Class 19

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-10-16

Chapter 2 Objectives: Numerical Data

- Describe the “shape” (i.e. distribution) of numerical variables
- Calculate means, medians, modes, variances, standard deviations, IQRs
- Learn the appropriate use of summary statistics (i.e. mean vs. median)
- Characterize the relationship between 2 numerical variables

Chapter 2 Objectives: Categorical Data

- Analyze contingency (e.g. 2x2) tables
- Summarizing categorical variables with proportions
- Comparison of numerical data between categorical groups

Chapter 2 Objectives: Visualizing Data

- Recognize common visualization techniques / plots
  - Numerical: Dot plots, histograms, density plots, box plots, violin plots
  - Categorical: bar plots, mosaic plots, tree map
- Build basic visualizations in `R` using `ggplot2`

Distribution Checklist

- Modality
- Symmetry
- Skew
- Outliers
- Summary Statistics

Modality

What is the modality of the distribution?

Modality

What is the modality of the distribution?

- ***Unimodal***: one peak
- ***Bimodal***: two peaks
- ***Multimodal***: many peaks
- ***Uniform***: no clear peak, flat distribution

Symmetry

Is the distribution symmetric or asymmetric?

Symmetry

Is the distribution symmetric or asymmetric?

- ***Symmetric***: "mirror image", the distribution to the left of center looks like the distribution to the right of center
- ***Asymmetric***: left half looks different than the right half

Skew

If the distribution is asymmetric, is it because it's skewed?

Skew

If the distribution is asymmetric, is it because it's skewed?

- Does the distribution "lean" towards the left or the right?
- Does the distribution have a long "tail" on one side but not the other?

Outliers

Are there outliers in this distribution?

Outliers

Are there outliers in this distribution?

- Are there any unusual data points?
- How extreme are the most extreme values?
- Outliers are *rare*
- When data points are unusual but not rare, they create *skew* or *modality*

Summary Statistics

Is the distribution normal or does it require robust statistics?

Summary Statistics

Is the distribution normal or does it require robust statistics?

- When the distribution is very close to normal, the mean + SD will describe the center ~70% of the data
- The mean + SD are sensitive to modality, asymmetry, skew, and outliers
- It's never wrong to use the median + IQR, but when the distribution IS normal, the mean + SD are better

Robust Statistics

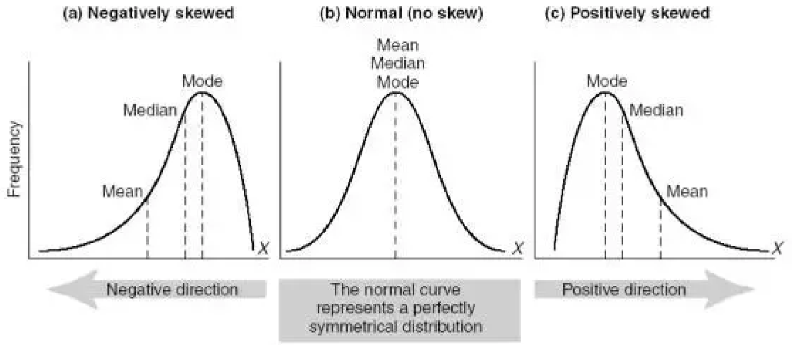


FIGURE 15.6 Examples of normal and skewed distributions

Figure 1: The ***median*** and ***interquartile range*** are considered to be ***robust statistics*** for the numerical summary of data because they are less sensitive to ***skew*** and ***outliers*** than the ***mean*** and ***standard deviation***.

