Liberian Mathematics Teacher Training Program 2023–2024

Andrew Obus[1]

CUNY Baruch College

October 13, 2023

[1]This program is partially supported by NSF CAREER Grant DMS-2047638

HW Exercise 1

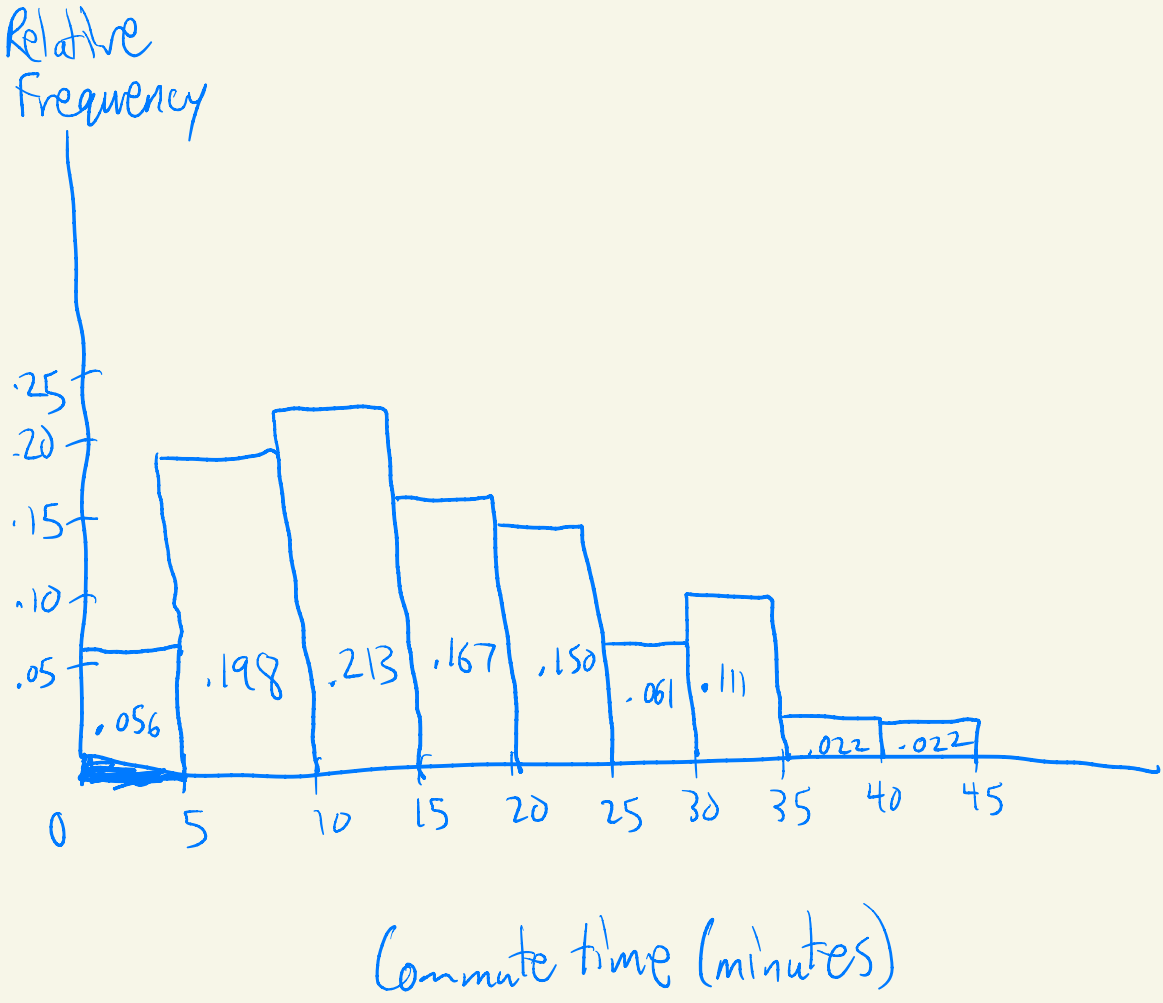
The following is a frequency distribution for commuting times to work (in minutes) for working adults in a county in California.

Time	Frequency	Relative Frequency.
0 to $<$ 5	5200	.056
5 to $<$ 10	18200	.198
10 to $<$ 15	19600	.213
15 to $<$ 20	15400	.167
20 to $<$ 25	13800	.150
25 to $<$ 30	5700	.061
30 to $<$ 35	10200	.111
35 to $<$ 40	2000	.022
40 to $<$ 45	2000	.022

(5200/92,100)

Add a column to this table for relative frequency, and then create a histogram for this data using relative frequency on the vertical axis.

↳ = absolute frequency / total number of data points = 92,100


Relative Frequency
.25
.20
.15
.10
.05
.056
.198
.213
.167
.150
.061
.111
.022
.022
0
5
10
15
20
25
30
35
40
45
Commute time (minutes)

HW Exercise 2

The following data represents total length of roads (in miles) in various subdivisions in the US: 1280, 360, 3350, 450, 1850, 3150, 5320, 3330, 540, 2250, 2460, 1890, 4390, 3380, 3870, 2320, 5850, 510, 2100, 340, 1250, 2400, 2700, 240, 1240, 1000, 2400, 3150, 2730, 3060. The data was used to help determine whether it would make more sense to build power lines above ground or underground.

