Univariate summaries and visualizations

Ryan Miller

Grinnell College
Statistics

Overview

1. Categorical variables
   - Frequencies and proportions
   - Bar charts vs. pie charts
2. Quantitative variables
   - Histograms
   - Center, shape, and spread

Categorical variables

Below is the categorical variable "Party" for 10 cases from a data set of all members in the 118th US Congress (2023-2025)

Name	Party
Grace F. Napolitano	D
Eleanor Holmes Norton	D
Harold Rogers	R
Bill Pascrell Jr.	D
Maxine Waters	D
Steny H. Hoyer	D
James E. Clyburn	D
Nancy Pelosi	D
Danny K. Davis	D
John Carter	R

How might you *summarize* the interesting aspects of this variable?

Frequencies, relative frequencies, and tables

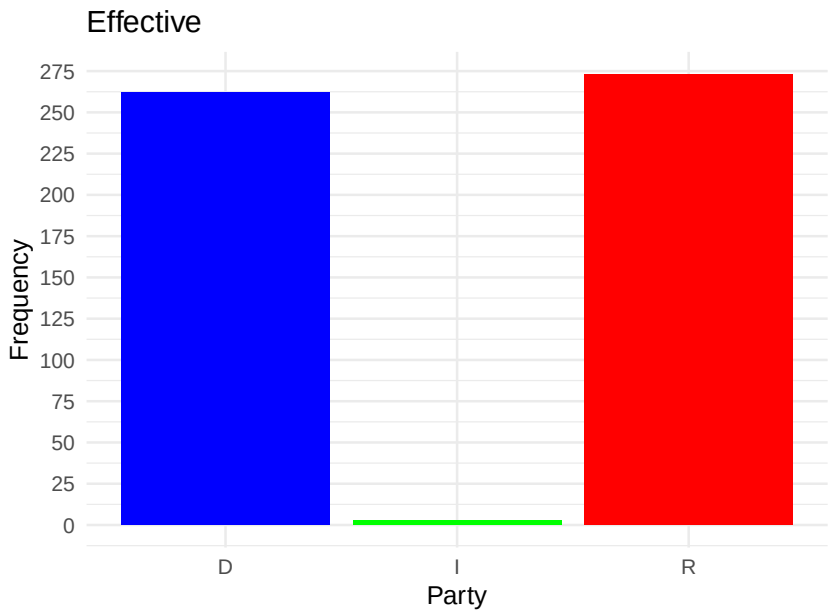
- **Frequencies** (counts) are simple tallies of how many times a category appears across cases
- **Relative frequencies** (proportions) is the ratio of a category's frequency to the total number of cases under consideration
- A **one-way table** is a common way to present the frequencies (or relative frequencies) for *all of the categories* of a *single categorical variable*

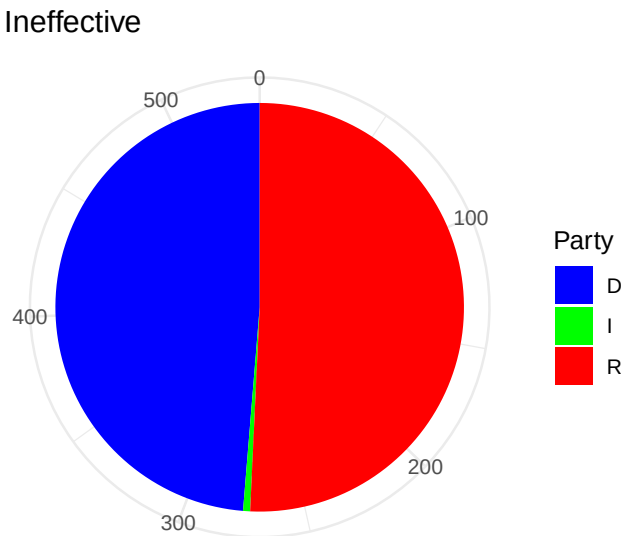
Table 1: One-way frequency table of 'Party'

Party	Frequency
D	262
I	3
R	273

Bar charts vs. pie charts

A **bar chart** is the preferred way to visualize the information in a frequency table.

