STAT 113
Variability

Colin Reimer Dawson

Oberlin College

September 14, 2017

Outline

Distribution of a Quantitative Variable

The distribution of a quantitative variable is characterized by:

A. **Shape (symmetric, skewed, bimodal, etc.)**
B. Center (mean, median)
C. Spread (Interquartile Range, Standard Deviation)
D. Outliers (if any)

Skewness

- A distribution is **skewed** when the extreme values on one side are more extreme than those on the other.
- We call a distribution **right-skewed** when the longer "tail" is on the right, and **left-skewed** when the longer tail is on the left.

Distribution of a Quantitative Variable

The distribution of a numeric variable is characterized by:

A. Shape (symmetric, skewed, bimodal, etc.)
B. **Center (mean, median)**
C. Spread (Interquartile Range, Standard Deviation)
D. Outliers (if any)

Resistance/Robustness

- The mean is strongly affected by skew and by **outliers**
- The mean is pulled toward the extreme values.
- In these cases, we generally prefer a measure of central tendency which is **resistant** to the influence of extreme values (also called **robust**).
- The **median** is a resist/robust measure of center.

Outline

Distribution of a Quantitative Variable

The distribution of a numeric variable is characterized by:

A. Shape (symmetric, skewed, bimodal, etc.)
B. Center (mean, median)
C. **Spread (Interquartile Range, Standard Deviation)**
D. Outliers (if any)

Measures of Variability

- We want to quantify the consistency, or lack thereof, of the data.
- A general term for "lack of consistency" is **variability**.
- We will look at:
  - Range
  - Interquartile Range
  - Variance / Standard Deviation

The Range

The range is easy to compute, but not very reliable.

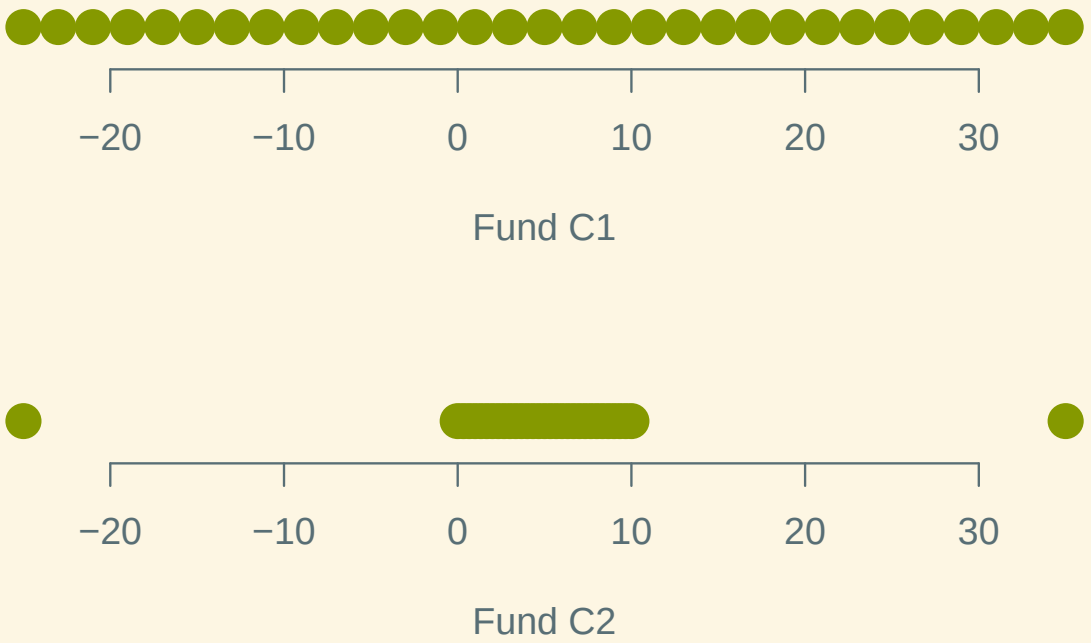


Figure: Historical Annual Returns for Two Hypothetical Index Funds

The Range

The range is easy to compute, but not very reliable.

