Econ 300 Homework 1

Note: for book exercises, solutions are in appendix A. Exercises 2.25 and 2.26 are good holistic examples, but a bit difficult for this course. The “chapter exercises” for chapter 2 are similarly good but somewhat difficult questions.

General data literacy (1.2.2)

The following book questions are examples of categorization of variables

1.3c, 1.4c, 1.5c, 1.6c, 1.10c, 1.11b 1.12b

1.4c

There is a variable that records whether they were in the treatment or control group, which is nominal. Quality of life, activity, asthma symptoms, and medication reduction are all ordinal variables. They’re on a numeric scale, but here there is no direct meaning to the scale, only that higher is more extreme.

1.6c

The number of candies they were reported to take is a standard numeric variable. It is discrete since the number they took is a counting number. They may have also recorded a separate variable of actual number of candies taken based on the research question, in which case the variable type would be identical. Socioeconomic status (SES) was identified via three survey questions, each comparing to others based on various characteristics. Because these are comparisons, they are ordinal variables, though it is unclear what the levels are from the actual description.

1.10c

Sex is nominal. age is numeric (it’s discrete since they round to the nearest integer). marital status is nominal. Gross income is ordinal since it’s binned into discrete categories (this is sometimes called censored data). Smoking is nominal. Amount is numeric - it’s theoretically a continuous variable, but is likely recorded as discrete (the NA values do not change the the type).

1.12b

Outcome variables are arms control/disarmament, colonialism, economic development, human rights, nuclear weapons and materials, and palestian conflict. These are all binary variables (nominal). Country is a nominal variable used for graphing, and year is a numeric variable used to create the time series.

Quantiles and Boxplots (2.1.5)

Find the median, first quartile, third quartile, and interquartile range of the following list

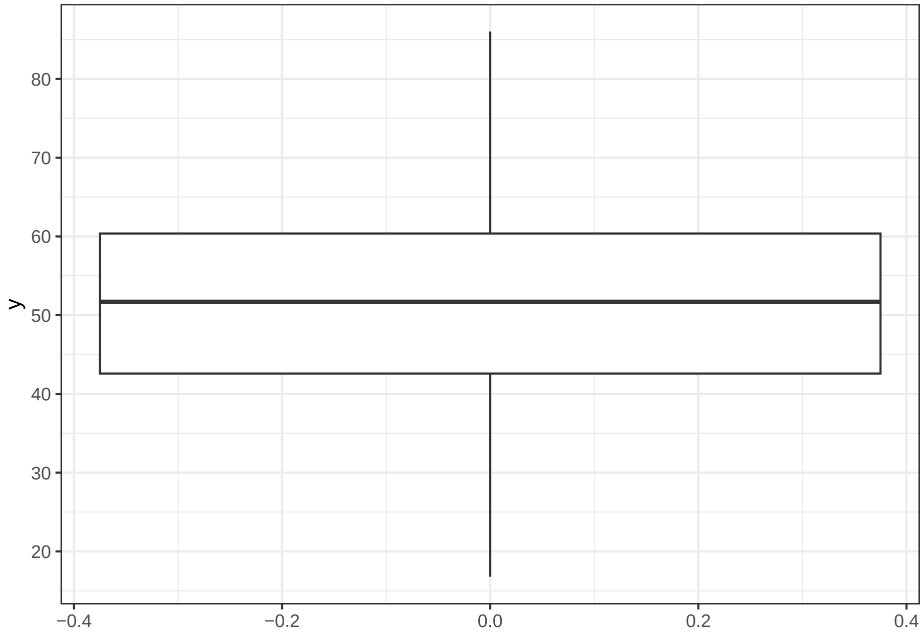
```
12 3 7 1 6 5 3 8 8
```

Answer

```
[1]  1  3  3  5  6  7  8  8 12
```

From this we see the median is 6 (4 numbers are below and 4 are above), Q1 is 3 (2 below, 6 above), and Q3 is 8 (2 above, 6 below). There are different methods of computing this, but here they'll yield the same answer

Find the median, first quartile, third quartile, and interquartile range of the following box plot



