Statistics 2000, Section 001, Final (300 Points)

Wednesday, December 9, 2009

Your Name: _____

based on: Midterm 2, Question 1, Part I

Question 1: Miscellaneous Short Questions (**40 Points**)

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

Please answer the following short questions. **Show your work!**

1. (**8 Points**) Universal blood donors: People with type O–negative blood are universal donors. That is, any patient can receive a transfusion of O–negative blood. About 12% of a particular population have O–negative blood. If 6 people appear at random to give blood, what is the probability that at least 1 of them is a universal donor?

$P(\text{none are O-negative}) = (1-0.12)^6$ (4) $= 0.4644$

$\Rightarrow P(\text{at least 1 is O-negative}) = 1 - 0.4644$ (4) $= 0.5356 = \underline{53.56\%}$

based on: HW 9, Exercise 1, Q 5.8

2. (**8 Points**) A coin is slightly bent, and as a result the probability of a head is 0.55. Suppose that you toss the coin 5 times. Use the binomial formula to find the probability of 4 or more heads.

$X \sim Bin(5, 0.55)$

$P(X \geq 4) = P(X=4) + P(X=5)$

$= \binom{5}{4} 0.55^4 \cdot 0.45^{5-4}$ (4) $+ \binom{5}{5} 0.55^5 \cdot 0.45^{5-5}$ (4)

$= \frac{5!}{4!\,1!} 0.55^4 \cdot 0.45^1 + \frac{5!}{5!\,0!} 0.55^5$

$= 5 \cdot 0.55^4 \cdot 0.45 + 0.55^5$

$= 0.2059 + 0.0503 = 0.2562 = \underline{25.62\%}$

based on: Midterm 2, Question 1, Part 3

3. **(16 Points)** Portfolio analysis. The means, standard deviations, and correlations for the annual returns from three Fidelity mutual funds for the 10 years ending in February 2004 are listed below. The following subscripts are defined as

$W =$ annual return on 500 Index Fund
$X =$ annual return on Investment Grade Bond Fund
$Y =$ annual return on Diversified International Fund

These subscripts (W, X, and Y) show which random variable/pair of random variables a numerical characteristic, indicated below, refers to.

$$\mu_W = 10.12\%, \quad \sigma_W = 17.46\%$$
$$\mu_X = 6.46\%, \quad \sigma_X = 4.18\%$$
$$\mu_Y = 11.10\%, \quad \sigma_Y = 15.62\%$$

Correlations:

$$\rho_{WX} = -0.22, \quad \rho_{WY} = 0.56, \quad \rho_{XY} = -0.12$$

Assume that a portfolio contains 40% Investment Grade Bond Fund (X) and 60% 500 Index Fund (W) stocks. Calculate the mean and standard deviation of the returns for this portfolio.

• $R = 0.4X + 0.6W$

• $\mu_R = 0.4\mu_X + 0.6\mu_W$

$= 0.4 \cdot 6.46\% + 0.6 \cdot 10.12\%$ (6)

$= \underline{8.656\%}$

• $\sigma_R = \sqrt{(0.4\sigma_X)^2 + (0.6\sigma_W)^2 + 2\rho_{XW}(0.4\sigma_X)(0.6\sigma_W)}$

$= \sqrt{0.4^2 \cdot 4.18^2 + 0.6^2 \cdot 17.46^2 + 2 \cdot (-0.22) \cdot 0.4 \cdot 4.18 \cdot 0.6 \cdot 17.46}\ \%$ (10)