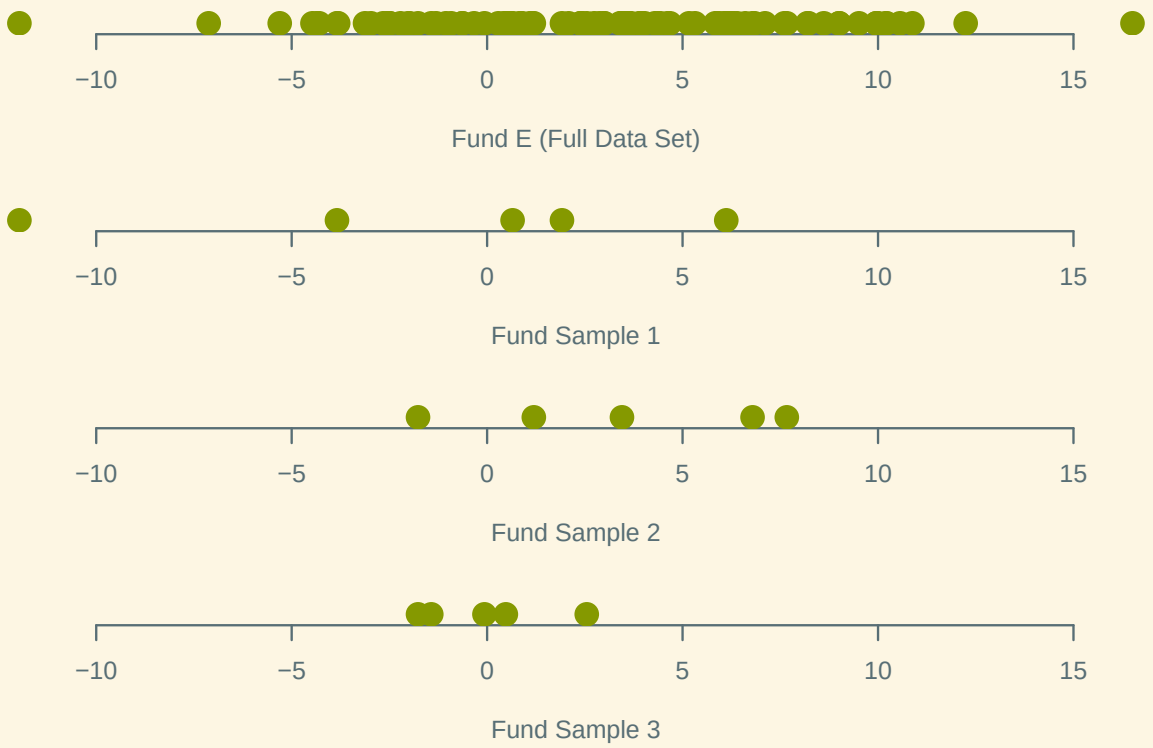


Figure: Annual Returns for 3 random samples of 5 years

Outline

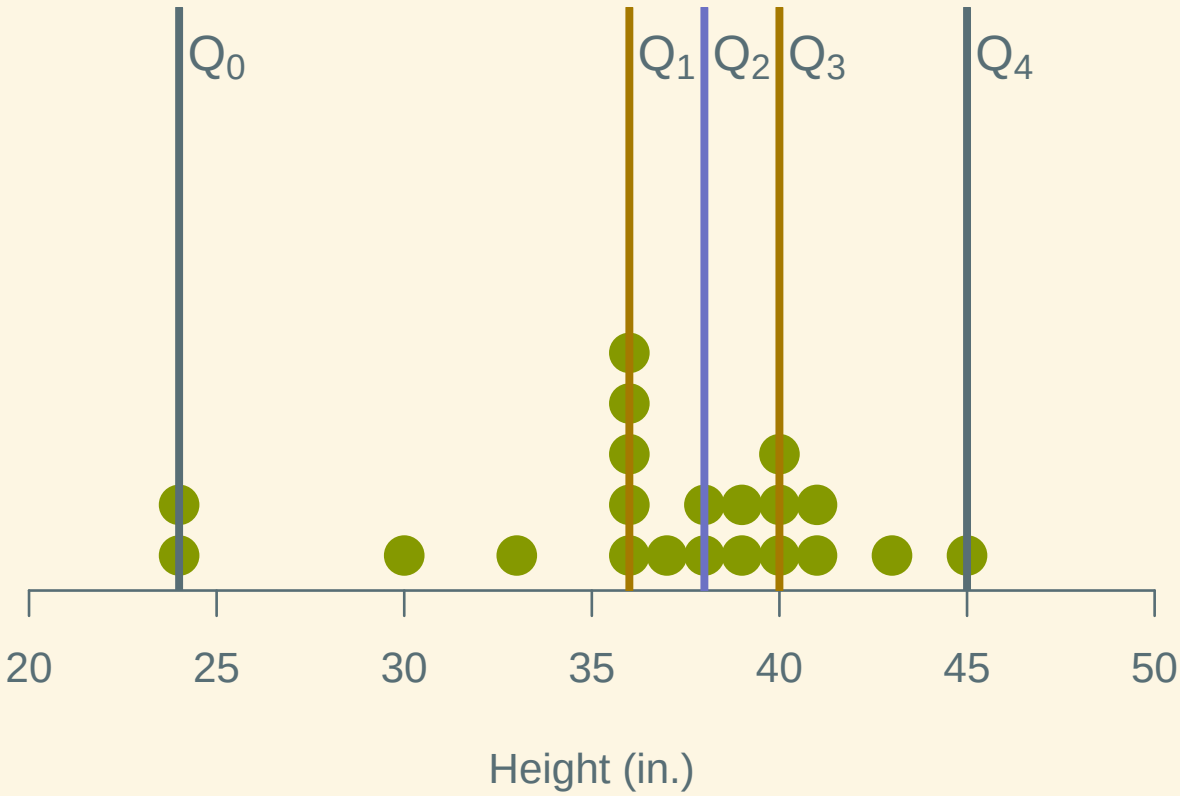
Robust Measures of Variability

- We'd like a more **robust** measure of variability, which is not affected so much by extreme values.
- Analogous to the median: describe the "middle" part of the data.
- The idea: find the "middle half" of the data, and then take its range.
- Specifically, exclude the lowest 25% and the highest 25%, and take the difference between the highest and lowest remaining values.

Quartiles