Answer

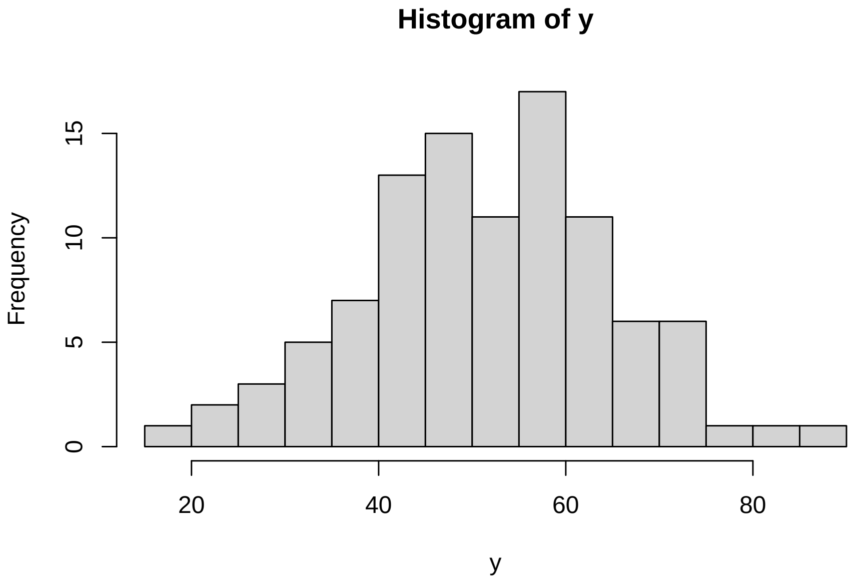
Approximate answers are fine. Exact answers end up being 42.59, 51.71, and 60.37 for Q1, the median, and Q3, respectively. The IQR is then $Q_3 - Q_1 = 17.78$

Book Problems: 2.8

a: These have the same median and IQR b: (2) has a higher median and IQR c: (2) has a higher median, but the same IQR d: (2) has a higher median and IQR

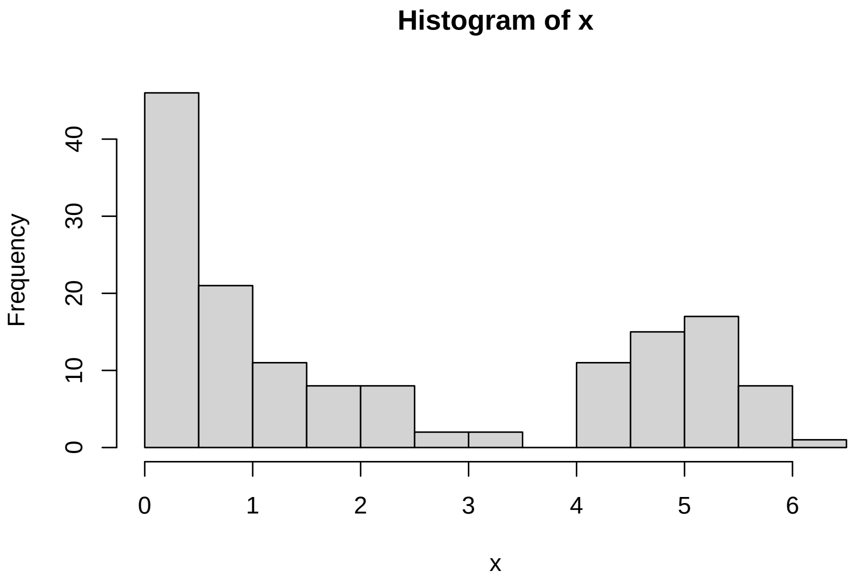
Histograms (2.1.3)

The following histogram has 100 total values. What percent of the data is between 30 and 40?



