MATH 316 (RUSIN) TEST 2, Mar 30 2012. **SOME ANSWERS**

1. Four identical coins are tossed, one after another.

1a. What is the probability that they will all come up "heads"?

1b. What is the probability that at there will be exactly three "heads" and one "tail"?

1c. Your answer in 1b should be greater than your answer in 1a. So does this mean that after three "heads" have already been tossed, you are more likely to toss "tails" next?

ANSWER: For (1a) it's $(1/2)^4 = (1/16)$ because the four coin tosses are independent.

For (1b) it's $4/16 = 1/4$ because there are 16 equally-likely outcomes (HHHH, HHHT, HHTH, HHTT, etc.) in the sample space, and exactly four of these (namely HHHT, HHTH, HTHH, and THHH) form the event that we are interested in.

For (1c) the answer is "no": the coin does not "know" that it is "behind" in the production of tails and so it doesn't "try to even things out". Each coin toss is independent of the previous ones, so that $Pr(H|HHH) = Pr(H) = (1/2)$. The reason we got a higher value in (1b) than in (1a) is because there are other ways to get a single tail other than HHHT; but none of those can occur after we have already observed that the first three tosses are H.

2. Pat is applying to several medical schools and has consulted the schools to find out what the chances of being admitted are. Pat discovers Johns Hopkins medical school accepts 50% of its applicants, and Harvard University medical school accepts 40% of its applicants.

2a. Assume that the acceptances into the two schools are independent events. What is the probability that Pat will be accepted into at least one of these two schools?

2b. In reality, Harvard and Hopkins are usually interested in the same students — a student accepted at one of them has a higher-than-average chance of being accepted at the other. Does this fact increase or decrease Pat's chances of being accepted by at least one of the two schools? (You may wish to illustrate with a numerical example.)

ANSWER: (2a) It's easier to find the probability of NOT getting accepted into either school, i.e. to be denied by JH *and* to be denied by HU; since those two events have been given as independent, the probability that both happen is the product of the individual probabilities: $(1 - .5) \times (1 - .4) = (.5)(.6) = .3$ so the probability of getting accepted by at least one of the two is $1 - .3 = .7$.

For (2b), let's write "JH" for the event that Pat is accepted by Johns Hopkins, and similarly for "HU". The information given is that $Pr(JH|HU) > Pr(JH)$. Then you might compute things as

$$\begin{aligned} Pr(JH \text{ } or \text{ } HU) &= Pr(JH) + Pr(HU) - Pr(JH \text{ } and \text{ } HU) \\ &= Pr(JH) + Pr(HU) - Pr(HU) \cdot Pr(JH|HU) \end{aligned}$$

Making that last factor larger makes the last product larger and thus makes the whole quantity smaller: Pat's chances are actually *lower* than 70% of being accepted at one of the schools.

3. You will be eligible to collect Social Security benefits on your 67th birthday; if you turn 18 today, that's exactly 49 years in the future. To invest for your retirement, you might buy a broad variety of U.S. stocks because there is an annual percentage gain x in the value of such an investment; historically, x has acted like a random value with a mean of about $\mu = 9$ percent per year and a standard deviation of about $\sigma = 21$ percent per year.

Assume that the values of x from year to year are uncorrelated, so that the net effect of investing for the next 49 years is the same as picking 49 values of x randomly. Then what is the probability that your stocks will lose value over that much time?

ANSWER: The standard deviation for sets of 49 x's is $\sigma/\sqrt{n} = 21/7 = 3$ percent per year, while the mean for sets of 49 is still $\mu = 9$. Thus only if our set-of-49 is at least 3 standard deviations below the mean would our $\bar{x}$ be smaller than $0 = 9 - 3 \cdot 3$. According to Table A, an event that's 3 standard deviations below the mean would occur only about .001 (i.e. 1/10 of 1 percent) of the time.