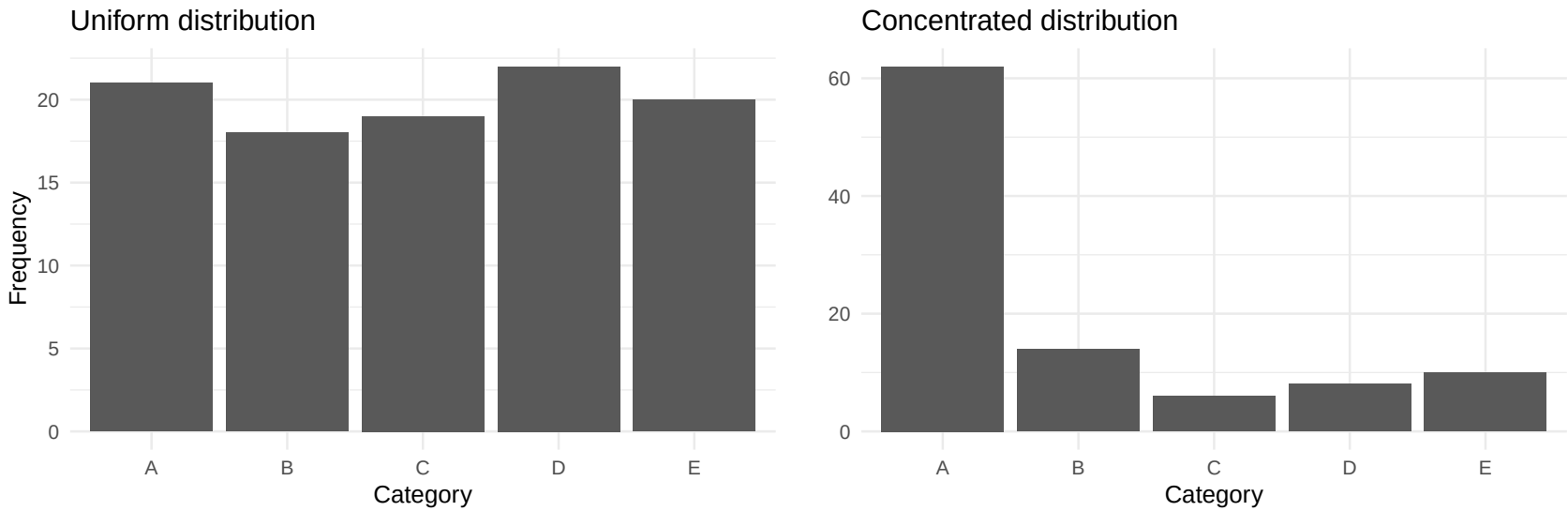




Pie charts, while popular, should be avoided because humans can more accurately judge heights (bars) than they can judge angles and areas (pie slices)

Distributions

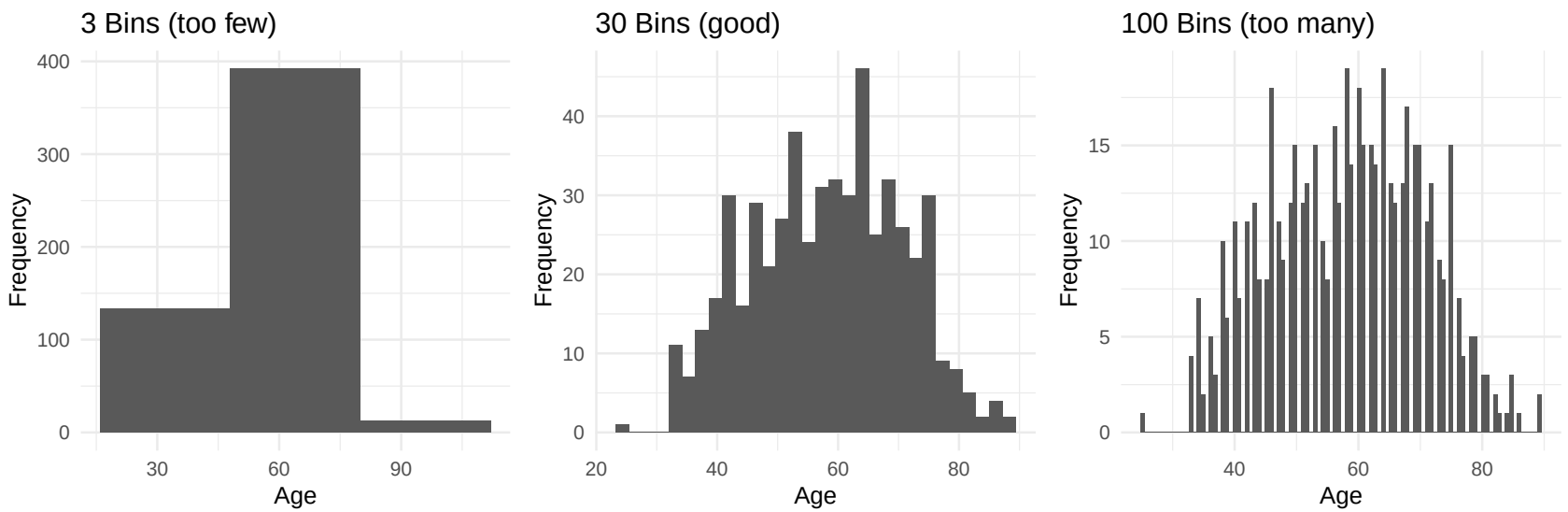
The **distribution** of a variable shows its possible values and how often they occur. Bar charts visually display the distribution of categorical variables:



When a categorical variable is *nominal*, we may describe it as *approximately uniform* (similar frequencies across categories) or *concentrated/not uniform* (some categories are more prevalent).