Construct a stem and leaf plot for this data using the thousands digit as a stem. Then, contruct a histogram for the data using the classes 0 to $<$ 1000, 1000 to $<$ 2000, 2000 to $<$ 3000, and so on.
We will construct this histogram using Microsoft Excel.

stem	leaves
0	234555
1	023399
2	1 33 44 5 77
3	1 2 2 3 44 9
4	4
5	39

Histograms in Excel from raw data

- First put the data entries in one column.
- Then put the *maximum* values for each frequency class in another column.
- Click on “Data” from the top menu, and then “Data Analysis”
- Select “Histogram” from the resulting menu.
- You will be asked for a range of cells for the data, and another range of cells for the bins. Input these using the format “top cell : bottom cell”
- This will create a frequency table. If you want a visual histogram as well, check the box for “chart output”.

Cumulative relative frequency tables

- Suppose we have a relative frequency table for a set of numerical data, broken up into frequency classes (or "bins").
- The *cumulative relative frequency* of a given class is the percentage of the total data in that class *as well as in all lower classes*. → FORMULA: Measurements in all class below or equal to given class ÷ total # of measurements
- Cumulative relative frequency tables can be made by hand, or using Microsoft Excel.
- Additionally, a line graph, called a "cumulative relative frequency plot" can be added to the historgram to represent the cumulative frequencies.
- The cumulative relative frequency plot has a point for each class corresponding to its cumulative relative frequency. These points are connected by lines.

Example from HW

Let's look at the example from homework problem #2.

Road length	Frequency	Relative frequency	Cumulative relative frequency
0 to 1000	7	.233 $= 7/30$	.233
$>$ 1000 to 2000	5	.167 $= 5/30$	$.233 + .167 = .400$
$>$ 2000 to 3000	8	.267 $= 8/30$	$.400 + .267 = .667$
$>$ 3000 to 4000	7	.233 $= 7/30$	$.667 + .233 = .900$
$>$ 4000 to 5000	1	.033 $= 1/30$	$.900 + .033 = .933$
$>$ 5000	2	.067 $= 2/30$	$.933 + .067 = 1$

Total Measurements $= 7+5+8+7+1+2 = 30$

Percentiles

- Percentiles give a way to help understand a numerical data set.
- They tend to work better with larger data sets (e.g., height and weight percentiles for children of a given age).
- Roughly, the Nth percentile of the data set is the number below which N percent of the data lie.
- Turning this into a precise definition is done differently in different books, but all methods give approximately the same answer (and the more data there is, the better the answers approximate each other).

Percentile Calculation

- We will use what is called the "interpolation method"
- Suppose we have a data set with N entries, and we want to calculate the Pth percentile.
- First we arrange the data in ascending order.
- Then, we calculate the *rank* R of the Pth percentile. This is done with the formula

$$R = \frac{P}{100}(N+1).$$

- If R is an integer, the Pth percentile is simply the Rth data entry.
- If R is not an integer, we look at the $\lfloor R \rfloor$th and the $\lfloor R \rfloor + 1$st data entries, call these x and y. (e.g. if $R = 3.75$, look at 3rd and 4th data entries)
- Let F be the fractional part of R. Then the Pth percentile is

$$x + F(y - x).$$

(e.g. $F = .75$; "interpolation")

Example

Consider a class whose scores on a quiz (graded out of 10 points) are given by the following frequency table:

Score	Frequency
4	2
5	4
6	3
7	3
8	2
9	3
10	3

Let us calculate the 30th and the 50th percentile of this data.

30th percentile: 1) N = total # of data points $= 2+4+3+3+2+3+3 = 20$.

2) $R = \frac{30}{100}(20+1) = \frac{3}{10} \cdot 21 = 63/10 = 6.3$

6th entry, 7th entry

3) DATA: 4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 7, 8, 8, 9, 9, 9, 10, 10, 10 $\quad x=5,\ y=6$

4) Interpolation Formula: $x + F(y-x) = 5 + .3(6-5) = \boxed{5.3}$ ← 30th percentile.

50th percentile

$N = 20$

$R = \frac{50}{100}(20+1) = \frac{1}{2} \cdot 21 = \frac{21}{2} = 10.5$

10th 11th

DATA: ~~4~~, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 7, 8, 8, 9, 9, 9, 10, 10, 10

$x = 7, \; y = 7$

Interpolation:

50^{th} percentile $= 7 + .5(7 - 7)$

$= 7$

$7 - 7 = 0$

MEDIAN

Percentiles and cumulative relative frequency

- Percentiles are related to cumulative relative frequency.
- In particular, if the cumulative relative frequency of a class is p, then the $100p$th percentile of the data should be roughly the maximum of that class.
- Let's try this with the second example from the homework. We will consider the cumulative relative frequency .4, corresponding to the 40th percentile.
- Without doing any calculation, what do we expect to be a good approximation to the 40th percentile?

↳ Expect 40th percentile to be about 2000.

Will do calculation next time.

Homework Exercises

1. Using the frequency table from slide number 9 (the quiz scores), add columns for relative frequency and cumulative relative frequency.
2. Construct a histogram for this data, as well as a cumulative relative frequency plot.
3. Calculate the 60th and 70th percentiles for the data.

Thank you for your attention. Next week we will discuss measures of central tendency. Class will again be one hour early (at 3:30)!