Lecture 14 Worksheet

Chrysafis Vogiatzis

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Descriptive statistics

The average high September temperature in Chicago over the last 15 years (from 2006 to 2020) has been as in Table 1 (in Fahrenheit).

Year	Temperature
2020	74
2019	75
2018	70
2017	70
2016	72
2015	71
2014	66
2013	69
2012	66
2011	63
2010	65
2009	66
2008	66
2007	67
2006	69

Table 1: The average high September temperatures over the last 15 years.

Let's use this small dataset today to describe the information presented and visually showcase it. In the remaining exercises you will only use this small part of the data. Of course, in real life, we would be given a bigger volume of data. In such instances, we would resort to using software such as Excel or Pandas on Python.

Problem 1: Sample average and variance

What is the average temperature in the sample? What is the sample variance? [1]

Answer to Problem 1.

Problem 2: Sample mode

What is the mode of the data? [2]

Answer to Problem 2.

Problem 3: Quartiles

What are the first, second, and third quartile? [3]

Answer to Problem 3.

Problem 4: Range and interquartile range

What is the range and the interquartile range? Based on your answers: are there any outliers in the data?

Answer to Problem 4.

[1] Recall how the calculation of variance is different for a sample compared to a population...

[2] The mode is the most frequently observed data point.

[3] The second quartile is also called the median.

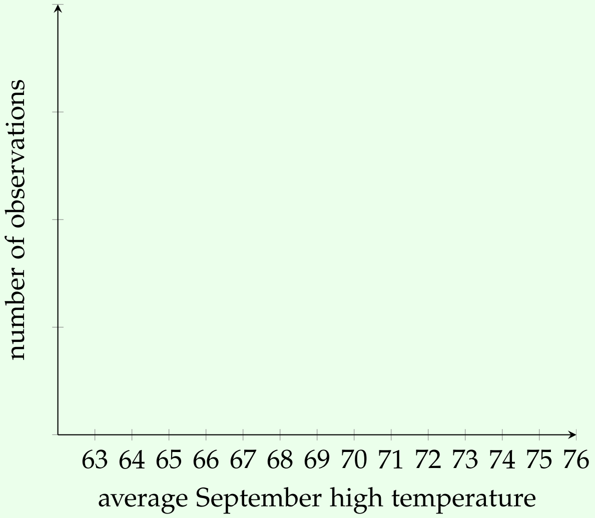
Activity 2: Graphical representations

In this activity, we will present the data given in a series of graphical tools. We omit some of them due to time constraints (for a full description, please read the Lecture 14 notes).

Problem 5: Dot diagrams

Dot diagrams [4] (as the name suggests) asks to place a dot on top of each data point. For example, say there are 5 observations for a specific data point, we would put 5 dots one on top of the other! Create a dot diagram for the data of Table 1.

Answer to Problem 5.



Note how easy it is to find the mode now! Simply look for the tallest set of dots.

Problem 6: Stem-and-leaf diagrams

A stem-and-leaf diagram [5] only makes sense when all of the data consists of at least two digits. How to construct one? We need to separate a numerical observation into a stem (the first, more important digits) and leaves (the least important digit). For example, the temperature of 32 Fahrenheit can be decomposed into a stem of "3" and a leaf of "2", or the number 538 can be decomposed into a stem of "53" and a leaf of "8".

Just for this example, consider the average low temperature in two different states:

[4] See Page 8-9 in the notes!

[5] See Page 9 in the notes!

State	J	F	M	A	M	J	J	A	S	O	N	D
State 1	15	19	33	41	51	61	65	63	54	43	33	20
State 2	36	40	47	53	62	68	71	71	65	60	51	40

Create a stem-and-leaf diagrams using the data provided for State 2. Observe my stem-and-leaf for State 1 as an example!

Answer to Problem 6.

