Midterm

COR1-GB.1305 – Statistics and Data Analysis

The exam is open book and notes. You are also permitted use of a calculator. Each part of each problem is worth 5 points, except for the matching problem, which is worth 10 points total. There are 75 points total. There is no penalty for guessing incorrectly on a multiple choice problem. Partial credit may be awarded if you show work.

For the problems involving calculations, you must show all work to get full credit. For short-answer problems, there should not be any symbols in your final answer (p, n, μ, etc.), but you do not need to fully simplify your answer. It is ok to have quantities like ${}_5C_2$, $\sqrt{3.1}$, etc. in your final answers on these problems.

NYU Stern Honor Code:

I will not lie, cheat or steal to gain an academic advantage, or tolerate those who do.

Signature: _____ Date: _____

Name: Solutions

Short Answer

1. (10 points) The NYU Office of Student Affairs wants to estimate the average salary of all current Langone students. To this end, they email all current Langone students, asking them to fill out an online survey and report their current salaries. Assume that there is no measurement error, so that all reported salaries are accurate. The sample consists of the reported salaries for the students who participated in the survey.

   (a) What is a reasonable population for this sample? (1 sentence.)

Salaries of all current Langone students

   (b) Is there bias in the sample? Why or why not? Answer in 1–2 sentences.

Yes. People with low salaries might be embarrassed and less likely to respond.

(other answers are also valid)

2. (15 points) I have 100 pennies in a bag. I shake the bag repeatedly so that the orientations of the coins are completely random. Then, I empty the bag onto my desk, and I count the number of pennies showing "heads".

(a) Let X denote the number of pennies showing heads. Compute the expectation $\mathrm{E}[X]$ and the standard deviation $\mathrm{sd}[X]$.

X is binomial with $n = 100$, $p = 0.50$

$E[X] = np = 100(0.50) = 50$

$sd[X] = \sqrt{np(1-p)} = \sqrt{100(0.50)(1-0.50)} = 5$

(b) What is the interpretation of $\mathrm{E}[X]$? Answer in 1–2 sentences.

The "long run average" of X. If we repeatedly shake 100 pennies in a bag, dump them onto the desk, and count the number of heads, then the average number of heads (after performing many trials) will be close to $E[X] = 50$.

(c) Would it be unusual to see 75 or more pennies showing "heads"? Justify your answer by computing a rough approximation of $\mathrm{P}(X \geq 75)$.

The binomial can be written as

$$X = Y_1 + Y_2 + \cdots + Y_n$$

where $Y_i = \begin{cases} 1 & \text{if } i\text{th coin is heads} \\ 0 & \text{otherwise.} \end{cases}$

Thus it would be very unusual to see $X \geq 75$

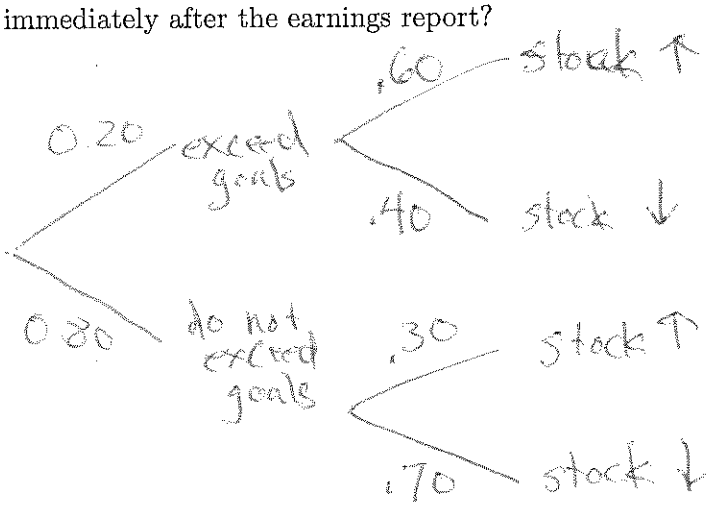
Since n is large, the CLT says X is approximately normal. Thus, $P(X \geq 75) = P\left(\frac{X-\mu}{\sigma} \geq \frac{75-\mu}{\sigma}\right)$ with $\mu = 50$, $\sigma = 5$. Hence, $= P\left(Z \geq \frac{75-50}{5}\right) = P(Z \geq 5)$

$\approx .0000002867$

Multiple Choice

3. (5 points) Twitter is going to release an earnings report at the end of the month. When they make this report there is a 20% chance that they will have met or exceeded their revenue goals. If this happens (if they meet or exceed their revenue goals), then there is a 60% chance that their stock price will go up immediately after the report, and a 40% chance that it will go down. Conversely, if they do not meet or exceed their revenue goals, then there is a 30% chance that their stock price will go up, and a 70% chance that it will go down. What is the chance that their stock price will go up immediately after the earnings report?

