Midterm 1: Elementary Statistics *(Solutions)*.

Prof. Jordan C. Hanson

August 2, 2020

1 Formula Area

1. Quantitative continuous data: sample data that can be measured

2. Quantitative discrete data: sample data that can be counted

3. Qualitative or categorical data: sample data that can be classified but not counted

4. Average/mean, definition 1: $\bar{x} = N^{-1} \sum_i x_i$ for a sample size N

5. Median: the value below which are half of the frequencies. Half of the frequencies are also above this value.

6. Mode: the value corresponding to the highest frequendy.

7. The quartiles $Q1$, $Q2$, and $Q3$ are the values that separate the frequencies into four bins of equal frequency. $Q2$ is equal to the median. The IQR is $Q3 - Q1$.

8. The k-th percentile: the value below which k percent of the data is located. Formula: $i = (k/100)(n+1)$, where k is the percentile, n is the total number of data, and i is the integer location of the k-th percentile.

9. Finding the percentile of a data value: $(x + 0.5 * y)/n(100)$, where x is the number of data values below the given data value, y is the number of data values equal to the given one, and n is the total number of data values.

10. Average/mean, definition 2: $\bar{x} = \sum_i^M f_{r,i} x_i$, where x_i are the bin centers of a histogram, or the discrete random variable data values, M is the number of bins, and $f_{r,i}$ are the relative frequencies. For a discrete random variable, $f_{r,i}$ is replaced with $p(x)$, the probability distribution function.

11. Probabilities of mutually exclusive and independent events: if two events have probabilities p_1 and p_2, then the probability that event 1 AND event 2 occur is $p_1 p_2$. The probability that event 1 OR event 2 occurs is $p_1 + p_2$.

12. The standard deviation s of a sample of size N is

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2 \tag{1}$$

13. Mean or expectation value of a binomial distribution: $\mu = Np$

14. Standard deviation of a binomial distribution: $\sigma = \sqrt{Np(1-p)}$

2 Unit 0

1. Suppose we measure 10 resting heart rates from 10 college students...during finals, and they've each had coffee. We find: 59, 60, 70, 75, 75, 75, 76, 77, 77, and 78 beats per minute. Provide the following:

   - What is the sample size? **10**
   - What is the mean heart rate? **72**
   - What is the standard deviation of the heart rates? **7**
   - Describe one issue with the sample that affects its randomness. How would you get a more complete sample of the student population at Whittier College? **There is some bias in the sample: it's finals week, and they all have caffeine in the system. Also $N = 10$, which is not large.**

2. Whittier College collects data on the number of students that apply each year, and the proportion we accept and who enroll. See Fig. 1, and copy the data into a spreadsheet like LibreOffice Calc or Microsoft Excel.

Whittier College			
Fall of	Applied	Admitted	Enrolled
2006	3120	1804	344
2007	2214	1485	312
2008	2206	1591	421
2009	2285	1638	359
2010	2900	2038	453
2011	2993	2139	427
2012	4125	2622	417
2013	4380	2771	446
2014	4850	3001	388
2015	5192	3267	445
2016	5146	3587	426
2017	5773	4277	520
2018	6220	4724	512
2019	7187	5369	493

Figure 1: A table of the number of freshmen who applied, were accepted, and were enrolled in Whittier College, versus year.

   - What is the mean number of newly enrolled freshmen per year from 2006 - 2019? **426 students.**
   - Define the *acceptance rate* as the second column of Fig. 1 divided by the first. What is the average acceptance rate from 2006 to 2019? **68 percent.**
   - What is the standard deviation of the acceptance rate from 2006 - 2019? Are there any outliers? **2019, 2018, and 2006 are all more than 1 standard deviation from the mean.**
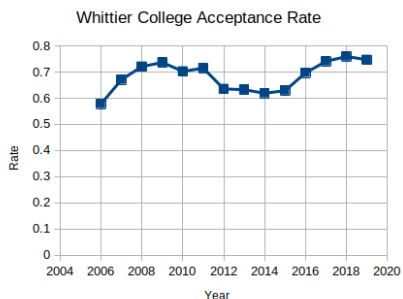   - Graph the *time-series* of the acceptance rate.

Figure 2: The acceptance rate of Whittier College freshmen for 2006-2019.

