Statistics 1040, Section 006, Midterm 2 (200 Points)

Friday, November 12, 2004

Your Name: _____

based on: Stat 1040, Fall 2004 !, Quiz 7, October 22, 2004

Question 1: Chance/Probability (**50 Points**)

A big box with sports equipment contains 20 balls of which 6 are soccer balls, 10 are basketballs, and 4 are volleyballs. A little boy, barely able to reach into the box, randomly picks 2 balls to play with. **Show your work!**

6 soccer (S)
10 basketballs (B)
4 volleyballs (V)
20

-1 for each calculation error (or no final result in %)
-9 if % > 100% or % < 0%

1. (**10 Points**) What is the chance that the first ball is a volleyball?

   The chance is: 20.0 %

   first V: $\frac{4}{20} = 0.20 = 20.0\%$ (4) (2) (4)

2. (**10 Points**) What is the chance that at least one of the balls is a soccer ball?

   The chance is: 52.1 %

   first not S: $\frac{14}{20}$ (2) — dependent — second not S, given first not S: $\frac{13}{19}$ (2)

   both not S: $\frac{14}{20} \cdot \frac{13}{19} = \frac{182}{380}$ (2)(2) multiplication rule

   at least one S: $1 - \frac{182}{380} = \frac{198}{380} = 0.521 = 52.1\%$ (4) opposite rule

3. (**10 Points**) What is the chance that both balls are for the same game?

   The chance is: 34.7 %

   (SS) both soccer: $\frac{6}{20} \cdot \frac{5}{19}$

   (BB) both basketball: $\frac{10}{20} \cdot \frac{9}{19}$

   (VV) both volleyball: $\frac{4}{20} \cdot \frac{3}{19}$

   (SS) & (BB) & (VV) mutually exclusive (2)(2)(2)

   same game: $\frac{6}{20} \cdot \frac{5}{19} + \frac{10}{20} \cdot \frac{9}{19} + \frac{4}{20} \cdot \frac{3}{19}$

   $= \frac{30+90+12}{380} = \frac{132}{380} = 0.347 = 34.7\%$

4. (**10 Points**) What is the chance that one ball is a volleyball and the other ball is a basketball?

   The chance is: 21.1 %

   (VB) first V: $\frac{4}{20}$ — dependent — second B, given first V: $\frac{10}{19}$

   (BV) first B: $\frac{10}{20}$ — dependent — second V, given first B: $\frac{4}{19}$

   (VB) & (BV) mutually exclusive

   one V, one B: $\frac{4}{20} \cdot \frac{10}{19} + \frac{10}{20} \cdot \frac{4}{19}$ (2)(4)(4)

   $= \frac{40+40}{380} = \frac{80}{380} = 0.211 = 21.1\%$

5. (**10 Points**) What is the chance that both balls are volleyballs?

   The chance is: 3.2 %

   (VV) first V: $\frac{4}{20}$ — dependent — second V, given first V: $\frac{3}{19}$

   both V: $\frac{4}{20} \cdot \frac{3}{19} = \frac{12}{380} = 0.032 = 3.2\%$ (2)(4)(4) multiplication rule

Question 2: EV, SE, and Normal Curve (**50 Points**)

During the 2004 presidential elections, Kerry needed to win the state of Ohio to become the next president. Early on Nov 3, the day after Election Day, Bush had a 51% to 49% lead over Kerry, which related to about 140,000 more votes for Bush in Ohio. However, there were possibly up to 250,000 uncounted provisional ballots at that time. If Kerry could have gotten 140,000 of those, plus 1/2 of the remaining 110,000, plus 1, i.e., a total of 195,001, he would have won Ohio and would have been the next president. However, Kerry eventually conceded to Bush later on Nov 3 (even with many of the provisional ballots still being uncounted) because Kerry's advisors figured out that it was *statistically impossible* for Kerry to win Ohio and thus the election. **Show your work!**

1. (**10 Points**) Assume you are a highly optimistic advisor of Kerry, assuming that he might win up to 70% of the uncounted provisional ballots because a huge majority of these votes come from a population group close to the Democrats. Find the box model.