   A. 36%
   B. 45%
   C. 72%
   D. 12%
   E. 90%



$P(\text{stock } \uparrow) = (0.20)(0.60) + (0.80)(0.30) = 0.36$

4. (5 points) Consider the dataset of $n = 162,997$ taxi trips from New York City. The mean fare ($) was 12.424, and the median fare ($) was 10.000. Suppose we take the square roots of all 162,997 fares, and then compute the mean and median of the square roots. Which of the following must be true for the $n = 162,997$ square roots (up to 3 decimal places)?

   A. mean is 3.525; not enough information to determine the median
   B. mean is 3.525; median is 3.162
   C. median is 3.162; not enough information to determine the mean
   D. mean is 3.525; median is 10.000
   E. not enough information to determine the mean or the median

$\sqrt{12.424} = 3.525$ $\quad \sqrt{10} = 3.162$

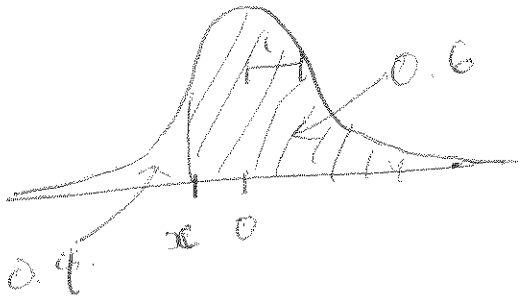
$\sqrt{EX} \neq E\sqrt{X}$ in general, mean $\neq 3.525$ necessarily

median of $\sqrt{\ }$ = $\sqrt{\ }$ of middle value

$= \sqrt{\text{median}}$ (since $\sqrt{\ }$ doesn't change order)

5. (5 points) If X is a normal random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$, what is the value x with $P(X > x) = 0.6$?

A. 0.2743
B. -0.2743
C. 0.6554
D. -0.2533
E. 0.2533



z - table: $x = -0.2533$

6. (5 points) You invest in 8 new startups. Each startup has a 10% chance of succeeding (giving you a positive return on your investment). Assume that, even though you are only investing in startups, your portfolio is diverse enough so that it is reasonable to assume that all 8 startups are independent of each other. Approximately what is the chance that 2 or more of your investments succeed?

A. 0.187
B. 0.383
C. 0.020
D. 0.813
E. None of the above.

Let X = # succeed

then X is Binomial, $n = 8$, $p = 0.10$

$P(X \geq 2) = 1 - P(X < 2)$

$= 1 - \{P(X=0) + P(X=1)\}$

$= 1 - (.4305 + .3826)$

$= 1 - .8131$

$= .1869$

$P(X=0) = {}_8C_0 (0.10)^0 (0.90)^8$

$= (0.90)^8$

$= .4305$

$P(X=1) = {}_8C_1 (0.10)^1 (.90)^7$

$= 8(0.10)(.90)^7$

$= .3826$

7. (10 points) A Gallop poll surveyed 100 likely voters, asking them which party they were likely to vote for in the next presidential election: Democrat, Republican, or Other. The results of the poll, segregated by age group, are tallied in the following tabe.

	Vote			
Age	Democrat	Republican	Other	Total
18–35	12	8	4	24
35–50	10	11	2	23
50–65	7	15	0	22
Over 65	13	13	5	31
Total	42	47	11	100

(a) What is the probability that a random survey respondent aged 35–50 says they will vote Republican?

A. $\frac{11}{100}$
B. $\frac{23}{100}$
C. $\frac{11}{47}$
D. $\frac{47}{100}$
E. $\frac{11}{23}$

$P(\text{Repub} \mid 35\text{–}50) = \frac{11}{23}$

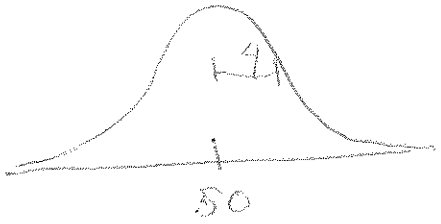
(b) Given that a random survey respondent says he or she will vote Democrat, what is the probability that the respondent is aged 18–35?

A. $\frac{12}{100}$
B. $\frac{24}{100}$
C. $\frac{12}{42}$
D. $\frac{42}{100}$
E. $\frac{12}{24}$

$P(18\text{–}35 \mid \text{Dem}) = \frac{12}{42}$

8. (5 points) Over the last 40 years, the average annual precipitation in New York City was 50 inches, and the standard deviation was 4 inches. The histogram of the rainfalls looks like a bell curve. How much precipitation would you expect to see in New York City in a typical year? (Take "typical" to mean "roughly 95% of the time.")

