Elementary Statistics: Math 080

Jordan Hanson

July 11, 2022

Whittier College Department of Physics and Astronomy

Stem	Leaves
0	
1	
2	
3	
4	[3.0]
5	[6.0]
6	[7.0, 9.0]
7	[8.0, 0.0, 8.0, 1.0, 2.0, 5.0, 7.0]
8	[8.0, 3.0, 4.0, 6.0, 2.0, 1.0, 2.0, 1.0]
9	[8.0, 7.0, 1.0, 4.0]

Table 1:

Unit 0 Outline

1. Topics from Chapter 1: 1.1, 1.2, 1.3
    - What is a statistic?
    - Probability examples
    - Data and sampling
2. Topics from Chapter 2: 2.1 - 2.4, 2.5 - 2.8
    - Data visualization
    - Location of the data in numerical space
3. Topics from Chapter 3: 3.1, 3.2, 3.3
    - Two rules of probability

Topics from Chapter 2

Stemplots

Useful for numbers like *grades*. Most significant digit is the category.

Stem	Leaves
0	
1	
2	
3	
4	[3.0]
5	[6.0]
6	[7.0, 9.0]
7	[8.0, 0.0, 8.0, 1.0, 2.0, 5.0, 7.0]
8	[8.0, 3.0, 4.0, 6.0, 2.0, 1.0, 2.0, 1.0]
9	[8.0, 7.0, 1.0, 4.0]

Table 2: A *stemplot* of a grade distribution.

Stemplots

Procedure:

1. Identify the approximate order of magnitude of the sample.
2. Within that order of magnitude, create ≈ 10 *stems*, corresponding to the base-10 digits.
3. For each data point, call the non-most significant digits the *leaves* and drop the leaves in the category with the matching leaf.

Professor example: What is the stemplot of

`[11, 22, 33, 44, 55, 66]`

Stemplots

Procedure:

1. Identify the approximate order of magnitude of the sample.
2. Within that order of magnitude, create ≈ 10 *stems*, corresponding to the base-10 digits.
3. For each data point, call the non-most significant digits the *leaves* and drop the leaves in the category with the matching leaf.

Let's create a stemplot of:

1. Our ages in MATH080
2. My age and the rest of my department

(Stemplots lead in to the topic of histograms)

Histograms

Histograms are a tool for measuring *probability distributions*. The inputs are the data points and the corresponding relative frequencies, or plain frequencies.

How many textbooks or books did you purchase for school last year? (Type in the chat).

1. Determine the bins, or *binning*
2. For each data point, drop it into the appropriate bin
3. Each time a measurement is dropped into a bin, the *count* increases by 1.
4. If a histogram displays plain frequencies, it is called *un-normalized.*
5. If a histogram displays relative frequencies, it is called *normalized.*

Histograms

1. Histogram of books, by hand
2. Repeat with Excel/Calc

Practice with the FREQUENCY function in Calc/Excel:

```
=FREQUENCY(A1:A99; B1:B11)
```

Then press control+shift+enter to execute on arrays of data and bins. To *normalize*, input the relative frequencies, or divide frequecies by *N*. Assume the data is in C column:

```
=C1/N ...
```

Histograms

For data that is appropriately “stationary,” we can use histograms to estimate the mean *faster*, since we only have to loop over bins rather than every data sample. Let H_i represent the counts in a given bin, and i represent the bin sample. We have:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{M} iH_i \tag{1}$$

To obtain the mean in signal *amplitude*, you’ll have to convert bin number to amplitude. **Professor example.**

When is a histogram appropriate? **Note**: There is a distinction between the *process or signal process* and the *the data*. Just because the data has a given $\bar{x}$ and s does not imply that the signal process has or will continue to have the exact same values of μ and σ. The underlying process could be *non-stationary*.

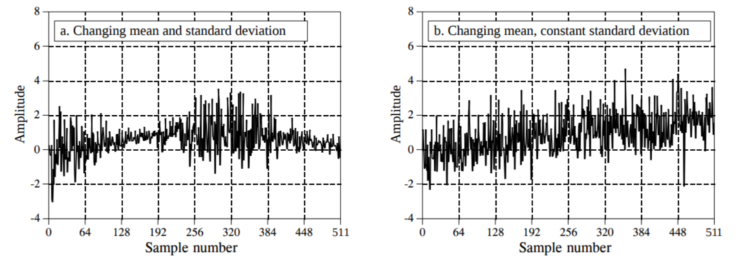

Figure 1: Signal processes in (a) and (b) are considered non-stationary because one or both of μ and σ depend on time.

Interactive Questions

Interactive Questions

Which of the following pairs of number have the same *stem*?