   1: Kerry
   0: Bush

70 × 1	30 × 0

   # draws: 250,000

   -2 if slightly incorrect number of 0/1 's in box
   -4 if box given as [0 1] etc.
   -6 if box contains something else than 0/1's
   -2 if # draws missing or incorrect

2. (**15 Points**) The expected number of votes for Kerry from the uncounted provisional ballots is 175,000 with an SE of 229.

   box avg = $\frac{70}{100} = 0.7$

   box SD = $\sqrt{\frac{70}{100} \cdot \frac{30}{100}} = \sqrt{0.21} = 0.458$

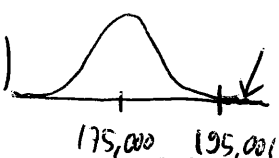
   $EV_{sum} = 250{,}000 \cdot 0.7 = 175{,}000$

   $SE_{sum} = \sqrt{250{,}000} \cdot 0.458 = 500 \cdot 0.458 = 229$

   -2 for each calculation error
   -2 for each minor mistake
   -3 for each major mistake (or step missing)

3. (**20 Points**) The chance that **at least** 195,001 of the uncounted provisional ballots are in favor of Kerry is about 0.0 %.

   S.U.: $\frac{195{,}001 - 175{,}000}{229} = \frac{20{,}001}{229} = 87.34$ (for all the table!)

   area between -4.45 and 4.45: 99.9991 %
   area between -87.34 and 87.34: almost exactly 100 %
   area above 87.34 %: basically 0.0 %

   -2 for each calculation error
   -4 for each incorrect curve parameter, i.e., anything else than EV & SE
   -4 for incorrect S.U.
   -4 for incorrect table value (i.e. not using area under the curve (i.e., not 0%)

4. (**5 Points**) So, do you agree that it was *statistically impossible* for Kerry to win Ohio and thus the election? (Yes) / No (5) - needs to match part "3."

   [Yes - we can really say that such an outcome is statistically impossible; moreover, our assumption that 70% of these votes may go to Kerry is already highly optimistic. Conceding was the best he could do from a statistical point of view.]

from: Stat 1040, Spring 2002, Midterm 2, March 27, 2002, Question 3
& Stat 1040, Fall 1999, Final Test, December 17, 1999, Question 8

Question 3: Sampling (**30 Points**)

In Web polls, anyone who views a certain Web page is allowed to vote by clicking on a button that represents their choice. In fact, there is nothing to stop someone voting as many times as they want. The results of one such poll suggest that almost 90% of the US population wants to ban firearm sales. The poll has a very large sample size (over 1 million).

1. (**21 Points**) Web based polls such as this are notoriously susceptible to bias. Give **three** possible sources of bias for this poll.

Answer on Web:

"1. People must own a computer - creates a bias towards younger, wealthier, more technologically-trained, etc., people.

2. People who feel strongly are more likely to vote + these people may vote multiple times.

3. The source will create a bias - what type of people read this Web page? Maybe it's a newspaper Web page - only people who find it will have the chance to vote."

+7 for each reasonable source of bias
-3 for repeating the same in different words
-5 if not a real source of bias

Additional explanation:
1. & 3. result in a selection bias,
2. is a typical problem of a voluntary response survey and leads to bias

2. (**9 Points**) Are the sources of bias you listed in Part 1. above a problem even with a very large sample, or does the sample size imply that they can be ignored? **Explain!**

Answer on Web:

"All a large sample does is repeat a mistake on a grand scale, if the data are subject to bias! So all we have is a large biased sample - we cannot ignore the problem."

-7 if no clear statement that sample is not the problem here

From: Stat 1040, Spring 2002, Final Test, April 30, 2002, Question 7

Question 4: Regression (**40 Points**)

-2 each calculation error

For a random sample of 20 car models, the average weight (in pounds) was 3236, with an SD of 523. The average gas mileage (in miles per gallon) was 21.4 with an SD of 4.2. The correlation between weight and gas mileage was −0.87. The scatter diagram was football shaped. **Show your work!**

-2 if x, y flipped
-1 if x, y not specified

1. (**10 Points**) Find the equation of the regression line for predicting gas mileage from weight.

	avg	SD
x: weight	3236	523
y: mileage	21.4	4.2

$r = -0.87$

$$\text{slope} = r \cdot \frac{SD_y}{SD_x} = -0.87 \cdot \frac{4.2}{523} = -0.00698x - 0.007 \quad (4)$$

$$\text{intercept} = \text{avg}_y - \text{slope} \cdot \text{avg}_x = 21.4 - (-0.007) \cdot 3236 = 21.4 + 22.652$$
$$= 44.052 \approx 44.1 \quad (4)$$

equation: $\boxed{\text{mileage} = 44.1 - 0.007 \cdot \text{weight}}$ or $\boxed{y = 44.1 - 0.007 \cdot x}$ (2)

2. (**10 Points**) Predict the gas mileage of a car that weighs 3500 pounds.