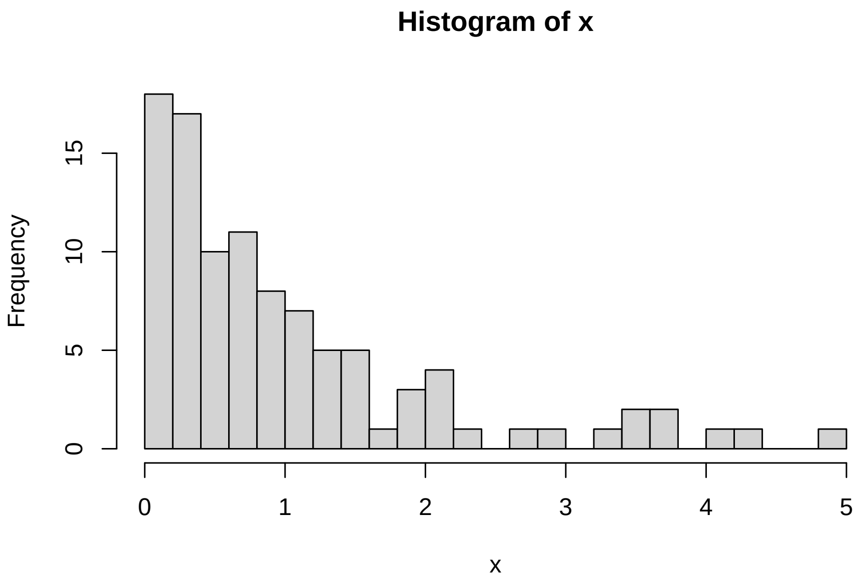
Exact answer is 12%. The height of the two bars immediately to the left of 40 are the answer: 5+7=12, and $\frac{12}{100} = 12\%$

What features are present in the following histogram?



This is both skewed right and bimodal

For the following distribution, which is larger, the mean or the median?



The mean is greater than the median. For a symmetric distribution, the mean and the median are equal. The right tail on this distribution strongly increases the mean but has little effect on the median. For this reason, the mean ends up getting dragged upward more than the median, leading to a larger value.

Book Problems: 2.15, 2.16 2.14 (this one pieces together information from multiple sections)

2.16

a: this is clearly right skewed. The median is more representative of a typical house. IQR is a better measure of variability for the 'typical' population since it is less influenced by the multimillion dollar outliers.

b: This is approximately symmetric (possibly a slight right skew from the description). Either the mean/median or IQR/standard deviation is fine here.

c: This will be right skewed since most of the density is at 0. The median is more representative of the typical student. Standard deviation is probably the better measure here, as IQR may very well be 0 in this dataset, which would imply less variability than is actually present

d: This is clearly right skewed. Median and IQR will better represent the typical employee and their variability.

2.14

Since the mean is greater than the median, this implies a right skew.

Mean (2.1.2)

Calculate the mean of the following data, where p is the proportion of data associated with each value

```
   x    p
1: 1 0.10
```

```
2: 2 0.20
3: 3 0.50
4: 4 0.15
5: 5 0.15
```

$.1(1)+2(.2)+3(.5)+4(.15)+5(.15)=3.35$

Suppose that an outlier is added to a dataset. Which is more affected, the mean or the median?

The mean, since the median only cares about rank, not size

Book Problems: 2.9

Variance (2.1.4)

