cdot \frac{2}{4} = 3/10 = 30\%$ (3)

$P(X = 3) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = 2/10 = 20\%$ (3)

$P(X = 4) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} \cdot \frac{2}{2} = 1/10 = 10\%$ (3)

3. **(4 Points)** We have discussed several probability rules in Chapter 4. Verify that the two most basic rules hold in your probability assignments above. Hint: Rule 1 is related to the probability of any event A and Rule 2 is related to the probability of the sample space S.

Rules (from p. 246):

Rule #1: $0 \le P(A) \le 1$ for any event A: yes ✓ (2)

Rule #2: $P(S) = 1$; yes, as $P(S) = \sum_{i=1}^{4} P(X=i) = \frac{4}{10} + \frac{3}{10} + \frac{2}{10} + \frac{1}{10} = 1$ ✓ (2)

4. **(8 Points)** What is the expected (average) number of warehouses the child has to visit until it finds the preferred kind of pralines?

$$\mu_X = \sum_{i=1}^{4} i \cdot P(X=i) = 1 \cdot \frac{4}{10} + 2 \cdot \frac{3}{10} + 3 \cdot \frac{2}{10} + 4 \cdot \frac{1}{10}$$

$$= \frac{20}{10} = 2$$

(8)

5. **(7 Points)** Given that the child **did not** find the preferred kind of pralines in warehouse 1, what is the probability that the child will find the preferred kind of pralines in warehouse 2?

$$P(X=2 \mid X \neq 1) = \frac{P(X=2 \text{ and } X \neq 1)}{P(X \neq 1)} = \frac{3/10}{3/5} = \frac{5}{10} = 50\%$$

(7)

[or directly from tree : $2/4 = 50\%$]

6. **(6 Points)** Does X have a $Bin(5, 2/5)$ distribution? Yes / **No.** **Circle your answer and provide an explanation.** (3)

A Binomial distribution requires the same (constant) probability of a success for each trial. Here, that probability is different for each trial ($\frac{2}{5}$, $\frac{2}{4}$, $\frac{2}{3}$, and $2/2$). (3)

The situation basically represents a scenario of drawing without replacement until the first "success" is encountered.

Statistics 2000, Section 001, Final (300 Points)

Wednesday, May 4, 2011

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. The first 25 questions have four choices, the last 5 questions have five choices. Each correct answer is worth 4 points.

Question	(a)	(b)	(c)	(d)	Question	(a)	(b)	(c)	(d)	(e)
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. (b) t–procedures are robust against nonnormality of the population except in the case of outliers or strong skewness. Guidelines are
   - Sample size less than 15: Use t–procedures if the data are close to normal. If the data are clearly nonnormal or if outliers are present, do not use t.
   - Sample size at least 15: t–procedures can be used except in the presence of outliers or strong skewness.
   - Large samples: The t–procedures can be used even for clearly skewed distributions when the sample is large, roughly $n = 40$ or 50.

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. (c) This is volunteer sampling as the *USA Today* readers decided themselves to participate by calling–in.

4. (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.

5. (c) When using t–procedures we are using two estimates based on the sample — both the sample mean and the sample standard deviation. This additional uncertainty is reflected in t multipliers being larger than z multipliers (see Table D). Therefore, t invervals are wider than z intervals based on the same data set.

6. (a) As degrees of freedom become larger, the difference between the t and z distributions becomes smaller. In fact, $t(\infty) = z$.

7. (c) The number of heads in $n = 4$ trials has a $B(4, 4/9)$ distribution. We must use the formula for computing binomial probabilities. We want to know the probability of exactly $k = 2$ heads. According to the formula, this probability is $P(X = 2) = \binom{4}{2}(4/9)^2(1 - 4/9)^2 = 0.366$.

8. (c) If the students' scores are independent, then the sum follows a normal distribution with mean $500 + 500 = 1000$ (the sum of the means for each student) and variance $100^2 + 100^2 = 20000$ (the sum of the variances for each student). Because $\sqrt{20000} = 141.42$, the sum of the scores is $N(1000, 141.42)$. The desired probability is $P(\text{sum} > 1200)$. The z–score of 1200 is $\frac{1200-1000}{141.42} = 1.41$, so the desired probability is $P(Z > 1.41) = 0.0793$.