-2 for old method, correct result
-8 for old method, incorrect result

mileage for 3500 pounds: $44.1 - 0.007 \cdot 3500 = 19.6$ mpg

3. (**10 Points**) Find the r.m.s. error for prdicting gas mileage from weight of cars.

$$\text{r.m.s. error} = \sqrt{1 - r^2} \cdot SD_y$$
$$= \sqrt{1 - (-0.87)^2} \cdot 4.2$$
$$= \sqrt{1 - 0.7569} \cdot 4.2$$
$$= \sqrt{0.2431} \cdot 4.2$$
$$= 0.493 \cdot 4.2$$
$$= 2.07 \text{ mpg}$$

-4 for each major mistake, e.g. SD_x instead of SD_y, $\sqrt{}$ of everything, r instead of r^2, etc.

4. (**10 Points**) Would you be surprised if someone told you that one of these cars weighing 3500 pounds got 27 miles per gallon? Explain your answer using the r.m.s. error.

$$\text{s.u.: } \frac{27 - 19.6}{2.07} = \frac{7.4}{2.07} = 3.57 \text{ s.u.} \quad (5)$$

27 is about 3.6 r.m.s. errors above the predicted value of 19.6; (2)

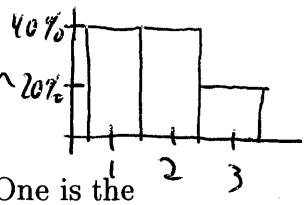
so, yes (3), this is very unusual and surprising...!

from: Stat 1040, Spring 2003, Midterm 2, March 28, 2003, Question 5
& FPP, p. 328, Review Exercise 5

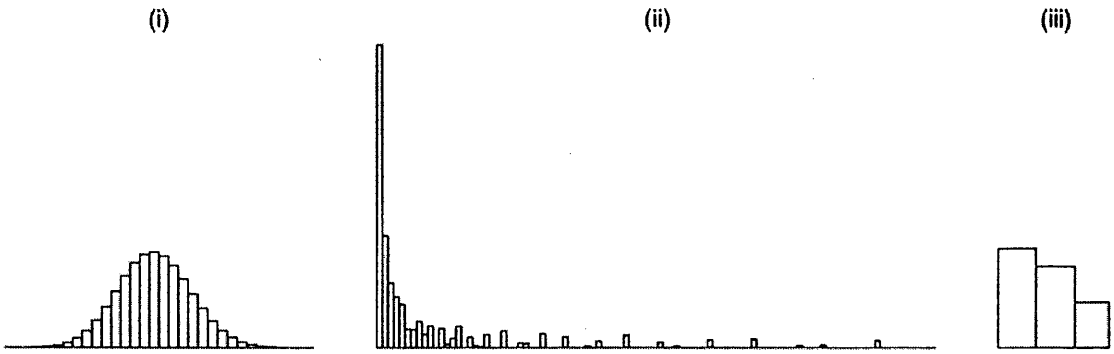
Question 5: Normal Approximation for Probability Histograms (**30 Points**)

Twenty–five draws are made at random with replacement from the box $\boxed{1}\boxed{1}\boxed{2}\boxed{2}\boxed{3}$.

Probability histogram for box:



One of the graphs below is an (empirical) histogram for the numbers drawn. One is the probability histogram for the sum. And one is the probability histogram for the product. Which is which? **Explain!**



- An (empirical) histogram for the numbers drawn is ___(iii)___. (7)
  Explanation:

  According to the probability histogram (see above), we should have 40% 1's, 40% 2's, and 20% 3's. The empirical histogram after 25 draws will somewhat resemble the probability histogram, but not too closely. (3)

- The probability histogram for the sum is ___(i)___. (7)
  Explanation:

  The probability histogram (see above) is not perfectly symmetric, but also not very asymmetric. The probability histogram for the sum will follow the normal curve even after only 25 draws. (3)

- The probability histogram for the product is ___(ii)___. (7)
  Explanation:

  The probability histogram for the product typically does not follow the normal curve. (3)