3. In a particular fund, there are 10 stocks, each with the following price per share in USD: 109.00, 108.00, 112.00, 113.00, 113.00, 120.00, 151.00, 170.00, 250.00, and 290.00. (a) What price represents the 75th percentile? (b) To what percentile does 113.00 dollars correspond? (c) What is the standard deviation and mean of the data? (d) Create a histogram of the data. Do you notice skew? **(a) 170 dollars. (b) 40th percentile. (c) 150 dollars and 65 dollars, respectively. (d) See Fig. 3. The median is 116 dollars but the mean is closer to 150, so there is skew to the left. Several prices are 0.5 standard deviations (approximately) below the mean.**
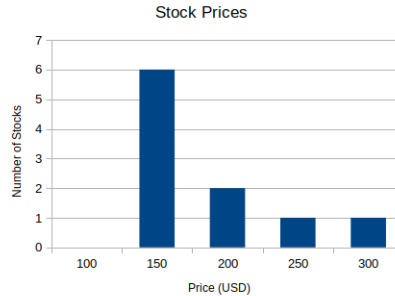

Figure 3: The prices of stocks in the group, as a histogram.

4. **Pirate Dice.** Imagine you are kidnapped by pirates, and to pass the time between scrubbing the deck and serving grog, you wager on dice. Each of four players has two dice. Each player rolls his or her dice but hides them from sight after looking at them. The players take turns stating how many instances of a number will appear among the 8 dice. If any player thinks the claim is false (a lie), then they call **liar**. If the player was wrong, they are out. For example, if each pair of dice are: (1,2), (4,4), (4,6), and (3,5), then there are three "fours" all together. Calling four "fours" would be a lie. (a) Your dice say (1,1). An opponent declares that "there be *five ones*" in play. What is the probability this is true? (b) Your dice say (2,3). An opponent declars that "there be *six fives*!" What is the probability? **(a) 5.36 percent. There are 20 ways this can happen, and each way has a probability of $(1/6)^3(5/6)^3$. (b) 2.1×10^{-5}. There's only one way this can happen, and it has a probability of $(1/6)^6(5/6)^0$.**

5. Consider a *fair coin* where the probability of a heads, P(H), is 0.5, and the probability of a tails, P(T), is 0.5. Suppose we model the random motion of a gas molecule in 1D like a fair coin, in which movement left by one unit has a probability L = P(T) and movement right by one unit has a probability R = P(H). (a) What is the probability that a molecule follows this path: LRLLRLRR? (b) What is the probability a molecule follows this path: RRRRRRRR? (c) Which is more common, a path that leads back to the starting point, or the path in part (b)? **(a) $1/2^8$. (b) The same as part (a). (c) The former. There are more ways that this can happen, so these probabilities would be summed.**

3 Unit 1

1. Suppose a stock trader agrees to purchase stock at a future date, according to a contract that stipulates she *must* purchase it, regardless of the price. However, she negotiates that the price will be measured into one of four bins, the centers of which are shown in Tab. 1. This makes the price a discrete random variable. She performs an analysis that gives the probability that the stock price will fall into each of the categories. If she buys one share, what is the *expectation value* of her profit? What would be her profit if she buys 1000 shares? **The expectation value of the discrete random variable x is 0.44 dollars per share. Purchasing 1000 shares will resolve to 440 dollars.**

2. Suppose a psychologist is studying whether or not people can decide if another person has an IQ score of greater than 100. Ten different people are shown ten different images of another person for 2.0 seconds. So you can think of an individual photo as a trial, and each participant as a run. The frequencies with which the participants said yes, $IQ > 100$ is shown in Tab. 2. (a) Fill in Tab. 2, given what you know about the experiment. (b) Are the participants guessing randomly? Why or why not? (c) Are the data distributed following a binomial distribution? Assuming they are, what is the p value that satisfies $\mu = Np$? **(a) See Tab. 2. (b) I do not think the participants are guessing randomly, because the expectation value is 6.4 positive IDs per run of 10 people. If they were guessing randomly, I would expect 5.0 positive IDs per run of 10 people. (c) Further, if the**

Outcome	x	$p(x)$	$x * p(x)$
Price bin 1	$90.00 per share	0.01	?
Price bin 2	$16.00 per share	0.49	?
Price bin 3	-$15.00 per share	0.49	?
Price bin 4	-$95.00 per share	0.01	?

Table 1: A table displaying a stock trader's assessment of the probability a stock will fall into one of four bins. (The bin centers are shown).

x	N_{good}	$p(x)$	$x * p(x)$
0	0	0	0.0
1	0	0	0.0
2	0	0	0.0
3	0	0	0.0
4	1	0.1	0.4
5	0	0	0.0
6	3	0.3	1.8
7	6	0.6	4.2
8	0	0	0.0
9	0	0	0.0
10	0	0	0.0

Table 2: A table displaying the frequencies with which participants decided random photos of people corresponded to an IQ score of larger than 100 or not. The left column is the discrete random variable x, the total number of times the participant decided a photo corresponding to a high IQ. The middle column is the frequency with which x occurs across 10 runs.

data are distributed binomially, then the standard deviation would be 1.5, so the mean is almost one sigma above the mean. The measured standard deviation of the sample is even smaller, 0.91. So the participants are definitely not randomly guessing. The p-value would be 0.64.