5-Number Summary of Numerical Data

1. Minimum value or Q1 - 1.5 x Interquartile Region
2. 1st quartile (Q1, 25th percentile)
3. Median (Q2, 50th percentile)
4. 3rd quartile (Q3, 75th percentile)
5. Maximum value or Q3 + 1.5 x Interquartile Region

Anatomy of a Boxplot

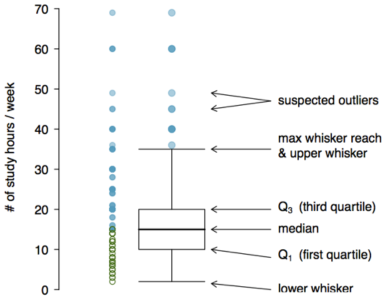


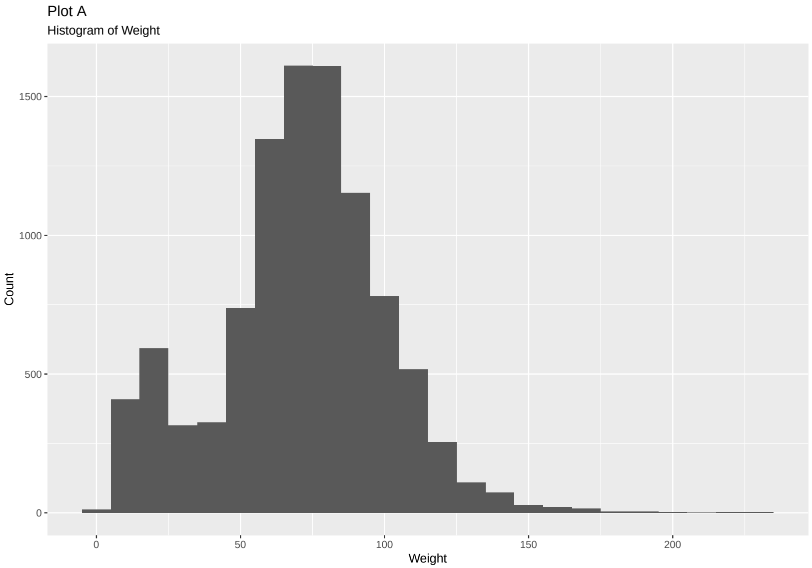
Figure 2: A boxplot is a visual representation of a 5-number summary. The “box” represents the middle 50% of the data, or the interquartile range. The line inside the box indicates the median or 50th percentile. The whiskers, the lines coming out from the box, extend 1.5 x IQR beyond Q1 and Q3. Values larger or smaller than that range are classified as outliers and appear as points.

Boxplot whiskers and outliers

- The ***whiskers*** of a boxplot (the lines extending out from the box) are 1.5 times the ***interquartile region*** long
  - Min whisker: Q1 - 1.5 x IQR
  - Max whisker: Q3 + 1.5 x IQR
- If a point is outside this range, it is considered to be a potential ***outlier***

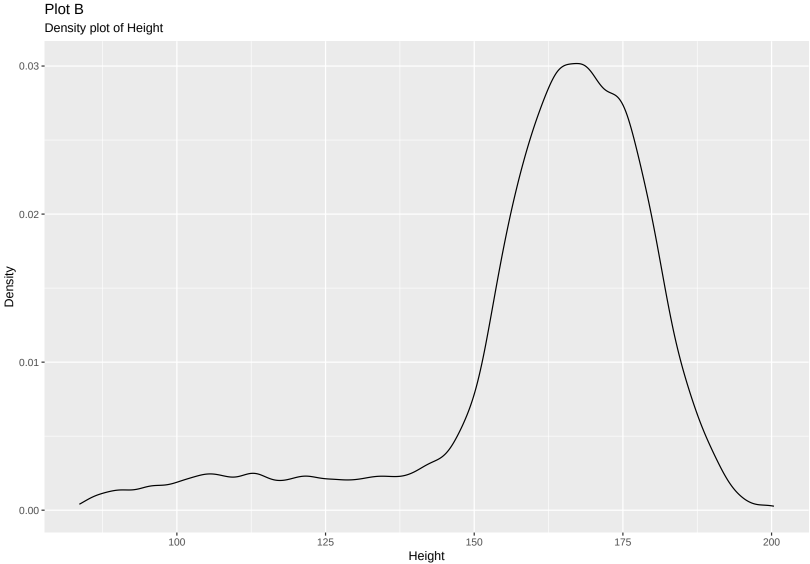
Homework 2, Plot A

The median of this distribution is 72.7, and the mean of this distribution is 71.