It's worth repeating that the only time you can use Table A is for variables which follow a normal distribution. We are not told that the variable x is normal! But the Central Limit Theorem tells us that the distribution of $\bar{x}$ when we take many sets-of-49 is much more like a normal distribution, so the use of Table A is justified. (However, economists do warn us that taking a single 49-year stretch is not the same as taking 49 years at random, so there is more uncertainty than our calculations suggest.)

4. In Chapter 14 we discussed statistical inference and encountered a lot of mumbo-jumbo like this: "The average body-mass index (BMI) of young Americans is 26.8 ± 1.9 with 95% confidence." Use the terms of Chapter 14 to explain what this means. (For example, why is there any uncertainty at all — can't the researchers do simple arithmetic and average all the BMI values? What does it mean to measure "confidence"? Can't they just replace the 1.9 by 2.0 and say what the average BMI is, with 100% confidence?)

ANSWER: I want to see that you understand the researchers had to contend with a *sample*, not the entire population, and that this introduces the possibility that the sample is not representative. Specifically, the numbers given tell us that 26.8 was the average BMI in their sample, and so this is the best single choice we could propose for the average, μ, of the BMI of all young Americans. The point of the 1.9 and the 95% is that if different, equally-large samples were randomly selected, we would expect that 95% of the time the average BMI in the sample would be between $\mu - 1.9$ and $\mu + 1.9$ but 5% of the time the sample average would be outside that range. So when we researchers give a range of values formed in this way, we are giving an interval that really does contain the right μ about 95% of the time (but 5% of the time that we report an interval like this, the correct value of μ is outside our range!) Remember, researchers who always do their computations with 95% confidence levels are likely to be wrong about once for every 20 statements they make!

Certainly changing the length of the interval just from 1.9 to 2.0 will barely increase the level of confidence we can report. More fundamentally we can NEVER give an interval with 100% confidence through sampling alone – you never know anything about the unsampled members of the population, and they can significantly change μ from $\bar{x}$. (As the next question demonstrates!)

5. You have data on an SRS of n recent graduates from the University of Texas that shows how long each person took to complete their degree. We want to use this SRS to estimate the mean, μ, of the statistical variable x = time-to-graduation, over all UT graduates. We would like to report our estimate to the press along with an 80% confidence interval.

While the study is being completed, a student named Chris Slacker finally manages to graduate after a record 23 years of continuous enrollment, so is now part of the population of all UT graduates, but is not part of the SRS that you have already selected.

Explain which of the following quantities change, and whether they get larger or smaller: n, μ, $\bar{x}$, the standard deviation σ, and the width of the 80% confidence level.

ANSWER: The quantities n and $\bar{x}$ are computed only from the SRS, so if poor Chris is not included in the sample, those quantities do not change.

But μ, the average length of time to graduation, is definitely changed: we are told that Chris set the record, so this new value of x is larger than any previous one, thus bringing up the average.

Note also that σ is sure to increase: since Chris's 23 years are a record, all the other deviations from average have to be less than Chris's. That is, an outlier like Chris is definitely increasing the "spread". (Here it's important to observe that time-to-graduation is not negative; if x were a variable that had some negative variables, then Chris's deviation from μ might actually be less than those of any very negative values of x.)

Finally, the width of the confidence level is $\sigma/\sqrt{n}$, so since Chris is not changing n but is increasing σ, as soon as Chris is included in the population, the confidence intervals have to be larger. (In fact that's obvious: Chris is single-handedly demonstrating that there is more uncertainty than we thought, in the process of predicting time-to-graduation...)

6. The Rusin Aspirin Company (RAC) makes the claim that taking an aspirin before taking a Statistics exam will substantially improve your score on the exam. Let's test this claim with the help of 1000 students taking Math 316. One hundred of them, selected at random, are given an aspirin before this exam. (The others are given a placebo. Those grading the exam have no idea which students got the aspirin.)

For each student we then compute the amount x that their exam 2 score improved from exam 1. This x is observed to follow an approximately Normal distribution. The average for all 1000 students is $\mu = +20$, and the standard deviation is $\sigma = 25$.