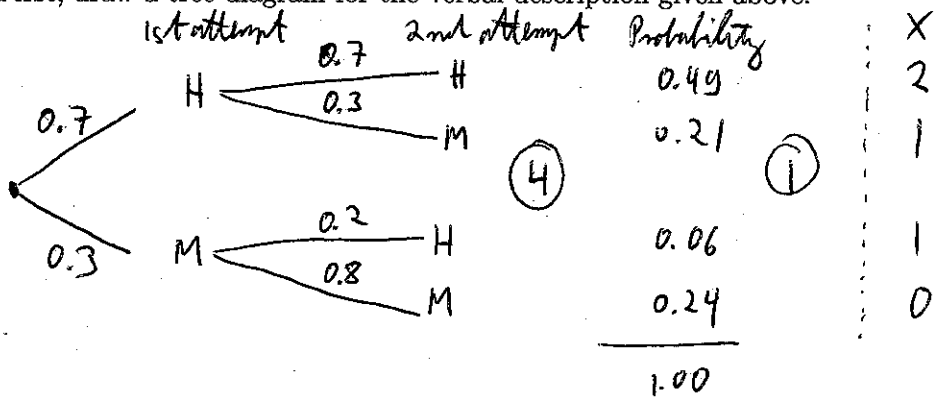
$= \sqrt{104.84}\ \%$

$= \underline{10.24\%}$

based on: Midterm, Question 2

4. **(8 Points)** A highly nervous basketball player has two free throw attempts. The player will make a hit with a probability of 0.7 in the first attempt. The outcome of the second attempt highly depends on the outcome of the first attempt: If the first attempt was a hit, the player will also make a hit with a probability of 0.7 in the second attempt. However, if the first attempt was a miss, the player will also miss in the second attempt with a probability of 0.8.

First, draw a tree diagram for the verbal description given above.



(4) (1)

Let X be a random variable that represents the number of baskets made out of the two shots. Then, use your tree diagram from above and fill in the following table. Don't forget to check whether the distribution in your table is legitimate (recall which conditions must hold for probabilities).

Value of X	0	1	2
Probability	0.24	0.21+0.06 = 0.27	0.49

(3)

(legitimate? - yes: sum = 1.0 ✓, each ≥ 0 ✓)

Question 2: Confidence Intervals and Hypothesis Testing I (**80 Points**)

A friend who hears that you are taking a statistics course asks for help with a chemistry lab report. She has made four independent measurements of the specific gravity of a compound. The results are:

3.82, 3.93, 3.67, 3.78

The lab manual says that repeated measurements will vary according to a Normal distribution with standard deviation $\sigma = 0.15$ (this standard deviation shows how precise the measurement process is). The mean μ of the distribution of measurements is the true specific gravity.

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

Show your work!

Part 1: The lab manual asks for a 95% confidence interval for the true specific gravity.

1. **(6 Points)** First, what would be a good estimator for the unknown population parameter? Indicate the name of this estimator (be specific!), the proper mathematical symbol, and the numerical values.

sample mean $\bar{x} = \frac{3.82+3.93+3.67+3.78}{4} = \frac{15.2}{4} = 3.8$ (2) (2) (2)

2. **(7 Points)** Calculate the margin of error for your 95% confidence interval.

$m = z^* \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{0.15}{\sqrt{4}} = 0.147$ (7)

3. **(6 Points)** Now, construct a 95% confidence interval for the true specific gravity.

95% CI: $\bar{x} \pm m = 3.8 \pm 0.147 = (3.653, 3.947)$ (6)

4. **(6 Points)** Can we be certain that the true population parameter falls in this interval? **Yes** / **No** ? Circle your answer. (6)

[no - we are just "95% confident" that the true mean μ falls into this interval]

5. **(6 Points)** What critical value from the Normal table would you use if you wanted 80% confidence instead of 95% confidence? You do not have to calculate this interval — just indicate the value from the Normal table that is needed.

from Table D, for $C = 80\%$: $z^* = 1.282$ (6)

6. **(6 Points)** Would the 80% confidence interval be **(i) wider** or **(ii) narrower** than your 95% confidence interval from 3. above? [Do NOT actually compute the 80% confidence interval — just circle the correct answer.] (6)

[a lower confidence level C results in a narrower interval]

7. **(7 Points)** How many measurements would be needed to estimate the true mean μ within $\pm$ 0.001 with 95% confidence, i.e., what would be the necessary sample size to obtain this margin of error?

$$n = \left(\frac{z^* \cdot \sigma}{m}\right)^2 = \left(\frac{1.96 \cdot 0.15}{0.001}\right)^2 = 294^2 = 86,436$$ (7)

Part 2: The lab manual also asks whether the data show convincingly that the true specific gravity is less than 3.9. To answer this question you decide to carry out a test of significance.