9. (c) The row total for males is 200 and the column total for those who thought a large salary is important is 260. The expected cell count is $\frac{200 \cdot 260}{400} = 130$.
10. (a) The degrees of freedom for a χ^2 test are $df = (r-1)\cdot(c-1) = (2-1)\cdot(2-1) = 1$.
11. (a) The null hypothesis in general is that there is nothing unusual going on and all we see in the sample is variation by chance. Nothing unusual in this framework is "no association".
12. (d) Substituting 1600 for n in the formula for the standard deviation, $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, gives 0.01.
13. (b) The model predicts an average y value for any x, not for individuals.
14. (c) To reduce the variability of a sample statistic, increase the sample size, so the largest sample size will have the smallest variability.
15. (d) Bias refers to whether or not a sampling distribution has a mean equal to the parameter. Sample size has an effect on sampling variability but no effect on bias — even large samples may be biased.
16. (b) We have $I = 4$ groups with 15 observations per group, so we have $N = 60$ observations overall. The degrees of freedom are $I - 1 = 3$ and $N - I = 56$.
17. (a) The null hypothesis for the test is $H_0 : \mu_1 = \mu_2 = \ldots = \mu_I$. With a p-value of 0.19, we do not reject H_0. Therefore, we can assume that all the groups have the same mean.
18. (d) ANOVA requires (b), i.e., that the data are approximately normal distributed with no outliers **and** (c), i.e., that the largest sample standard deviation is no more than twice as big as the smallest sample standard deviation. Both are fulfilled: (b) is stated in the text and (c) is true because $4.5 < 2 \cdot 2.3 = 4.6$.
19. (b) The principles are: comparison, randomization, and repetition. See Section 3.1 (pp. 183–184) for more details.
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. (c) A voluntary response. Students must be willing to express their opinion.
22. (c) The 12 students interviewed make up the sample.
23. (b) Resort the data from smallest to largest:

    −10   −5   5   8   10

    Then determine $x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{5+1}{2}\right)} = x_{(3)} = 5$.
24. (a) Calculate $5 + (-5) + (-10) = -10$.

25. (a) Resort the data from smallest to largest:

    -10   -5   5   8   10

    Then calculate $-10 + (-5) + 5 = -10$.

26. (e) Let b = boy and g = girl. Then events $A = \{\text{the next two babies are boys}\} = \{(b,b)\}$ and $B = \{\text{at least one of the next two babies is a boy}\} = \{(b,g),(g,b),(b,b)\}$. A and B are not disjoint as they have (b,b) in common. They are not complements as $B^C = \{(g,g)\}$. They are not independent because when I know that A has occurred, that guarantees that B also has occurred, i.e., $P(B|A) = 1$. Finally, $P(A) = 1/4$ and $P(B) = 3/4$. So, none of the 4 previous options is correct.

27. (c) It is $\sigma_U^2 = \sigma_X^2 + \sigma_Y^2 = 6^2 + 4^2 = 52$ and $\sigma_U = \sqrt{\sigma_U^2} = \sqrt{52} = 7.2$.

28. (b) For a scatterplot, we don't have to fix which of the two variables is the explanatory variable and which is the response variable. This is only required when we calculate the regression equation. Moreover, we can use different symbols or colors to display additional categorical variables in the scatterplot, such as gender or natural habitat in some of the examples used in class. So, only (b) is correct.

29. (e) is correct. (a), (c), and (d) are nonsense as the graph does not show how many times people brush their teeth nor is there any information regarding time. (b) also makes no sense as the correlation/regression is based on averages (which is sometimes called an ecological correlation). We may not be able to say much about individuals. In fact, it may be the case that people who eat large amounts of sugar regularly brush their teeth (and therefore prevent caries) whereas people who eat little sugar don't brush their teeth regularly (and have more caries) — and suddenly the association within a particular country may be much weaker or it may even be negative. Clearly, there are multiple possible confounding factors. So, none of the 4 previous options is correct. A correct answer would be something like this: "The larger the average sugar consumption (per person per day) in a country, the larger the average number of decayed teeth per person in that country" or simply "There is a positive association between average sugar consumption (per person per day) in a country and the average number of decayed teeth per person in that country".

30. (d) We have to calculate

$$P(X=8) = \binom{10}{8} 0.6^8 \cdot 0.4^2 = \frac{10!}{8! \cdot 2!} 0.6^8 \cdot 0.4^2 = \frac{10 \cdot 9 \cdot 8!}{8! \cdot 1 \cdot 2} 0.6^8 \cdot 0.4^2 =$$

$$\frac{10 \cdot 9}{1 \cdot 2} 0.6^8 \cdot 0.4^2 = 45 \cdot 0.6^8 \cdot 0.4^2.$$