State 1

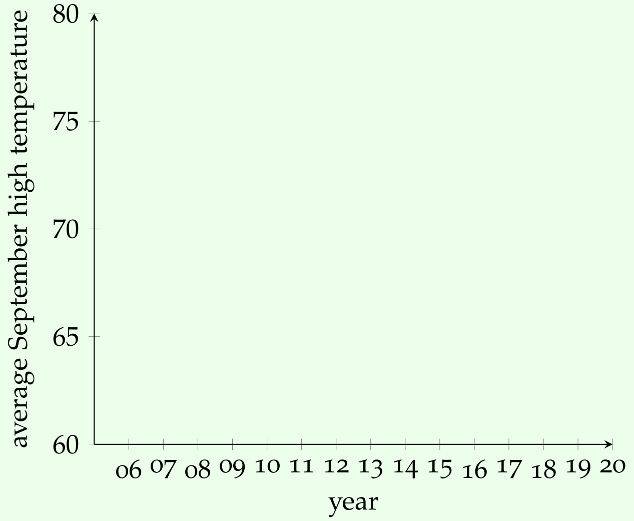
6	1 3 5
5	1 4
4	1 3
3	3 3
2	0
1	5 9

It is a *striking* visual tool to showcase frequency.

Problem 7: Time series plots

A time series plot [6] is especially useful when the data are recorded in the order of time. For example, in our original dataset of Table 1, all temperatures are given in order of time from 2006 to 2020. Create a time series plot by adding each observation (y coordinate) in the appropriate time (x coordinate), and then connect the observations using straight lines.

Answer to Problem 7.



[6] See Pages 11-12 in the notes.

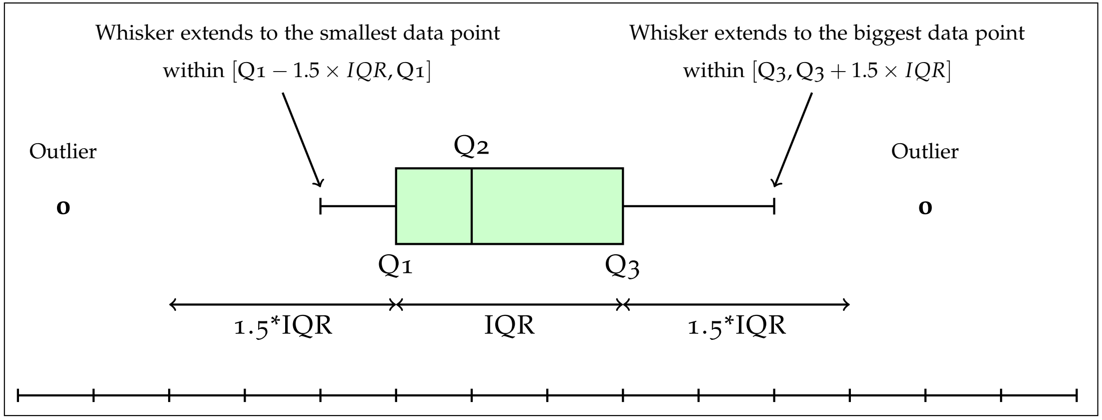
Activity 3: Box plots

Box plots [7], sometimes also called box-and-whisker plots, are graphical devices built to reveal multiple interesting properties at once. Seeing a box plot reveals the center, the spread, the shape, and the outliers in our data!

To build one, follow the next few steps:

1. Identify Q_1, Q_2, Q_3 and the IQR. Create a small box with three lines: the left most line is at Q_1, the middle one is at Q_2, and the right most one is at Q_3.
2. Calculate $Q_1 - 1.5 \cdot IQR$ and $Q_3 + 1.5 \cdot IQR$. Points that are outside these limits are *outliers*. Mark every outlier with a "o".
3. Extend the one whisker all the way to the smallest observation within $[Q_1 - 1.5 \cdot IQR, Q_1]$ and the other whisker all the way to the largest observation within $[Q_3, Q_3 + 1.5 \cdot IQR]$.

For convenience, we present here the same figure from the notes, where the details of a box plot are explained visually.


Whisker extends to the smallest data point
within [Q1 − 1.5 × IQR, Q1]
Whisker extends to the biggest data point
within [Q3, Q3 + 1.5 × IQR]
Outlier
o
Q2
Outlier
o
Q1
Q3
1.5*IQR
IQR
1.5*IQR

Problem 8: Designing a box plot

Design a box plot based on the data of Table 1.