A. 46 to 54 inches
B. 42 to 58 inches
C. 48.74 to 51.26 inches
D. 49.37 to 50.63 inches
E. None of the above.



typical / 95%:

$50 \pm 2 \cdot (4)$

50 ± 8

$(42, 58)$

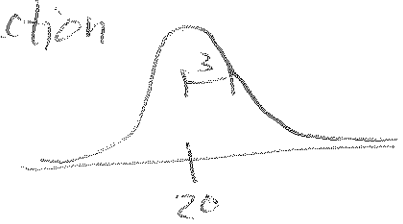
9. (5 points) Daisy the Dairy Farmer has five milk-producing cows. Each cow behaves similarly, producing a random amount of milk each day. The milk production for a single cow is approximately normally distributed with mean 20 liters and standard deviation 3 liters. If the milk productions of all cows are independent of each other, approximately what is the probability that Daisy's five cows will produce a total of at least 110 liters of milk tomorrow?

A. 0.06681
B. 0.93319
C. 0.0004835
D. 0.2420
E. 0.7580

X = single cow production

$\mu = 20$

$\sigma = 3$



$\text{total} = X_1 + X_2 + X_3 + X_4 + X_5$

$\bar{X} = \frac{\text{total}}{n}$ with $n = 5$

$P(\text{total} \geq 110) = P(\bar{X} \geq \frac{110}{5}) = P(\bar{X} \geq 22)$

Now, $\mu_{\bar{X}} = \mu = 20$; $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{5}} = 1.3416$

$P(\bar{X} \geq 22) = P(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \geq \frac{22-20}{1.3416}) = P(Z \geq 1.49)$

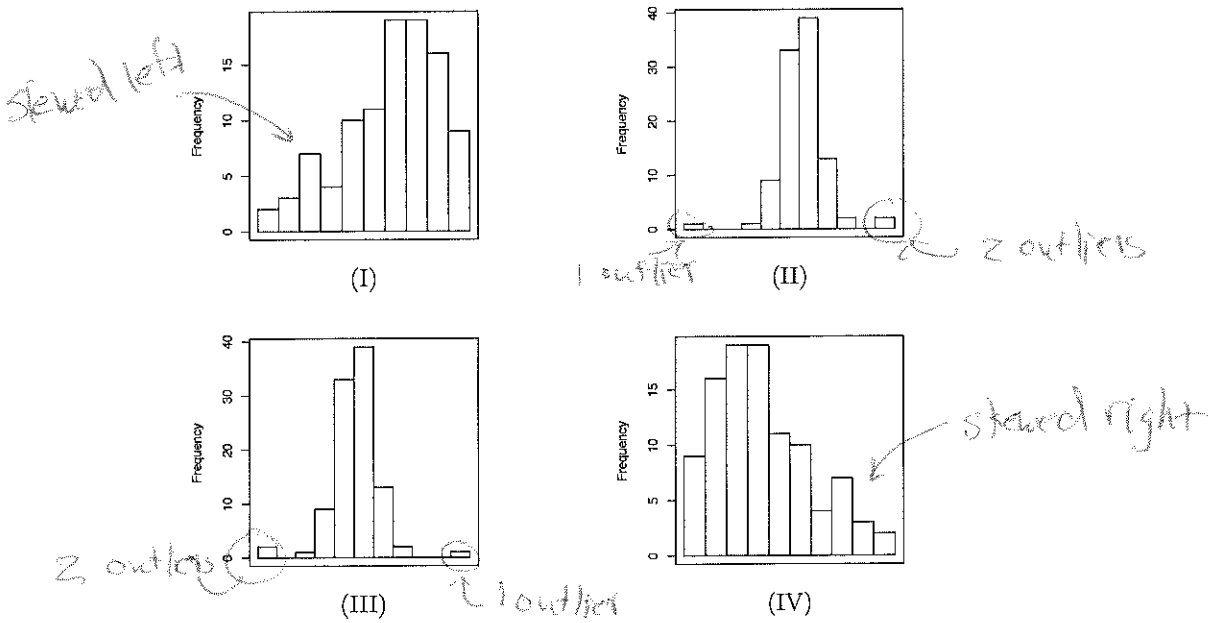
$= 1 - P(Z < 1.49)$

$\approx 1 - .93319$

$= .06681$

Matching

10. (10 points) Here are the histograms for four different datasets, labeled (I)–(IV):



For parts (a)–(d), match the appropriate histogram with the given box-and-whisker plot. Note that the scales for the plots are all different, and they do not necessarily match the scales of the histograms.