Homework 1

PSTAT 5A: Spring 2023, with Ethan P. Marzban

Instructions

- Please submit your work to Gradescope by no later than **11:59pm on Tuesday, April 11**. As a reminder, late homework will not be accepted.
- Recall that you will be asked to upload a **single** PDF containing your work for *both* the programming and non-programming questions to Gradescope.
  - You can merge PDF files using either Adobe Acrobat, or using adobe's online PDF merger at this link.

Problem 1: Data Classification

Part (a): Classify each of the following variables as discrete, continuous, ordinal, or nominal. Justify your answers.

i. The number of accidents occurring between noon and 1pm at a particular traffic light, over the span of 100 days.

ii. The amount of soda (in liters) consumed daily by 247 randomly selected UCSB students.

iii. The zip codes of 1,000 randomly selected households in California.

Part (b)

i. Give an example of a continuous variable. Pick one that wasn't covered in lecture or part (a) of this problem.

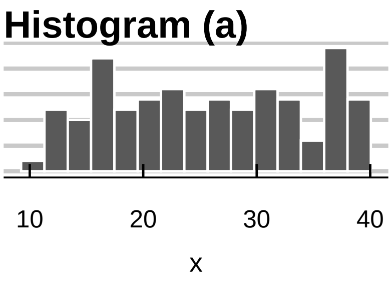
ii. Give an example of a discrete variable. Pick one that wasn't covered in lecture or part (a) of this problem.

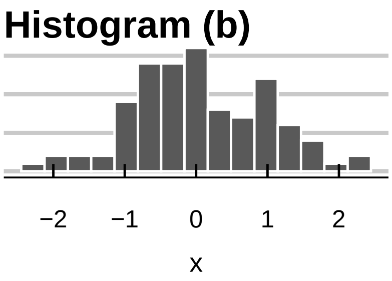
iii. Give an example of a nominal variable. Pick one that wasn't covered in lecture or part (a) of this problem.

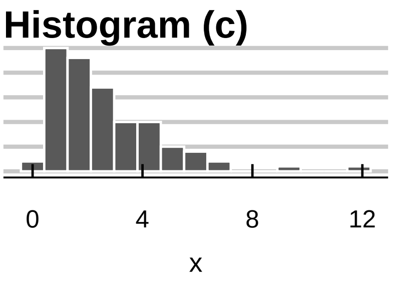
iv. Give an example of an ordinal variable. Pick one that wasn't covered in lecture or part (a) of this problem.

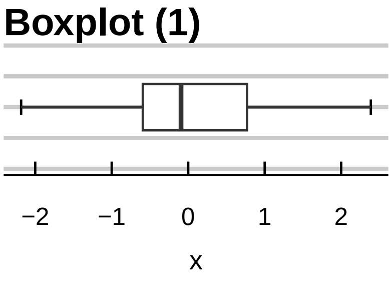
Problem 2: Matching

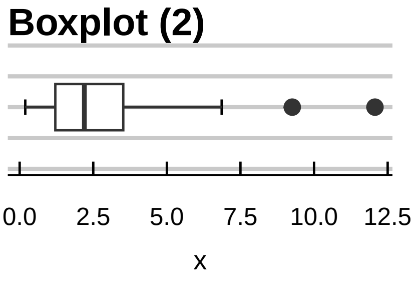
Below you will see 3 histograms, labeled (a) through (c), and 3 boxplot, labeled (1) through (3). Match each histogram to its corresponding boxplot, and justify your answers. **Hint:** In each case, think about both spread as well as how the mean compares with the median.

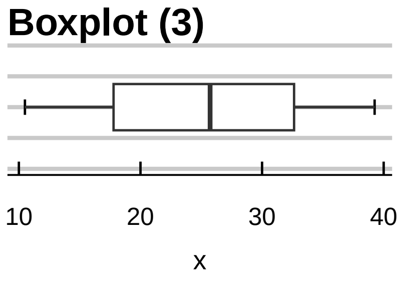












Problem 3: Appropriate Visualizations

In the parts below, you will be provided with a pair of variables x and y. For each part, identify the appropriate type or types of visualization for x, along with the appropriate type of plot to visualize the relationship between x and y.

a. x is monthly salary; y is zip code.

b. x is average commute time (in hours, including decimals); y is amount of sleep (also in hours, including decimals)

c. x is month of the year, y is amount of ramen (in pounds) consumed.

Problem 4: Associations

a. Give an example of two variables you think would have a positive association.

b. Give an example of two variables you think would have a negative association.