- The median divides the data in two.
- Percentiles divide the data into 100 pieces.
- Quartiles divide the data into _____. The k^{th} **quartile** (written Q_k) is the point below which k *quarters* of the data lies.
- So, in terms of quartiles, the median is _____, the minimum value is _____, the maximum value is _____.
- We can calculate the range using quartiles as _____.

Quartiles



The Inter-Quartile Range (IQR)

The Inter-Quartile Range (IQR)

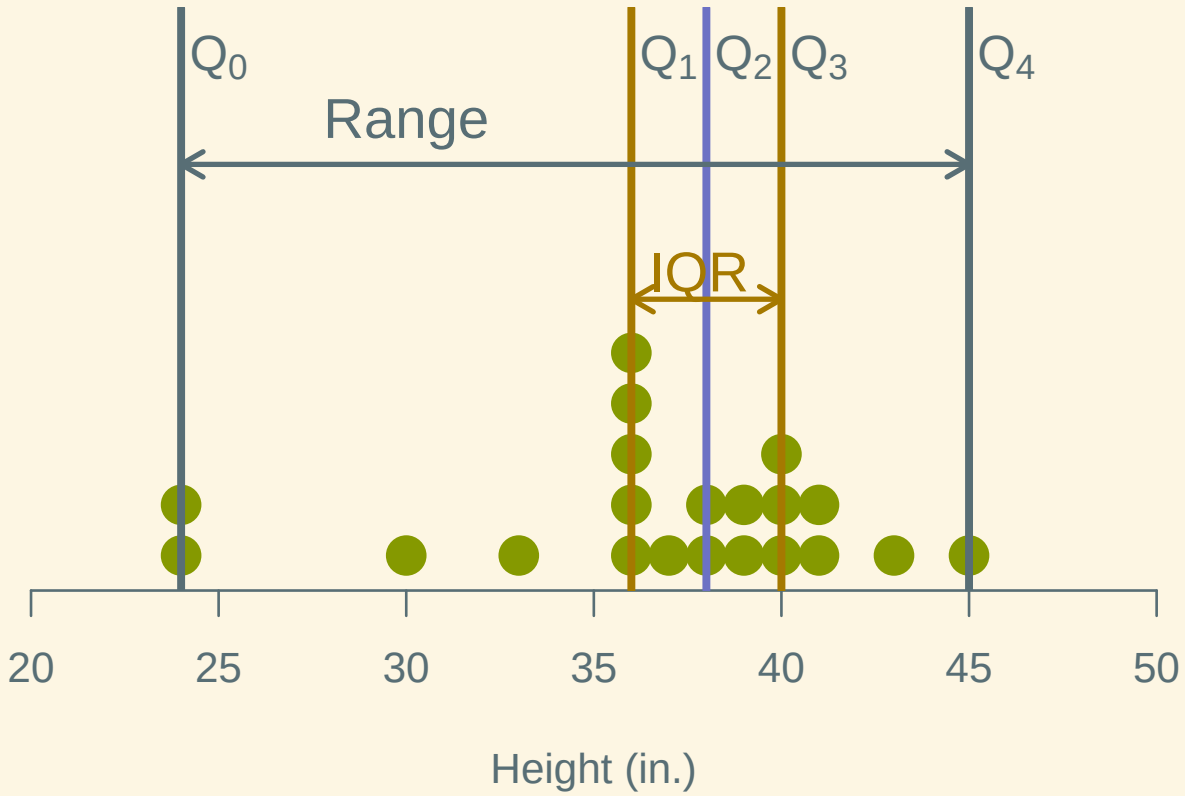
The **Inter-Quartile Range** (or **IQR**) is the distance between the first and third quartiles:

