Taylor Watanabe

WC I.D. · 20594796

Midterm 1: Elementary Statistics

Prof. Jordan C. Hanson

July 30, 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: $\sigma = sqrtNp(1-p)$

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?
   - What is the mean heart rate?
   - What is the standard deviation of the heart rates?
   - 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?

Sample size : Randomly selected group of students enrolled @ a school during finals who've drank coffee (10 students)

Sample = "Randomly selected group of students enrolled @ a school during finals who've drank coffee"

Sample space = $\{59, 60, 70, 75, 75, 75, 76, 77, 77, 78\}$

Mean ($\bar{x}$) heart rate : $\bar{x} = 72.2$ beats per minute

Standard deviation (σ) of heart rate : $(s^2) = (59-72.2)^2+(60-72.2)^2+(70-72.2)^2+(75-72.2)^2+\ldots(1/N-1) \quad = 49.5$

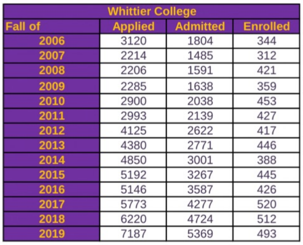
$$\sqrt{s^2} = \sqrt{49.5}$$

$$= 7.04$$

72.2 ± 7.04

Issue : "they've each had coffee" - affects the randomness of the sample. To get a more complete sample of Whittier College's student population, you could change your sample to be 10 resting heart rates of 10 students during finals week that enter the libary.

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.

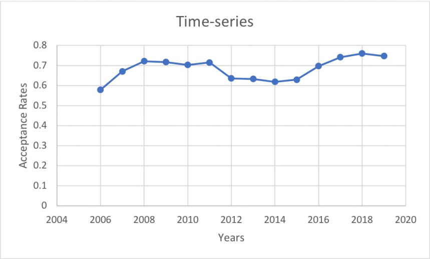
- What is the mean number of newly enrolled freshmen per year from 2006 - 2019?
- Define the *acceptance rate* as the second column of Fig. 1 divided by the first. What is the average acceptance rate from 2006 to 2019?
- What is the standard deviation of the acceptance rate from 2006 - 2019? Are there any outliers?
- Graph the *time-series* of the acceptance rate.

Mean of enrolled : 425.9 students

Average acceptance rate : 0.68 = 68% acceptance

Std. dev. of acceptance rates : 0.05

Outliers : if 0.68 ± 0.05 ; years 2017-2019 could be classified as outliers as they are above the standard deviation of the mean w/ 2017 = 0.74. 2018 = 0.76 and 2019 = 0.75. Year 2006 can also be said to be an outlier being less than the calculated variance amongst the mean with it having an acceptance rate of 58% (0.58).



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) 75^{th} percentile : $i = (k/100)(n+1)$

$= (75/100)(10+1)$

$i = 8.25 \quad = \quad$ 170.00 USD

b) $x = 3$, $y = 2$, $n = 10$

$(x + 0.5y)/n \times 100$

$\frac{3+0.5(2)}{10} \times 100 = 40^{th}$ percentile

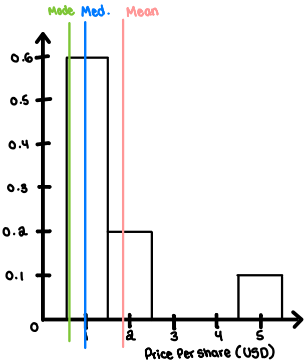
c) μ = 153 6 USD

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} = 65.5$$

153.6 ± 65.5

d)

	Bin	Rel. Freq.
1.	100 - 140	0.6
2.	141 - 182	0.2
3.	183 - 224	0
4	225 - 266	0.1
5.	267 - 308	0.1



Histogram shows a positive skew.

∴ direction of outliers is to the right.

$Q_3 - Q_2 > Q_2 - Q_1$

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) P(5 ones in play) = P(3 more ones in 3 rolls)

$P(1,1) = 1/6 \times 1/6$ $\quad = (1,1), (1,x), (x,x)$

$P(1,x) = 1/6 \times 5/6$ $\quad = (1/36)(5/36)(25/36) \overset{\text{same}}{\underset{\text{as}}{\longleftrightarrow}} (1/6)(1/6)(1/6)(5/6)(5/6)(5/6)$

$P(x,x) = 5/6 \times 5/6$ $\quad$ P(5 ones in play) = 125 / 46656 = 0.00268

b) Event : (2,3), (5,5), (5,5), (5,5)

$= (1/6)(1/6)(1/6)(1/6)(1/6)(1/6)$

$= (1/36)^3$

P(6 fives) = 1 / 46656 = 2.14×10^{-5}

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) $P_{Tot.} = (1/2)^8 = 1/256 \approx 0\ 39\%$

b) $P_{Tot.} = (1/2)^8 = 1/256 \approx 0.39\%$.

c) While it's intuitively known that a path that leads back to the starting point is more common, flipping a fair coin is an independent event so the individual probability of the molecule to go right or left is the same thus both arrangements are equally likely to occur.

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?

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).

$x * p(x) = 0.9\ ,\ 7.84\ ,\ -7\ 35\ ,\ -0.95$

Expectation value $= \sum_i x * p(x)$

a) Expectation value = \$0.44 per share

b) 0.44 dollars / ~~share~~ X 100 ~~shares~~

Profit = \$44

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 $\mu = Np$?

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

a) $p(x) = N/n$ where n = tot. # of runs

b) No, the photos shown are randomly selected but the participants are deciding the IQ level of the people in the photos based on each participants own set of criteria/opinions, not randomly guessing.

c) There's a fixed number of trials (100 photos), the random variable is discrete (x), Only two possible outcomes, thus $p+q=1$, and the n trials are independent. So yes that data distributed is following a binomial distribution.

$$\mu = Np$$

$$p = \mu/N = 6.4/10$$

$$p = 0.64$$