Calculate the (population) standard deviation of the following data

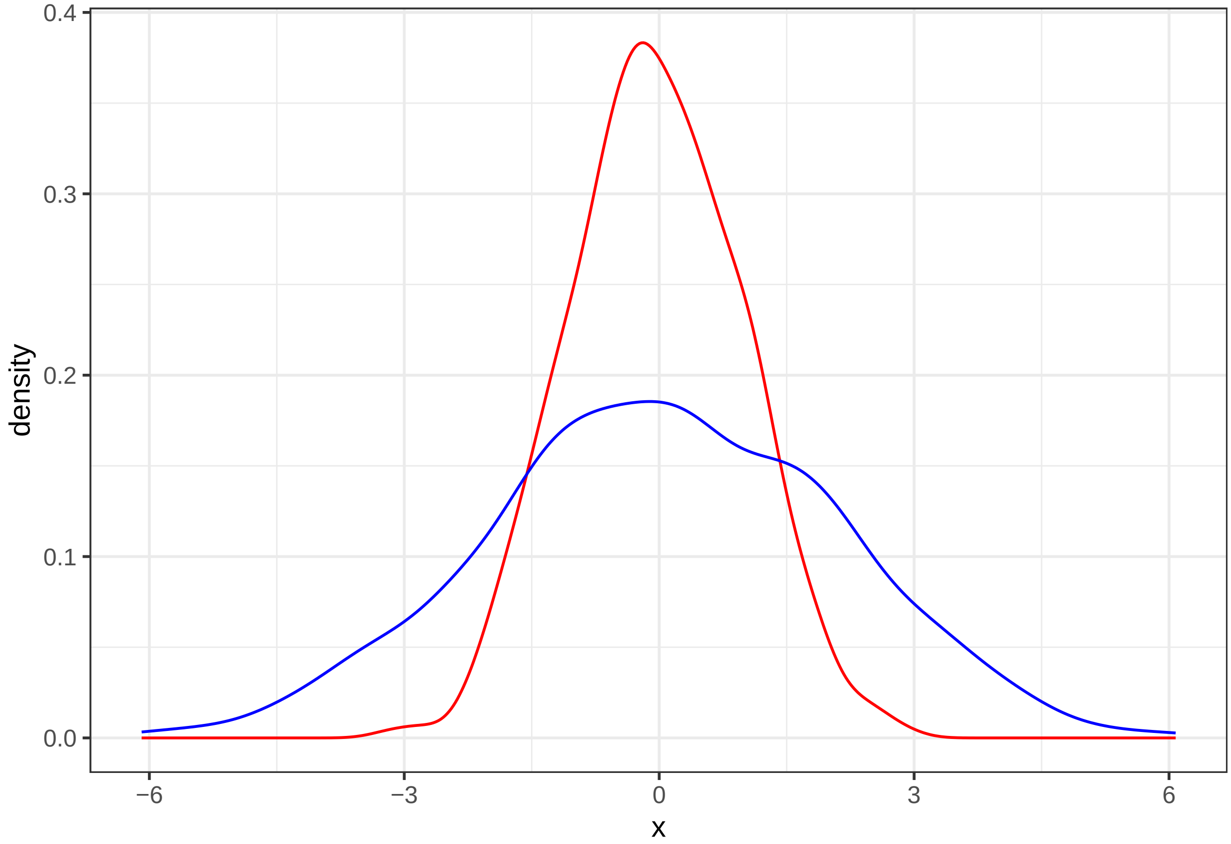
```
5 10 12
```

The mean is $\frac{5+10+12}{3}=9$. The variance is then $\frac{(5-9)^2+(10-9)^2+(12-9)^2}{3}=\frac{16+1+9}{3}=\frac{26}{3}$. Standard deviation is then the square root of this $\sigma=\sqrt{\frac{26}{3}}\approx 2.94$

The ACT is a standardized test with mean 18 and standard deviation 6. You score a 27 on the ACT. Calculate your z-score, i.e. your standardized test score

$\frac{27-18}{6}=1.5$

Which of the following distributions has a larger standard deviation?



The blue one has a larger stnadard deviation since it is wider

Contingency Tables (2.2.1)

The following (made-up) table gives the joint distribution of econ majors (vs non-econ majors) and whether they are a dog or a cat person.

- What percent of people prefer dogs?
- What percent of people are non-econ majors who prefer cats?
- Of people who are econ majors, what percent prefer dogs?
- Of people who prefer dogs, what percent are econ majors?

```
      Dog Cat Total
Econ   25  20    45
Other  15  40    55
Total  40  60   100
```