$$IQR = Q_3 - Q_1$$

Pedantic Note

The IQR is a *single number*, not the two quartiles themselves.

The Inter-Quartile Range (IQR)



The Five-Number Summary

Five-number Summary

- The quartiles are very natural to report together to describe the center and spread of a distribution.
- Q_0 through Q_4 collectively form the **five-number summary** of a quantitative distribution.

$$\begin{aligned}\text{Five Number Summary} &= (x_{\text{min}}, Q_1, \text{Median}, Q_3, x_{\text{max}}) \\ &= (Q_0, Q_1, Q_2, Q_3, Q_4)\end{aligned}$$

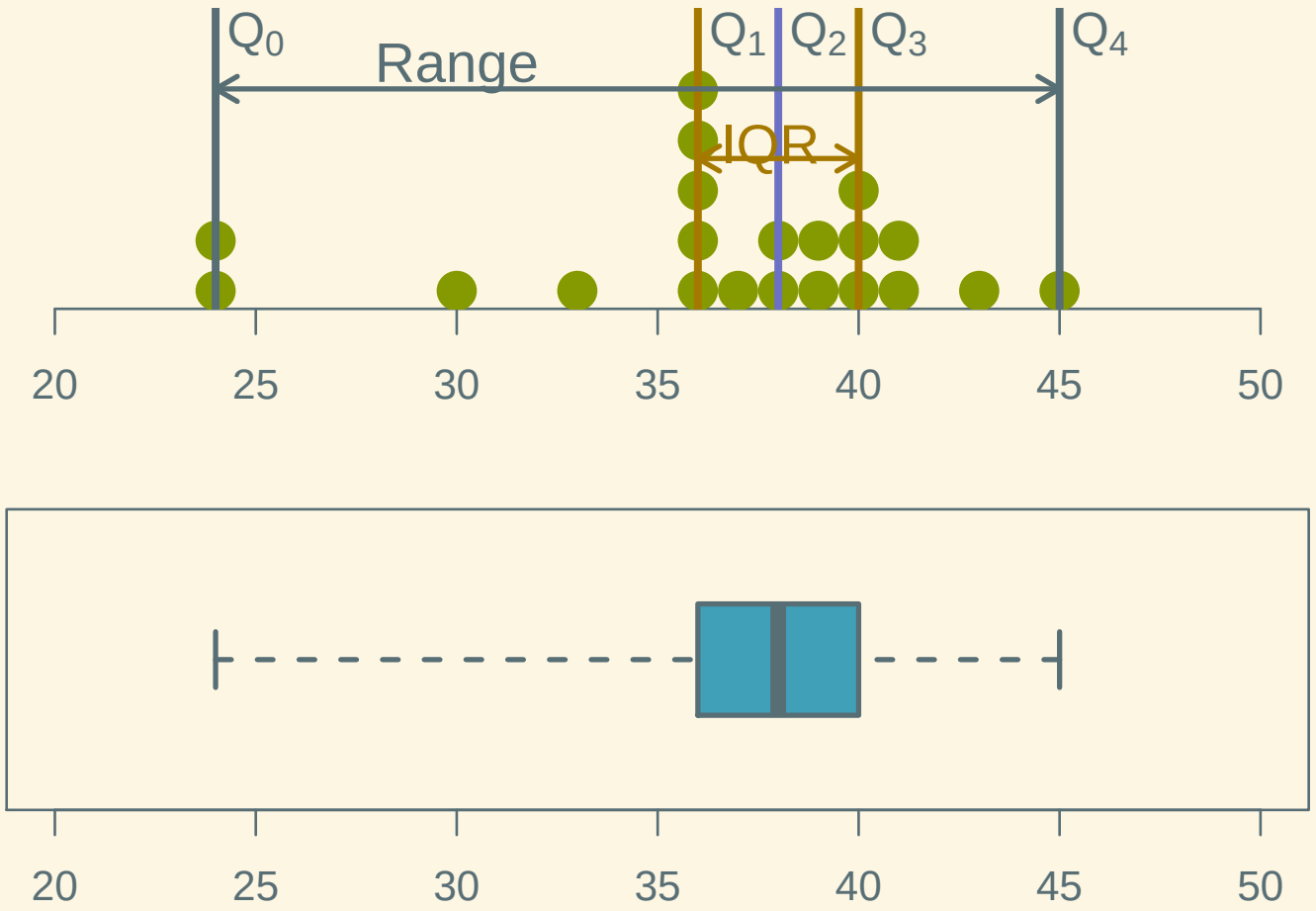
Box-and-Whisker Plots

Box-and-Whisker Plots

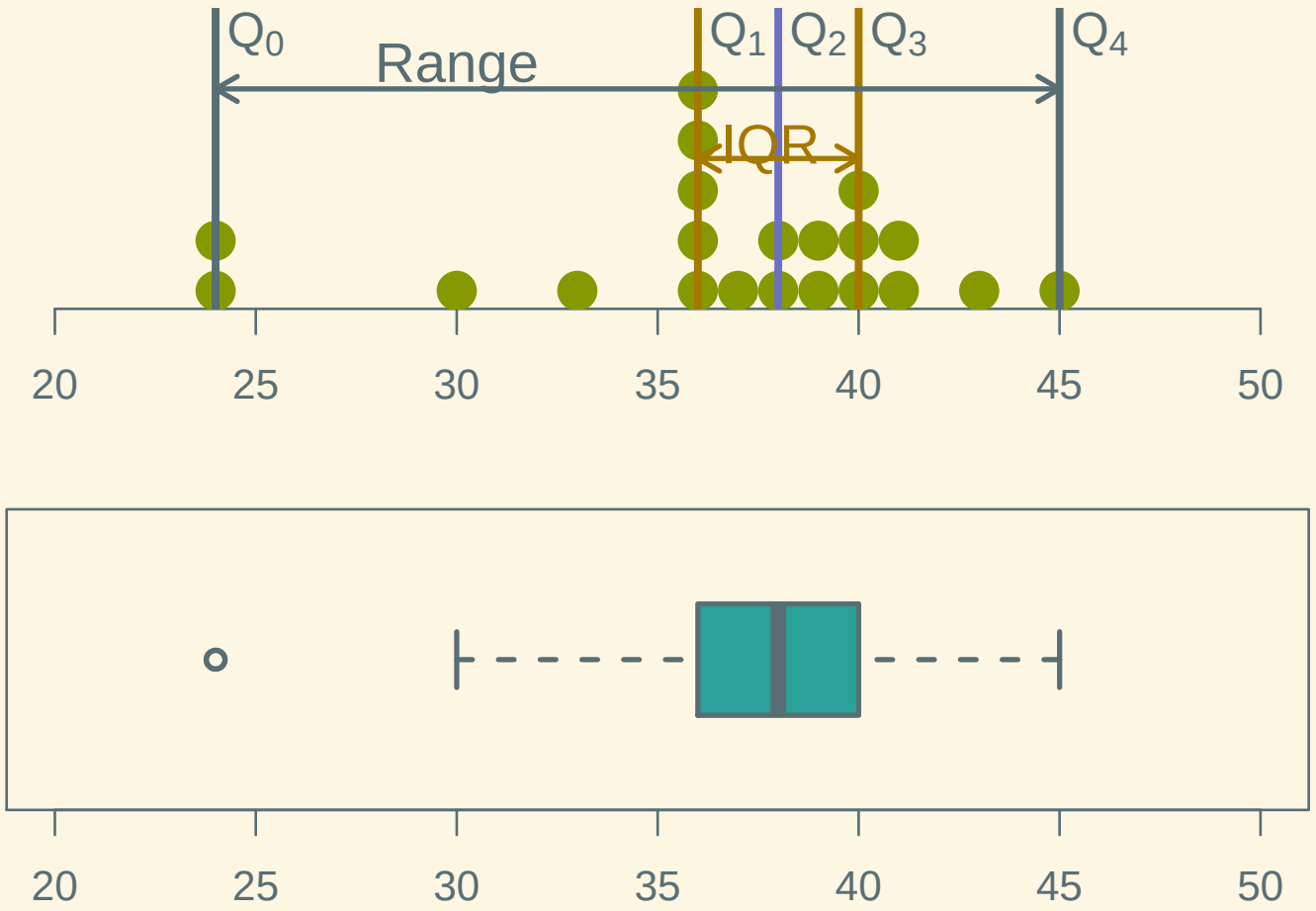
From the five-number summary, we construct a graph called a **box-and-whisker plot** (or just **box plot**, for short)

