Study Guide for Midterm 1: Elementary Statistics

Prof. Jordan C. Hanson

July 26, 2022

1 Formula Area

1. Quantitative continuous data: sample data that can be measured. Quantitative discrete data: sample data that can be counted. Qualitative or categorical data: sample data that can be classified but not counted.

2. Average/mean, definition 1: $\bar{x} = N^{-1} \sum_i x_i$

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

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

5. 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 $Q3 - Q1$.

6. 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.

7. 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.

8. 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, 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.

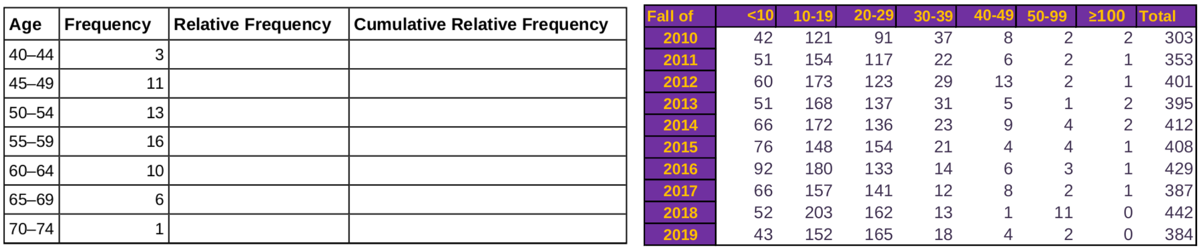
9. Probabilities of mutually exclusive and independent events: if two events have probabilities p_1 and p_2, then the probability that event 1 AND event 2 occur is $p_1 p_2$. The probability that event 1 OR event 2 occurs is $p_1 + p_2$.

10. The standard deviation s of a sample is

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

2 Unit 0

1. Forbes magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, have a stock price of at least \$5 per share, and have reported annual revenue between \$5 million and \$1 billion. Figure 1 (left) shows the ages of the chief executive officers for the first 60 ranked firms.



Age	Frequency	Relative Frequency	Cumulative Relative Frequency
40–44	3		
45–49	11		
50–54	13		
55–59	16		
60–64	10		
65–69	6		
70–74	1		

Fall of	<10	10-19	20-29	30-39	40-49	50-99	≥100	Total
2010	42	121	91	37	8	2	2	303
2011	51	154	117	22	6	2	1	353
2012	60	173	123	29	13	2	1	401
2013	51	168	137	31	5	1	2	395
2014	66	172	136	23	9	4	2	412
2015	76	148	154	21	4	4	1	408
2016	92	180	133	14	6	3	1	429
2017	66	157	141	12	8	2	1	387
2018	52	203	162	13	1	11	0	442
2019	43	152	165	18	4	2	0	384

Figure 1: (Left) A table of the ages of CEOs of the top 60 ranked small firms list, according to *Forbes* magazine. (Right) The number of classes at Whittier College, binned by their size, in number of students, and by year.

- What is the frequency for CEO ages between 54 and 65?

- What percentage of CEOs are 65 years or older?
- What is the relative frequency of ages under 50?
- What is the cumulative relative frequency for CEOs younger than 55?
- Graph the relative the cumulative relative frequency below.
- Create a Pareto graph of the relative frequencies below.

2. Consider Fig. 1 (right).
   - Create a time-series plot that contains both the *smallest class size* column, and the *30-39 student* column. What do you notice?
   - What was the mean class size of Whittier College in 2019?
   - In a particular fund, there are 10 stocks, each with the following price per share in USD: 14,14,15,17,19,21,25,50,72,90. (a) What is the median? (b) What price represents the 60th percentile? (c) To what percentile does 19 dollars correspond? (d) What is the standard deviation and mean of the data?
3. A lottery is constructed by labeling tokens with all the letters of the alphabet. There are 26 tokens, and five are drawn, *without replacement.* If you have to match all five tokens, what are your odds of winning? (Your card can have any string of five letters, without repeating a letter).
4. Suppose a student is applying to five colleges, each with an equal probability p of accepting him. (a) What is the probability that he is accepted to all five? (b) What is the probability that he is accepted to any two of the five colleges? (Think of it like: the first college OR the second college). (c) **Harder question**: what is the probability he gets accepted to none of them?

3 Unit 1

1. Some stock traders engage in *high frequency trading*, in which they write an algorithm that executes a pre-designed purchase or sale of a stock via computer code that runs for several microseconds. Suppose the TelCo stock is fluctuating rapidly between \$10 dollars per share and \$20 dollars per share, with an average of \$15 dollars. Consider Tab. 1 below, which explains the high-frequency strategy. (a) What is the expectation value of the discrete random variable x, the money made per trade? (b) If this code earns the expectation value once per day, after how many

Outcome	x	$p(x)$	$x * p(x)$
Buy	-\$14.95 per share	0.49	?
Sell	+\$15.05 per share	0.51	?

Table 1: The code buys stock when it has a price \$14.95 49, and this occurs 49 percent of the time. The code sells stock when it has a price of \$15.05, and this occurs 51 percent of the time.

days will the profit exceed 100 dollars? (c) If this code earns the expectation value every 100 microseconds, and runs for 12 hours, what is the profit?

2. Suppose the trading goes on from the previous problem and the profits for one week days in USD are 1.20, 4.05, 3.45, 0.90, 1.10, 2.40, and 0.50. What is the standard deviation in the profits? What is the mean? How many standard deviations is the last data point *below* the mean? (That is, how rare is this?).

3. **Extra: continuous probability density functions.** A subway train arrives every 8 minutes during rush hour. We are interested in the length of time a commuter must wait for a train to arrive. The time follows a uniform distribution. (a) Define the random variable. (b) Graph the distribution. (c) What, theoretically, is the mean of the distribution? (d) Theoretically, what is the standard deviation? (e) Ensure the PDF is normalized. (f) What is the probability that passengers must wait between 2 and 6 minutes?

4. **Same scenario as the prior question.** Suppose we find that, if we cast the passenger wait times as integers, and it turns out the distribution actually follows a *normal distribution* $N(4, 2)$ (minutes). Using Microsoft Excel or LibreOffice Calc, estimate the probability that passengers will wait between 6 and 8 minutes. (*Hint for Calc: use the menu options Sheet → Fill Cells → Fill Random Number. Then choose the normal distribution.*)