Answer to Problem 8.

[7] See Pages 12-13 in the notes!

Problem 9: Missing whiskers?

The compressive strength of concrete is the subject of a test by civil engineers. Nine different specimens were tested and the civil engineers obtained (in psi): 2210, 2230, 2200, 2240, 2250, 2240, 2330, 2250, 2210. Describe the data as a **box plot**.

Answer to Problem 9.

Note then how when there are no points in $[Q_1 - 1.5IQR, Q1]$ or $[Q_3, Q_3 + 1.5IQR]$ then the whisker simply disappears!

Activity 4: Histograms

A histogram [8] is a graphical construct that presents data by placing them in *bins*. Histograms possess three important characteristics:

1. modality.

2. heavy/light tailedness.

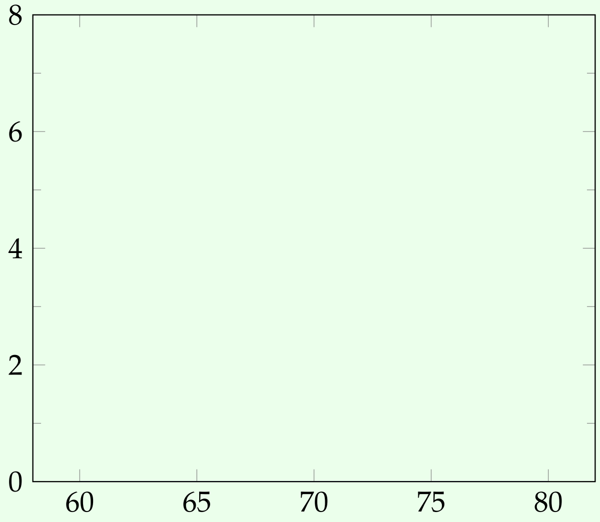
3. skewness.

We will investigate all three of them in the next few activities.

Problem 10: Designing a histogram

First of all, we may create different histograms depending on the number of bins and the type of bins we create. Assume we create 4 bins corresponding to temperatures $[60, 65)$,$[65, 70)$,$[70, 75)$,$[75, 80)$. Design this histogram.

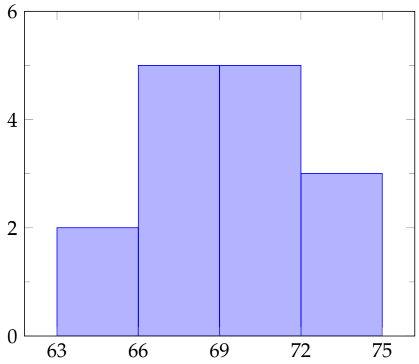
Answer to Problem 10.



[8] See Pages 13-19 in the notes!

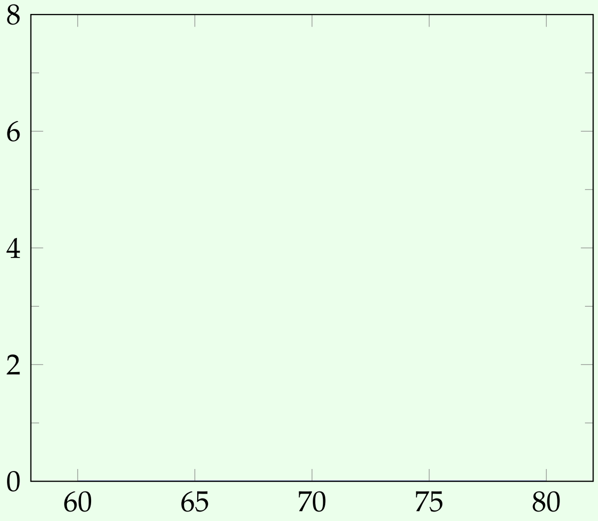
Problem 11: Different histograms

What if we change the number of bins? Assume we want to generate n bins. What we could do, is find the range of values and divide this by n: call this (fractional, probably) number q. Then, create bins as follows: $[min, min + q), [min + q, min + 2q), \ldots, [max - q, max]$. For example, if we created $n = 4$ bins with our data, we would have $q = 12/4 = 3$ and the following bins: $[63, 66), [66, 69), [69, 72), [72, 75]$. The corresponding histogram would be:



Draw a histogram with $n = 6$ bins.