For the students who took the aspirin first, it is found that the average change in their test scores was $\bar{x} = +25$, which appears to support the claim made by RAC; but is this finding significant at say the $\alpha = 0.10$ level? At the $\alpha = 0.01$ level? Explain.

ANSWER: Since $\sigma = 25$, the standard deviation for samples of 100 is $\sigma/\sqrt{n} = 25/10 = 2.5$. Our sample thus reported a mean test improvement which is two standard deviations above average. According to Table A, a random Normal variable will be 2.0 or more standard deviations above average about 2.3% of the time. Since $2.3 < 10$, at the $\alpha = 0.10$ level we consider this finding to be significant; we agree that taking an aspirin has helped. But at the $\alpha = 0.01$ level (where we attribute to random fluctuation any event that can happen at least 1% of the time) we consider this finding to be insignificant; we believe that the observed results could "likely" have resulted by accident, and so we don't believe that it has been proved that aspirin help.

7. NASA wishes to determine the expected lifetime of rats in space. It is very expensive to send a rat into space so only 24 rats were included in the experiment. (A careful process was followed to make sure the rats were selected without any bias.)

The lifetime x (measured in days) for each of the rats was noted, and NASA computed the average $\bar{x} = 595$ and the standard deviation $s = 84$ of this sample. So what should NASA report as its estimate of the average lifespan of a rat in space, if it wants to give a 95% confidence interval?

ANSWER: Normal random variables following a normal distribution are "usually" "close" to their average value; among the possible ways to make that statement concrete are that 95% of the time, the random variable will be within 1.96 standard deviations of the average value. (See Table C, or the implicit line "$z = 1.96, Area = .975$" that can be interpolated in Table A.) So if we knew the standard deviation σ of all rat lifespans, we would report the estimated lifespan as $595 \pm 1.96\sigma/\sqrt{n}$ weeks.

But we don't know the standard deviation σ, although we can estimate it from the observed standard error s. In order to account for the additional uncertainty that results from using s instead of σ, we use Table C, reading the line with $n - 1 = 23$ degrees of freedom; the correct multiplier is now not 1.96 but rather 2.07. So we report our prediction of the average lifespan of a space rat as "$595 \pm (2.07)(84)/\sqrt{24}$ days, with 95% confidence." This works out to 595 ± 35.49 days, i.e. we are 95% confident that the true μ is between 559.5 days and 630.5 days.

Here are some of the entries from Table A:

z	$Area$	z	$Area$	z	$Area$	z	$Area$
−5.0	.0000	−1.2	.115	0.0	.500	1.3	.903
−4.0	.0001	−1.1	.136	0.1	.540	1.4	.919
−3.0	.001	−1.0	.159	0.2	.579	1.5	.933
−2.5	.006	−0.9	.184	0.3	.618	1.6	.945
−2.2	.014	−0.8	.212	0.4	.655	1.7	.955
−2.0	.023	−0.7	.242	0.5	.691	1.8	.964
−1.9	.029	−0.6	.274	0.6	.726	1.9	.971
−1.8	.036	−0.5	.309	0.7	.758	2.0	.977
−1.7	.045	−0.4	.345	0.8	.788	2.2	.986
−1.6	.055	−0.3	.382	0.9	.816	2.5	.994
−1.5	.067	−0.2	.421	1.0	.841	3.0	.999
−1.4	.081	−0.1	.460	1.1	.864	4.0	.9999
−1.3	.097	−0.0	.500	1.2	.885	5.0	1.0000

And here are some of the entries from Table C:

Degrees of Freedom	Conf Level C 80%	90%	95%	99%
1	3.08	6.31	12.7	63.7
2	1.89	2.92	4.30	9.93
11	1.36	1.80	2.20	3.11
12	1.36	1.78	2.18	3.06
23	1.32	1.72	2.07	2.81
24	1.32	1.71	2.06	2.80
25	1.31	1.71	2.06	2.79
z^*	1.28	1.65	1.96	2.58
One-sided P	0.10	0.05	0.025	0.01
Two-sided P	0.20	0.10	0.05	0.02