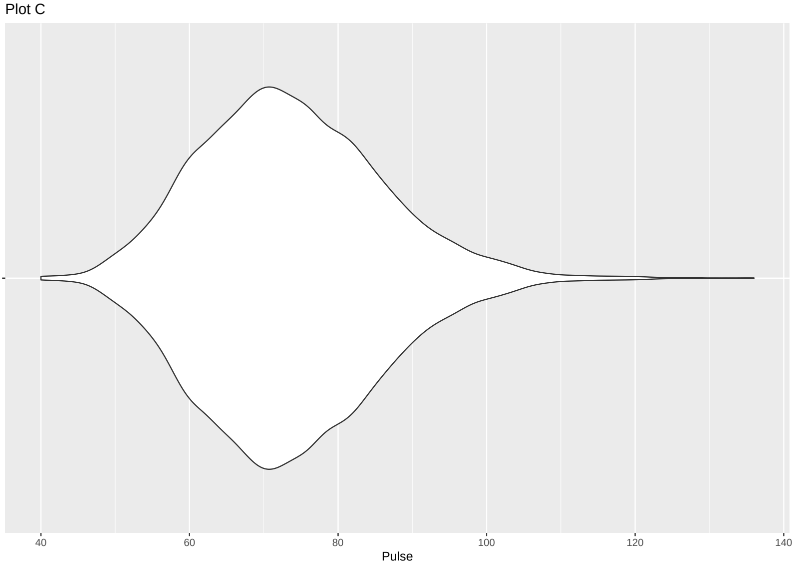
Homework 2, Plot

The median of this distribution is 166, and the mean of this distribution is 161.9.



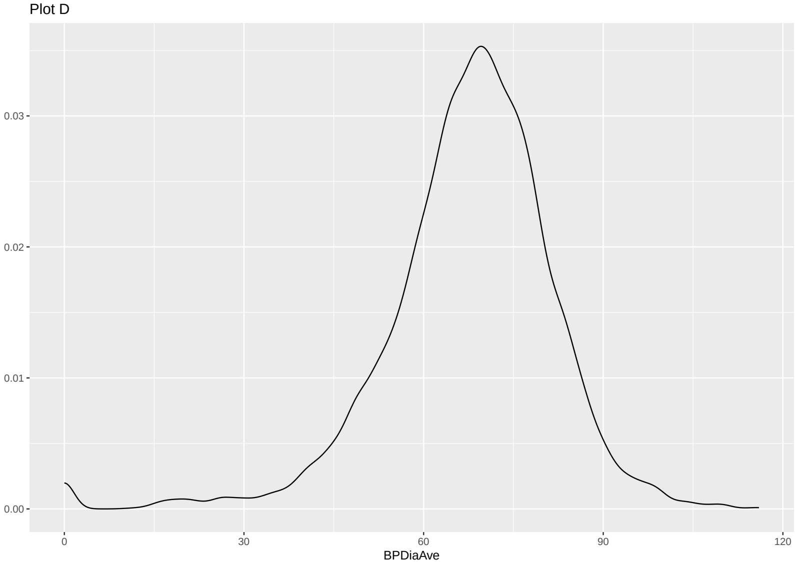
Homework 2, Plot C

The median of this distribution is 72, and the mean of this distribution is 73.6.