Quantitative (numeric) variables

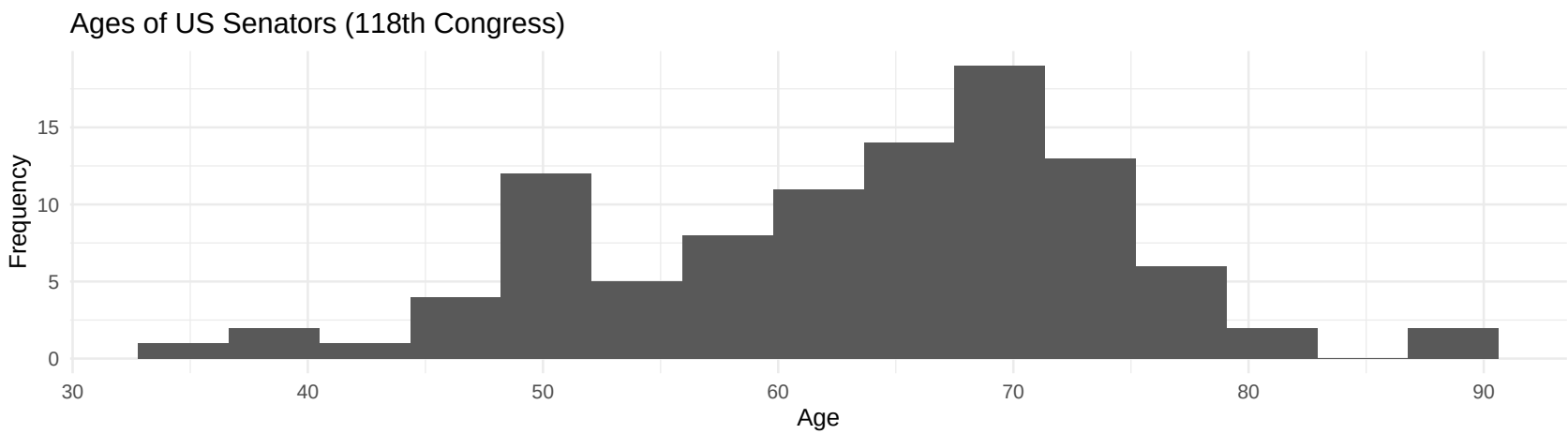
- The main idea of bar charts can be applied quantitative variables by dividing the variable's scale into *equal-sized bins*, then finding frequencies for each bin
  - This type of graph is known as a **histogram**
  - Choosing an appropriate number of bins is essential for accurately assessing a variable's distribution



Distributions of quantitative variables

When describing the distribution of a quantitative (numeric) variable we should address the following:

1. **Shape** - is the distribution symmetric or skewed? is it bell-shaped?
2. **Center** - where is the distribution centered at?
3. **Spread** - how much do values of the variable tend to vary?
4. **Unusual points** - are there any outliers? excessive zeros or anomalies?



Describing a quantitative variable's "shape" and "outliers"

- You will not be responsible for describing shape or outliers *quantitatively*, but you should know how to describe them *qualitatively*. Common descriptions include:
  - "skewed-right" if there's a long tail on its right (positive) side, or "skewed-left" if there's a long tail on its left side
  - "bell-shaped" for a central peak with roughly even tails on both sides
  - "bimodal" or "multimodel" if the distribution has two peaks or multiple peaks (respectively)
- Data-points that are more than 3-standard deviations from a variable's mean are often considered outliers

Describing a quantitative variable's "center"

We'll focus on two different ways of numerically describing a quantitative variable's center:

- **Mean** - the arithmetic average of a variable, if we have n observations the mean of variable "X" is given by: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Median** - the middle value if the data were arranged in ascending order

The median is often called a **robust** measure of center because it tends not be influenced by outliers. In contrast, the mean is pulled towards outliers.

Describing a quantitative variable's "spread"

We have several ways to summarize a variable's spread:

- **Standard deviation** - the average deviation (distance) of individual data-points from the distribution's mean
  - $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$
- **Range** - the difference in the data's maximum and minimum values
- **Interquartile Range (IQR)** - the difference in the 75^{th} and 25^{th} percentiles of the data (also called Q3 and Q1 respectively)

The standard deviation and range are *greatly* influenced by outliers, while the IQR is resistant/robust.

Practice

For each of the following variables (visualized below):

1. Determine whether the mean or median is larger.
2. Decide whether it's more reasonable to describe the variable's "spread" using standard deviation or IQR.