8. **(6 Points)** First, state the appropriate null and alternative hypotheses used to answer this question. Use the proper mathematical notation and symbols.

H_0: $\mu = 3.9$ (3) — 4 if wrong

H_a: $\mu < 3.9$ (3)

9. **(6 Points)** Calculate the test statistic.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{3.8 - 3.9}{0.15/\sqrt{4}} = -1.33$$ (6)

10. **(6 Points)** Determine the P–value.

from Table A:

$P(Z \leq -1.33) = 0.0918$ (6)

11. **(6 Points)** If you use the usual 5% significance level, should you reject the null hypothesis? **Yes** / (**No**) ? Circle your answer and explain why/why not. (4)

No, we fail to reject H_0 at the 5% significance level because the P-value is 0.0918 which is greater than 0.05. (2)

12. **(6 Points)** If you use the 10% significance level instead, should you reject the null hypothesis? (**Yes**) / **No** ? Circle your answer and explain why/why not. (4)

Yes, we reject H_0 at the 10% significance level because the P-value is 0.0918 which is less than 0.10. (2)

13. **(6 Points)** State your conclusion in terms of the context of the problem for the usual 5% significance level, i.e., how would you explain the result to your friend?

As we do not reject the null hypothesis at the 5% significance level, we have no reason to mistrust the lab manual. The true mean μ could indeed be 3.9. (6)

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

Question 3: Hypothesis Testing II (**60 Points**)

I am a little skeptical of the claim that the average weight of a particular type of tomato is 5.0 ounces. I think it might be somewhat greater than 5.0 ounces. I select 3 tomatoes at random and record their weights:

6.0, 4.8, 7.2

Then, I create a normal quantile plot. This plot only shows some minor wiggles without significant curvature or outliers, i.e., we can assume that the weights follow a normal distribution. Our main question now is: Is there evidence based on the significance level $\alpha = 0.05$ that the average weight of this type of tomato is greater than 5.0 ounces?

Show your work!

1. **(10 Points)** State which test of significance are you going to use (e.g., z, t, chi–square, one–sided or two–sided, one–sample or two–sample, etc.). Clearly indicate the crucial assumption(s) allowing you to use the test you chose.

one-sided (1), one-sample (1) t-test (4), because:
- the population standard deviation σ is unknown (2)
& - the data follows a Normal distribution (2)

2. **(8 Points)** Clearly state the null and alternative hypotheses. Use the proper mathematical notation and symbols.

-6 if swapped

H_0: $\mu = 5.0$ (4)

H_a: $\mu > 5.0$ (4)

3. **(6 Points)** Calculate the sample mean $\overline{x}$ and the sample standard deviation s. You can use the statistical mode of your calculator or you can do this by hand. No need to write down any formulas or intermediate results for this step. Just indicate the final results:

$$\overline{x} = \frac{6.0 + 4.8 + 7.2}{3} = \frac{18}{3} = 6.0 \quad (3)$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{3}(x_i - \overline{x})^2} = \sqrt{\frac{1}{2}\left[(6.0-6.0)^2 + (4.8-6.0)^2 + (7.2-6.0)^2\right]}$$

$$= \sqrt{\frac{1}{2}\left[0^2 + (-1.2)^2 + (1.2)^2\right]} = \sqrt{1.44} = 1.2 \quad (3)$$

4. **(8 Points)** Calculate the appropriate test statistic.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{6.0 - 5.0}{1.2/\sqrt{3}} = 1.4434$$ (8)

5. **(8 Points)** Determine the P–value.

$df = 3 - 1 = 2$ (3)

from Table D: $1.386 < t = 1.4434 < 1.886$ (2)

thus $0.15 > p > 0.10$ (3)

6. **(6 Points)** If you use the usual 5% significance level, should you reject the null hypothesis? Yes / **No** ? Circle your answer and explain why/why not. (4)

No, we fail to reject H_0 at the 5% significance level because the P-value is between 0.10 and 0.15 which is greater than 0.05 (2)

7. **(6 Points)** State your conclusion.