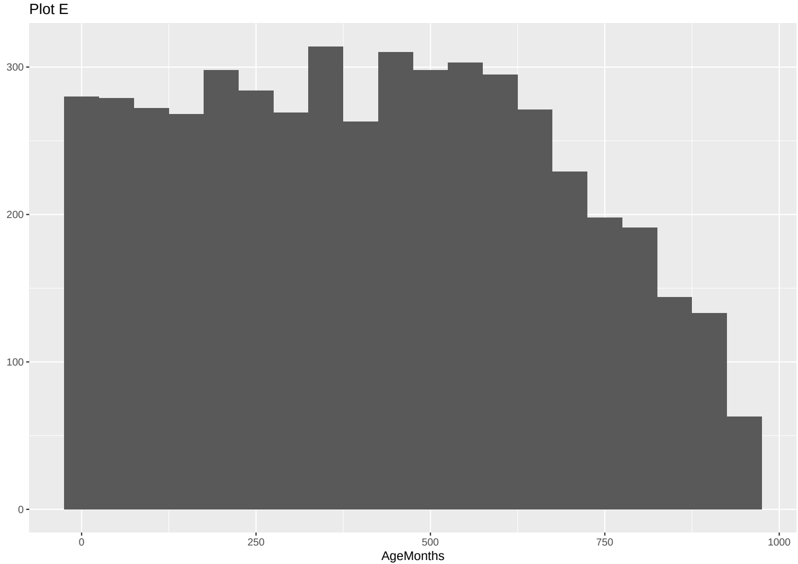
Homework 2, Plot D

The median of this distribution is 69, and the mean of this distribution is 67.5.



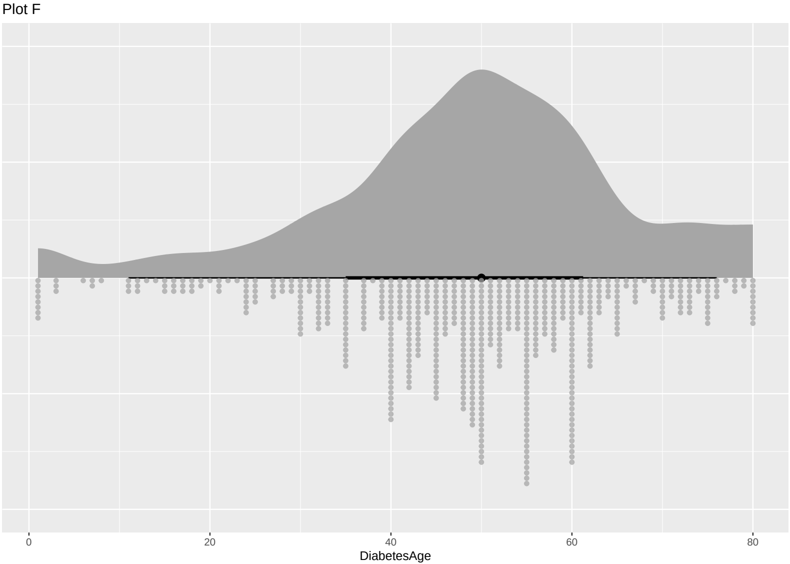
Homework 2, Plot E

The median of this distribution is 418, and the mean of this distribution is 420.1.