Answers

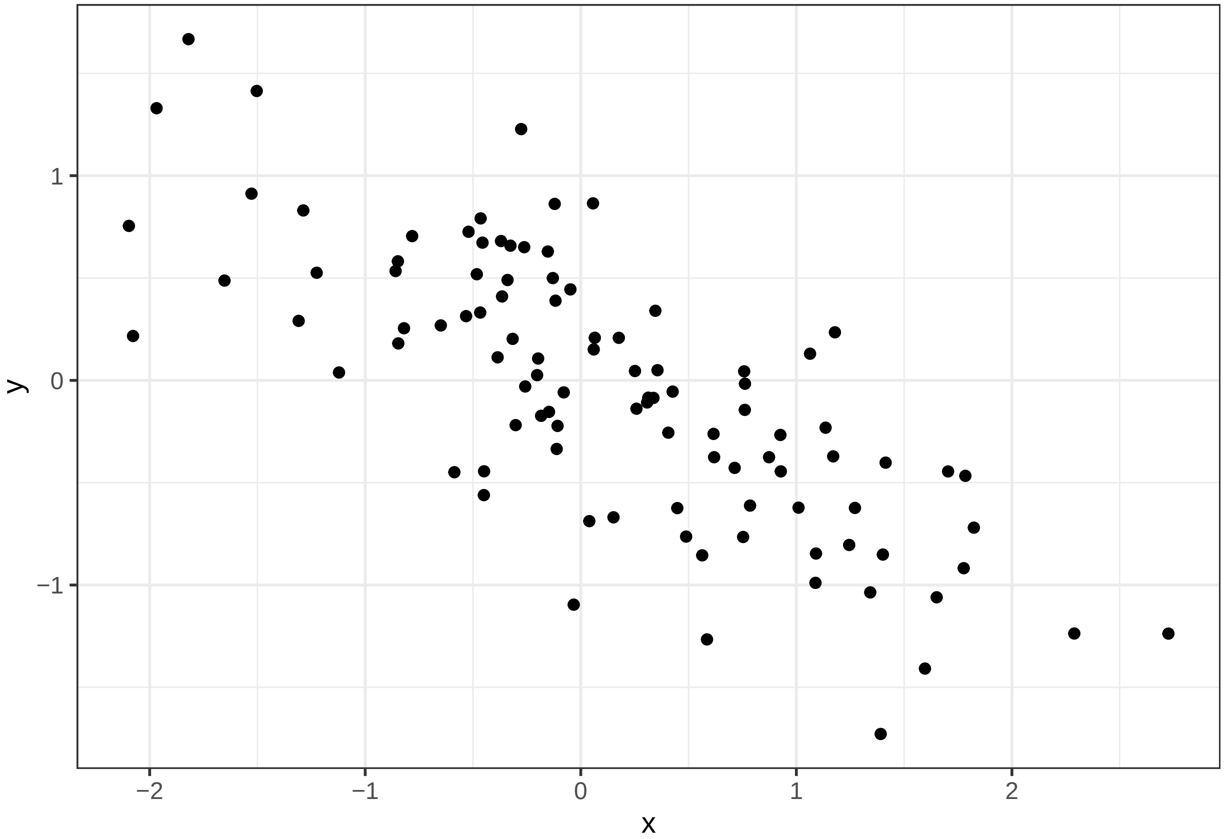
a: There are 40 people in the dog column out of 100 total, or 40% b: This is the cell corresponding to (other, cat), or 40 out of 100; 40% c: Looking across the Econ row, 25/45 prefer dogs, or around 55% d: Looking down the Dog column, 25/40 are econ majors, or 62.5%

Book Problems: 2.22 (e is difficult)

a: 372/910=40.9% b: 278/910=30.5% c: 57/910=6.3% d: 57/372=15.3%; 120/363=33.1%; 57.7% e: Answer d shows that these are not independent. If these views were independent, then the conditional probability of being in favor of citizenship would be (approximately) the same for conservatives, moderates, and liberals. We see above, however, that these numbers are very different.

Scatterplots (2.1.1)

Is the correlation between the following variables positive, negative, or approximately 0?



It’s negative. The exact value is -0.75

Suppose $r = -0.9$. What percent of the variation in y is explained by x?

r^2 gives the percent of the variation in y that is explained by x. $(-0.9)^2 = 0.81$, so this is 81%.

Book problems: 2.1, 2.2, 2.4 (positive vs negative association)

2.2

Positive, none, positive (nonlinear), negative

Graphical Comparison (2.2.6)

Book problems: 2.10, 2.13