As we do not reject the null hypothesis at the 5% significance level, we have no reason to believe that the true population mean weight μ is greater than 5.0, i.e., we can assume that the true population mean is indeed 5.0. [However, we may want to confirm this result with a larger sample if possible ...] (6)

8. **(8 Points)** If it was believed that the weights of tomatoes do not follow the normal curve, would your significance test still be valid? Yes / **No** ? Circle your answer and explain briefly. (6)

Our sample size ($n = 3$) is less than 15 and if the data do not follow the Normal curve for such a small sample size, we should not use a t-test! (2)

Stat 2000, Final, Question 4 — Solutions

1. (b) 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 $(60-62)/8 = -0.25$. The area to the left is 0.4013, or about 0.40, which corresponds to the proportion who complete the exam in under an hour.

2. (d) The z–score corresponding to the 90th percentile is 1.28. We need to multiply the standard deviation by 1.28 and add it to the mean, which gives $62+1.28\cdot 11 = 76.08$.

3. (c) The median is the average of the third and fourth smallest observations. The third smallest must be at least 1.3. It will be 1.3 when $x \leq 1.3$, or else it will be larger. It can't possibly be smaller than 1.3. Similarly, the fourth smallest observation cannot exceed 2.2. So the median must be between these numbers. (In fact, you can show that the median must actually be between 1.6 and 2.05, but this requires a very careful argument — if you really want to test your understanding, try to see why this is true.)

4. (c) With most regressions, we could not say this because of extrapolation, but we were told the data included people who have had from 0 to 20 years of experience, so this is not extrapolation.

5. (a) Calculate $y = -30 + 60 \cdot 18 = 1050$.

6. (b) There were 6 students who earned A's who also studied more than 10 hours per week and 20 students who said they studied more than 10 hours per week. $6/20 = 0.30$.

7. (b) The model predicts an average y value for any x, not for individuals.

8. (b) The sampling distribution of $\hat{p}$ is centered about p, the chance of a die showing a 1 or a 2. The chance of any face of the die appearing is 1/6, so the chance of showing a 1 or 2 is $1/6 + 1/6 = 1/3$.

9. (c) The IQR measures the spread of the middle half (the distance between Q1 and Q3) so any outliers will not affect this measure.

10. (a) As long as the population is much larger than the sample (say, at least 10 times as large), the spread of the sampling distribution for a sample of fixed size n is approximately the same for any population size. Here $n = 1000$ and the populations of both the U.S. and Canada are at least 10 times this large. So the sampling variability associated with the proportion satisfied among the 1000 sampled is about the same.

11. (b) X takes the value \$0 with probability 6/10 (the probability of a blue ball), the value \$2 with probability 3/10, and the value \$4 with probability 1/10. The mean of X is given by $\mu_X = \$0 \cdot 6/10 + \$2 \cdot 3/10 + \$4 \cdot 1/10 = \$1$.

12. (c) The degrees of freedom for a χ^2 test are $(r-1)\cdot(c-1) = (4-1)\cdot(5-1) = 12$.

13. (b) Here, 73% is the percentage of all registered voters in the district, so it is a parameter. 68% was obtained from the sample of 500 voters, so it is a statistic.

14. (c) The rule for the variance of the difference of two independent random variables applies here, since the oranges are a SRS of size two from a large population. Recall the rule is $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$. You are given the standard deviation of the weight, and you must square it to get the variance. So $\sigma^2_X = 1.2^2 = 1.44$ and similarly $\sigma^2_Y = 1.2^2 = 1.44$. Thus $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y = 1.44 + 1.44 = 2.88$. The problem asks for the standard deviation of the difference which is the square root of the variance. The answer is then $\sigma_{X-Y} = \sqrt{2.88} = 1.70$ (which is greater than the original standard deviation of 1.20).

15. (d) Probability rule 5 states that two events are independent if $P(A \text{ and } B) = P(A) \cdot P(B) = 0.8 \cdot 0.7 = 0.56$, not 0.6, so the events are not independent.