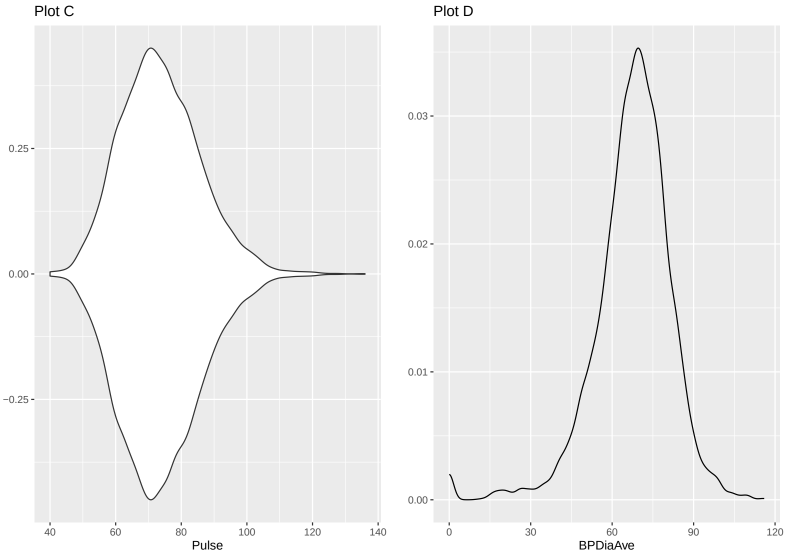


Homework 2, Plot F

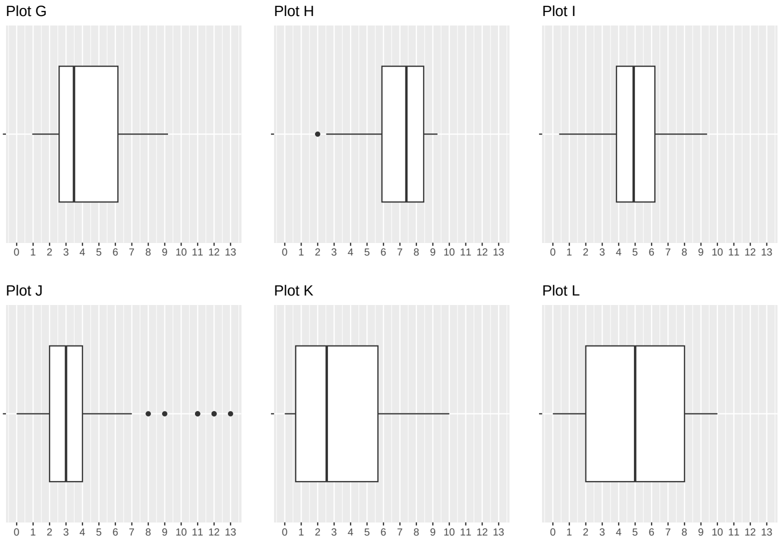
The median of this distribution is 50, and the mean of this distribution is 48.4.



Homework 2, Summary Statistics



Homework 2, Boxplots



Contingency Tables: Counts

Variable 2

Variable 1	Category 1	Category 2	Total
Category 1	A	B	A + B
Category 2	C	D	C + D
Total	A + C	B + D	A + B + C + D

Figure 3: How to construct a contingency table with counts for 2 categorical variables.

Calculating Proportions by row

Variable 2

Variable 1

	Category 1	Category 2	Total
Category 1	A / (A + C)	B / (B + D)	1
Category 2	C / (A + C)	D / (B + D)	1
Total	(A + C) / (A + B + C + D)	(B + D) / (A + B + C + D)	1

Figure 4: The row totals are all 1, which is the maximum value of a proportion. This indicates that the denominator for the proportions is the row total for each cell.

Calculating Proportions by Column

Variable 2

Variable 1

	Category 1	Category 2	Total
Category 1	A / (A + B)	B / (A + B)	(A + B) / (A + B + C + D)
Category 2	C / (C + D)	D / (C + D)	(C + D) / (A + B + C + D)
Total	1	1	1

Figure 5: The column totals are all 1, which is the maximum value of a proportion. This indicates that the denominator for the proportions is the column total for each cell.

Chapter 3 Objectives

- Define probability, random processes, and the law of large numbers
- Describe the sample space for disjoint and non-disjoint outcomes
- Calculate probabilities using the General Addition and Multiplication Rules
- Create a probability distribution for disjoint outcomes

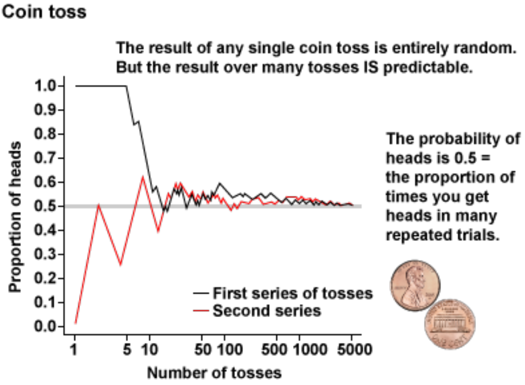
Defining the sample space

The **sample space** is the total collection of possible outcomes for a **random process**.

- Die rolls: 1, 2, 3, 4, 5, 6
- Coin flips: heads, tails
- Stock market: up, down, no change

Law of Large Numbers

As more observations are collected, the sample statistic $\hat{p}$ or $\bar{x}$ of a particular outcome approaches the population proportion p or population mean μ for that outcome.



