Distributions

The distribution of a variable tells us what values the variable takes and how often it takes this values

We often represent distributions graphically as well as numerically

What do they tell us?

The distribution of a variable provides us with quantitative and qualitative info that allows us to make inferences/predictions

What values of the variable occur most frequently? Least frequently?

What are the “middle” and average values?

How spread out is the data? Does it tend to cluster around any specific values?

Are there any “outliers” or unusual/rare values?

Distribution of a categorical variable

The distribution of a categorical variable lists the possible categories (ex. Values of the variable) and gives either a count/percent of the individuals that fall in each category

We can represent the distribution of a single categorical variable graphically by using either a bar chart/pie chart

Bar charts- better for counts; pie charts- better for percentages

Comparative bar chart- comparison of bar charts

Bar charts are used to visualize the “distribution” of categorical/qualitative data

Histograms are similar to bar charts, but they are used to plot the distribution of quantitative data

Data: 9, 10, 22, 25, 32, 33, 39, 39, 42, 50, 58, 70

We will divide the # line into classes (bins)

We will use class (bin) widths of 10 starting with age 0, so our classes will be:

0 to 10, 10 to 20, 20 to 30 etc.

We will use the convention that:

A data point X is in the first class (0 to 10) if Z is greater than/equal to 0 but strictly less than 10

A data point X is in the second class (10 to 20) if X is greater than/equal to 10 but strictly less than 20 etc.

X = age

Class | Frequency | Relative frequency

0 < X < 10 | 1 | 1/13

\*when the frequency is 0 just put a “0” height bar

Frequency = total count

Relative frequency = proportion

What percentage of people in the sample are less than 30 years old? Between 30 and 40?

30.77%. 30.77%.

If a person is randomly selected from this sample, what is the probability he/she is at least 40?

38.46%.

Symmetric, unimodal

Uniform

Skewed right, unimodal

Skewed left, unimodal

Bimodal

Multimodal

\*if unsure of distribution being symmetrical/skewed, look at the end values

Larger widths —> fewer details, data too summarized

Smaller widths —> too may details, data not summarized enough

25.

6 bins of width 6 each

x = length of time (in months) between onset of illness

1. Bin | Rel Freq.

0<x<6 | 24/50

6<x<12. | 13/50

12<x<18| 5/50

18<x<24| 5/50

24<x<30| 2/50

30<x<36| 1/50

B) skewed right

C) 4/50

How to make a stem plot

* Separate each data pint into a stem and leaf. The last digit of the data point will be the leaf. The digits in front determine the stem.
* Write the stems in a vertical column and draw a vertical line to the right of the stems.
* Write each leaf in the row to the right of its stem in increasing order from the stem.

Stem- numerical base

Stem-and-leaf plot (basically a numerical histogram turned on its side)

0 | 9

1 | 0

2 | 2, 5

3 | 2, 3, 9, 9

4 | 2, 9

5 | 0, 8

6 |

7 | 0

24a) draw a stem and leaf plot for the data. Describe the shape of the distribution and look for outliers.

Stem-and-leaf plot for weekend gross sales of movies for 8/25/17-8/28/17 (in millions of $)

0 | 9

1 | 2, 4, 5, 5, 7, 7, 8

2 | 3, 4, 5, 6, 7, 8

3 | 9

4 | 2, 6, 7

5 |

6 |

7 | 7

8 |

9 |

10| 3

Skewed right

Outlier