- A: 17 and 27
- B: 33 and 43
- C: 16 and 11
- D: -1 and 1

Interactive Questions

How many *leaves* are there for the stems, given the data set?
Data set: 67, 77, 72, 74, 90, 91, 94, 88, 82.
Stems: 6, 7, 8, and 9

- A: 1, 3, 3, 2
- B: 1, 3, 2, 3
- C: 1, 3, 3, 3
- D: 1, 2, 2, 3

Consider the following relative frequencies below. Is the corresponding histogram *normalized*?
Relative frequencies: 0.1, 0.1, 0.25, 0.1, 0.1, 0.05

- A: Yes
- B: No

Interactive Questions

What is the *mean* of the histogram data below?
Bins: 0, 2, 4, 6, 8, 10
Data: 10, 60, 20, 5, 1, 1

- A: 0.98
- B: 2.56
- A: 4.11
- B: 10

More on Histograms

Normalization - To convert all the frequencies to relative frequencies.

- Looking at fractions is helpful for *relative* questions about data. (Professor example).
- Makes calculating the mean simple, the idea of a *weighted average*. (Professor example).
- Summing a subset of bins is a *probability*, not a *count*. (Professor example).

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 2: A table of class sizes at Whittier College.

More on Histograms

1. In your notebook, create a normalized histogram of of the 50-99 column of Fig. 2.
2. For your histogram class size, what fraction of all classes of this size come from the years 2010-2014? What is the fraction that come from 2015 onwards?
3. What is the *mean* of the histogram?

More on Histograms

Two-dimensional histograms. There's no reason to restrict to one dimension… (Professor: draw a 2D histogram of Fig. 2 below).

How do you think about the *mean*?

More on Time-Series Data

More on Time-Series Data

We also think of the left-most column as *time slices*, and then we can frame the rest of the data as a time-series.

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 3: A table of class sizes at Whittier College.

Graphing time-series to look for trends. (Make a time-series of the class-size data below).

Locating the Center of the Data

1. **Median** - The data value that halves a sorted list of data
2. **Mode** - The data value with the highest frequency
3. **Mean** - The average using either definition
4. **Quartiles** - The values Q_i that divide a sorted list into quarters of equal frequency
5. **IQR** - $Q_3 - Q_1$
6. **k-th Percentile** - $i = k/100(n+1)$, where i is the index of the k-th percentile, and n is the number of data points in a sorted list
7. **Percentile of a value** - (next slide)

An algorithm for finding the percentile of a particular value:

- Order the data from smallest to largest.
- x = the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
- y = the number of data values equal to the data value for which you want to find the percentile.
- n = the total number of data.
- Calculate: $(x + 0.5y)/n \times 100$ and round to nearest integer.

The Spread of the Data

The *mean*, μ, and *standard deviation*, σ, of a data set $\{x_i\}$ are defined as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2}$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \mu)^2 \tag{3}$$

Octave commands:

```
x = randn(100,1);
mean(x)
std(x)
```

One nice theorem: *The variance is the average of the squares minus the square of the average.* Let $\langle x \rangle$ represent the average of the quantity or expression x. We have

$$\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2 \tag{4}$$

Proof: observe on board.

Note: There is a distinction between the *process or signal process* and the *the data*. Just because the data has a given μ and σ does not imply that the signal process has or will continue to have the exact same values of μ and σ. The underlying process could be *non-stationary*.

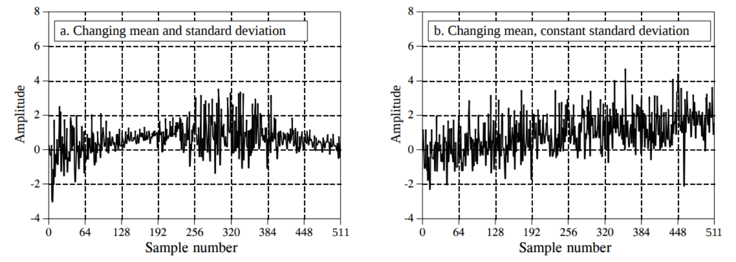

Figure 4: Signal processes in (a) and (b) are considered non-stationary because one or both of μ and σ depend on time.

A histogram is an object that represents the frequency[1] of particular values in a signal. For example, below is a histogram of 256,000 numbers drawn from a probability distribution:

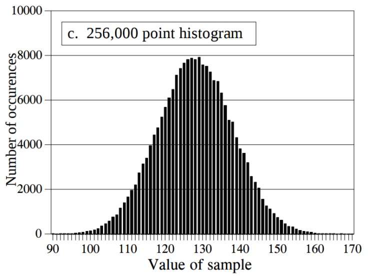

Figure 5: The histogram contains counts versus sample values.

[1]Careful: the word frequency refers to the number of occurences in the data, not a sinusoidal frequency.

Conclusion

Unit 0 Outline

1. Topics from Chapter 1: 1.1, 1.2, 1.3
    - What is a statistic?
    - Probability examples
    - Data and sampling
2. Topics from Chapter 2: 2.1 - 2.4, 2.5 - 2.8
    - Data visualization
    - Location of the data in numerical space
3. Topics from Chapter 3: 3.1, 3.2, 3.3
    - Two rules of probability