The General Addition Rule

The probability of event A ***or*** event B occurring is the sum of the probability that A occurs and the probability that B occurs minus the probability that A *and* B occurs.

$$
\begin{aligned}
P(A \,\text{or}\, B) &= P(A) + P(B) - P(A \,\text{and}\, B) \\
&= P(A) + P(B) - P(A \cup B) \\
&= P(A \cap B)
\end{aligned}
$$

The Addition Rule for Disjoint Events

When events A and B are ***disjoint***, the probability of event A ***or*** event B occurring is just the sum of the probability that A occurs and the probability that B occurs, because the probability that event A *and* event B occurs is 0.

$$
\begin{aligned}
P(A \,\mathrm{or}\, B) &= P(A) + P(B) - P(A \,\mathrm{and}\, B) \\
&= P(A) + P(B) \\
&= P(A \cap B)
\end{aligned}
$$

Dependent Processes

- If random process B is ***dependent*** on random process A, then the probability of random process B varies based on the outcome of random process A

Dependent Processes

- If random process B is ***dependent*** on random process A, then the probability of random process B varies based on the outcome of random process A
- *Knowing the outcome of A provides additional information about the probability of B*

The General Multiplication Rule

The probability of event A ***and*** event B occurring is the product of the probability that A occurs and the conditional probability that B occurs given that A has already occurred.

$$
\begin{aligned}
P(A \,\text{and}\, B) &= P(A) \times P(B \,\text{given}\, A) \\
&= P(A) \times P(B|A) \\
&= P(A \cup B)
\end{aligned}
$$

Independent Processes

- If random process B is ***independent*** of random process A, then the probability of random process B does NOT vary based on the outcome of random process A

Independent Processes

- If random process B is ***independent*** of random process A, then the probability of random process B does NOT vary based on the outcome of random process A
- *Knowing the outcome of A does NOT provide additional information about the probability of B*

Multiplication Rule for Independent Processes

The probability of event A ***and*** event B occurring is the product of the probability that A occurs and the probability that B occurs, because the probability of B does not change based on the outcome of A.

$$\begin{aligned} P(A \,\text{and}\, B) &= P(A) \times P(B \,\text{given}\, A) \\ &= P(A) \times P(B|A) \\ &= P(A) \times P(B) \\ &= P(A \cup B) \end{aligned}$$

How do you know if two random processes are independent?

How do you know if two random processes are independent?

- Compare the conditional probabilities of B given the different possible outcomes of A. If $P(B|A) \approx P(B)$ for all values of A, then the two random processes are likely independent.

How do you know if two random processes are independent?

- Compare the conditional probabilities of B given the different possible outcomes of A. If $P(B|A) \approx P(B)$ for all values of A, then the two random processes are likely independent.
- Calculate the probability that event A and B occur under both an independence model $(P(A \, \text{and} \, B) = P(A) \times P(B))$ and a dependence model $(P(A \, \text{and} \, B) = P(A) \times P(B|A)$.
  - If $P(A) \times P(B) \approx P(A) \times P(B|A)$, then A and B are likely ***independent processes***.
  - If $P(A) \times P(B) \neq P(A) \times P(B|A)$, then A and B are likely ***dependent*** processes.

Standardizing Normal Distributions with Z-Scores

A ***Z-score*** is the number of standard deviations a value falls above (when positive) or below (when negative) the mean of the data

- Center the data at 0 by subtracting the mean from each score
- Scale the units of the data to 1 by dividing the centered data by the standard deviation

$$
\begin{aligned}
Z &= \frac{\text{observed value} - \text{mean}}{\text{standard deviation}} \\
&= \frac{x - \mu}{\sigma}
\end{aligned}
$$

Probabilities with the Standard Normal Distribution

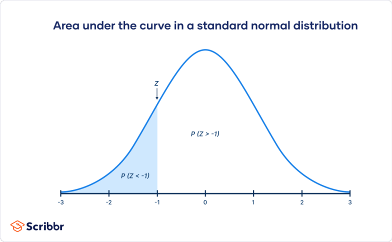


Figure 6: The shaded area under this normal probability distribution is the proportion of observations which are ***less than*** a given threshold.

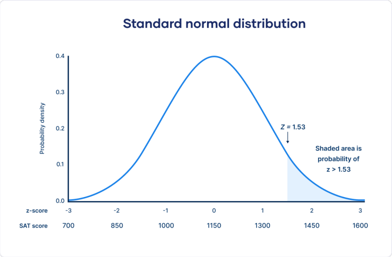


Figure 7: The shaded area under this normal probability distribution is the proportion of observations which are ***greater than*** a given threshold.