Answer to Problem 11.



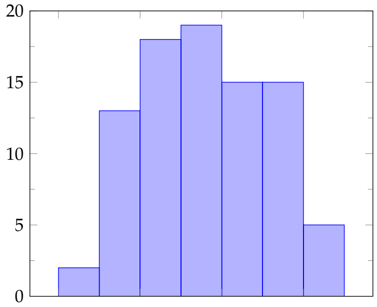
Problem 12: Histogram details

What do you observe about your two histograms (in Problems 10 and 11)? What are their modalities (i.e., number of discernible peaks)? Are they heavy-tailed? Are they left or right skewed? [9] After you answer these for your histograms, then match the following histograms (a, b, c, and d) with the histogram below.

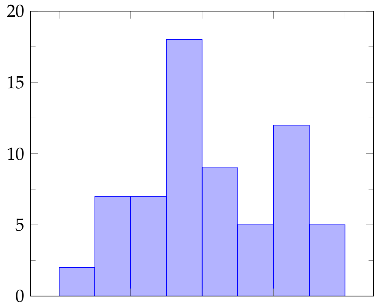
Which of the following four histograms:

1. are unimodal?
2. are bimodal?
3. are uniform?
4. are right-skewed?
5. are left-skewed?
6. are symmetric?

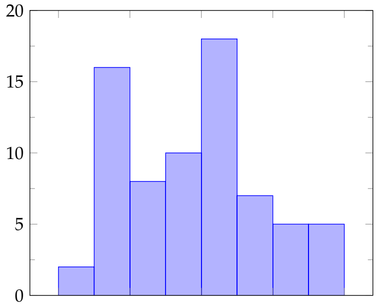
(a)



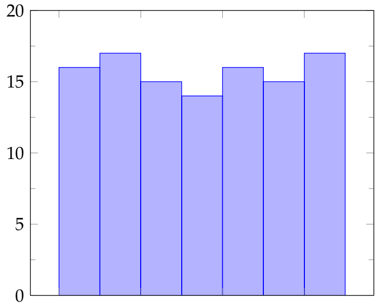
(b)



(c)



(d)



Answer to Problem 12.

- **unimodal**:
- **bimodal**:
- **uniform**:
- **right-skewed**:
- **left-skewed**:
- **symmetric**:

[9] For example, the histogram with $n = 4$ bins is **unimodal** (one discernible peaks, one around 66–72), heavy-tailed, and it appears to be right-skewed (i.e., with a tail to the right).

Activity 5: Congressional districts – a study of PVI

This is for **extra credit** only. You do not need to submit your answers for this at this point, unless you'd like to practice your Python and your `pandas` knowledge.

For this question, you will need to use a **usdistricts.csv file** that is available under **Worksheet 14 – extra credit** on Canvas. This file contains all of the congressional districts of the United States, complete with their populations and voting indices (PVI). Search for PVI online: it is an interesting index that measures how much more Republican or Democrat a district is, compared to the national average.

To successfully complete this part, perform the following using `pandas` in Python.

(a) Construct the box plot for the district populations over 18 for all of the United States.
(b) Construct the histograms for the district populations over 18 for all of the US. Report your histograms with 10 bins, 25 bins, 50 bins, and 100 bins.
(c) Construct the box plots of the PVI parameter for the states of CA, NY, TX, OH, and WV.
(d) Tell me a story on the data: compare the PVI box plots of WV and TX. Compare the PVI box plots of WV and OH. What are the similarities and the differences between them? What can we infer by looking at them? What else did you find by analyzing the data?

Send me your code, your answers, and your thoughts/stories through email at chrys@illinois.edu by the Exam 2 date (at around 10/21).

Good luck!