16. (c) X, the number correct if you were guessing, has a B(100, 0.25) distribution. The standard deviation of X is $\sqrt{n \cdot p \cdot (1-p)} = \sqrt{100 \cdot 0.25 \cdot 0.75} = \sqrt{18.75} = 4.33$.

17. (c) X, the number correct, has a B(20, 0.25) distribution. The mean of X is $\mu_X = 5$. The student's score on the exam is $5 \cdot X$, so the student's mean score on the exam should be $5 \cdot \mu_X = 25$.

18. (d) The distribution is approximately normal with mean 0.2. The variance σ^2 is $p \cdot (1-p)/n = 0.2 \cdot 0.8/100 = 0.0016$. So, the standard deviation σ is $\sqrt{0.0016} = 0.04$.

19. (d) We cannot say the Central Limit Theorem applies in this case because a sample size of 10 is not "large", so we cannot determine the probability with the provided information.

20. (a) For a sample with 10 pieces of data, there are $10 - 1 = 9$ degrees of freedom. Find the correct t^* from Table D from the column with 0.025 in the tail.

21. (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).

22. (c) t procedures are robust against non–normality 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.

23. (a) The sample mean is the center of the confidence interval.

24. (b) The z test statistic has value $z = (\overline{x} - \mu_0)/(\sigma/\sqrt{n})$. For the data given we compute $z = (5.01 - 5.00)/(0.02/\sqrt{16}) = 0.01/(0.02/4) = 2.00$. The P–value is then the probability that a standard normal random variable would exceed 2.00 in absolute value, i.e., it is the probability that a standard normal random variable is either greater than 2.00 or less than -2.00. This probability is equal to twice the probability of being greater than 2.00. From Table A in the text, the probability of exceeding 2.00 is 0.0228. Doubling gives the P–value, i.e., 0.0456.

25. (c) For the P–value to be meaningful, we would need to know that the students in the state are a simple random sample (or can be considered a simple random sample) of the population of all students who took the SAT exam. Since these students were not selected by simple random sampling, we have no basis for regarding them as a simple random sample. In addition, comparing the average score in the state to the national average is not the correct comparison. To see if scores have risen, we would need to compare the mean score to scores in the state before the curriculum was introduced. Finally, even if we could regard the students as a simple random sample and even if we made the appropriate comparison, this is an observational study, not an experiment. We cannot safely conclude that the new curriculum caused scores to increase from such a study.

26. (d) Confidence intervals such as we have studied require the data be a simple random sample. Since he used the computer system to get all the GPAs, he has really done a census of the population, and so the mean calculated from his data is exactly the population mean μ (and not the sample mean $\overline{x}$).

27. (d) The confidence level states the probability that the method will give a correct answer in repeated use. In other words, if you use a 90% confidence interval often, in the long run 90% of your intervals will contain the true parameter value. That is the interpretation of the confidence level. So, none of the provided answers is correct.

28. (b) If X and Y are any two random variables, then $\mu_{X+Y} = \mu_X + \mu_Y$. If X is the time to complete the mathematics homework and Y is the time to complete the language arts homework, then the mean time to complete the entire homework assignment is $\mu_{X+Y} = \mu_X + \mu_Y = 30 + 40 = 70$ minutes. The correlation does not change the mean of $X + Y$, although it does affect its standard deviation.

29. (b) The correct degrees of freedom are $(r-1)\cdot(c-1) = (2-1)\cdot(3-1) = 2$. From Table F we have

p	.025	.02
Chi*	7.38	7.82

Because the value of the χ^2 statistic falls between the tabled critical values of 7.38 and 7.82, the P–value must lie between 0.02 and 0.025, which are included in the listed wider interval of 0.01 and 0.05.

30. (b) The expected cell count is (row total) · (column total)/(grand total). The row total for those who said they studied between 5 and 10 hours per week is 23, there are a total of 13 B's (the column total), and the grand total is 64. So, the result is $23 \cdot 13/64 = 4.672$.