Problem 5: Transformations

The notion of **transforming data** is an incredibly important one. As an example, suppose $F = \{f_i\}_{i=1}^{n}$ denotes a set of temperature measurements, as recorded in Fahrenheit. If we wish to convert the measurements to Centigrade, we obtain a new set of data $C = \{c_i\}_{i=1}^{n}$ where each c_i is linked with a corresponding f_i through the formula

$$c_i = \frac{5}{9}(f_i - 32)$$

In general, we consider a set $X = \{x_i\}_{i=1}^n$ and a function $g : \mathbb{R} \to \mathbb{R}$, and construct a new set of data as $Y = \{y_i\}_{i=1}^n$ with $y_i = g(x_i)$. This is all we mean by a **transformation:** a function that we apply to each point in a dataset to obtain a new dataset.

a. Suppose we take a linear transformation g; i.e. we take $y_i = ax_i + b$ for some fixed constants a and b. (A concrete example of such a transformation is the conversion from Fahrenheit to Centigrade mentioned above). Show that $\overline{y} = g(\overline{x})$, using the following steps:
   i. Write down the definition of $\overline{y}$
   ii. Substitute $ax_i + b$ in place of y_i
   iii. Perform algebraic manipulations to obtain $\overline{y} = a\overline{x} + b = g(\overline{x})$ to complete the argument.

b. Is it always true that $\overline{y} = g(\overline{x})$? If so, provide a short proof/justification. If not, give a specific counterexample. As a hint: think about nonlinear transformations as well!

Problem 6: The Median

There is another measure of central tendency: the **median**. Here's how we compute the median of a set $X = \{x_i\}_{i=1}^n$:

i. Line up the numbers in ascending order
ii. Cross off the first and last numbers.
iii. Cross off the first and last numbers that are not crossed off.
iv. Continue until you are either left with a single number (in which case this number is the median), or we are left with a pair of numbers (in which case the median will be the mean of these two numbers).

As an example, to compute the median of the set $S = \{1, 2, 3, 3, 5, 6, 10, 11\}$: we write:

- $\cancel{1},\ 2,\ 3,\ 3,\ 5,\ 6,\ 10,\ \cancel{11}$
- $\cancel{1},\ \cancel{2},\ 3,\ 4,\ 5,\ 6,\ \cancel{10},\ \cancel{11}$
- $\cancel{1},\ \cancel{2},\ \cancel{3},\ 3,\ 5,\ \cancel{6},\ \cancel{10},\ \cancel{11}$

So, the median is $(3 + 5)/2 = \boxed{4}$.

a. Compute the median of the set $\{1, 2, 3, 4\}$.

b. Compute the median of the set $\{1, 2, 3, 4, 5\}$.

c. Compute the median of the set $\{1, 2, 3, 4, 5, 6\}$.

d. Generalize your answers to parts (a) - (c) above to find a formula for the median of the set $\{1, 2, \cdots, n\}$, for some fixed natural number n.

e. It turns out (as a result from mathematics) that

$$\sum_{k=1}^{n} k = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

Use this fact to compute the mean of the set $\{1, 2, \cdots, n\}$. How does this compare to the median you computed in part (d) above?

Problem 7: Programming

Instructions

- Write your answers to this question in a new Jupyter notebook, and export your work to a PDF using the steps you saw in Lab01. **Be sure to merge this PDF with your PDF containing your work to the above questions before submitting!** (See instructions at the top of this homework).

a. Write a line of code that results in a `NameError`. Don't use the same code from `Lab01`!

b. Run the following code in a new cell:

```
x = 2
y = 3
x = y + 2
```

Using a comment, write down what you think `print(x)` will return after having run the above code. Then, run `print(x)` and comment on whether your initial guess was correct or not.

c. Use Google to find out where the name `Jupyter` comes from. Write your answer in a Markdown cell.

d. Create a new cell, copy the code `type(abs)`, and comment on the result. Specifically, based on the result of this code, can you identify yet another data type in addition to `float` and `int`?

e. Navigate to this (https://docs.python.org/3/library/functions.html) link, which contains a list (and description) of the functions that come built into Python (i.e. that can be used without needing to import any modules). Pick **two** functions to read up on, and write a brief description of both functions as well as an example call of each. For example:

- `abs()`: Computes the absolute value of a number. Example call:

```
abs(-103.203)
```

```
103.203
```

(Don't use `abs()` as one of the functions you choose!)