1. Draw an axis
2. Draw a rectangle (box) from Q_1 to Q_3
3. Draw a line across the box (or place a dot) at Q_2
4. Draw lines (whiskers) extending outward from the box on both sides to either
   (a) (Simplest version) x_{min} and x_{max}.
   (b) (R default) $Q_1 - 1.5IQR$ and $Q_3 + 1.5IQR$.
5. In version (b), plot points beyond the whiskers individually.

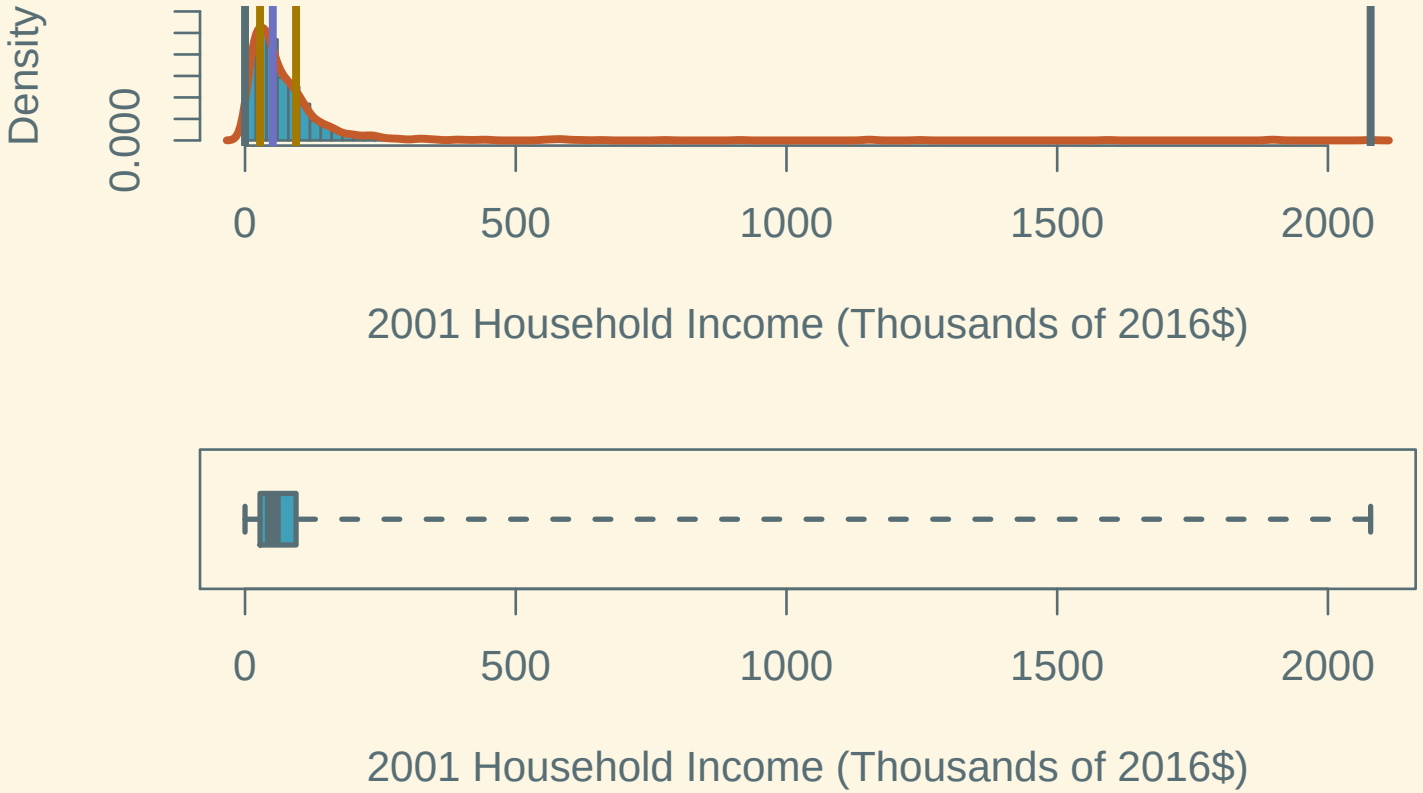
Box-and-Whisker Plot: Version 1



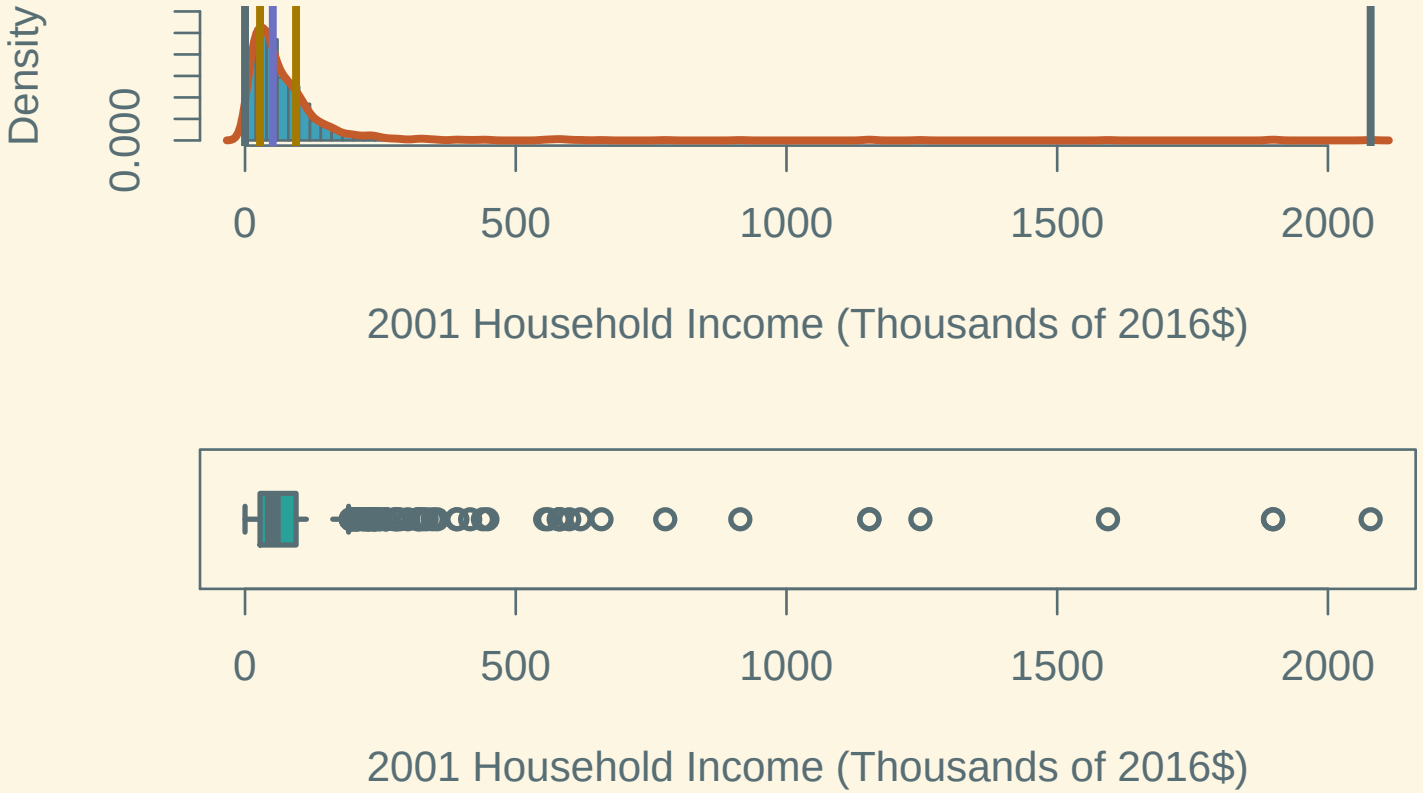
Box-and-Whisker Plot: Version 2



Box-and-Whisker Plot: Right Skew



Box-and-Whisker Plot: Right Skew



Matching Graphs to Variables

Handout

Outline

Deviations

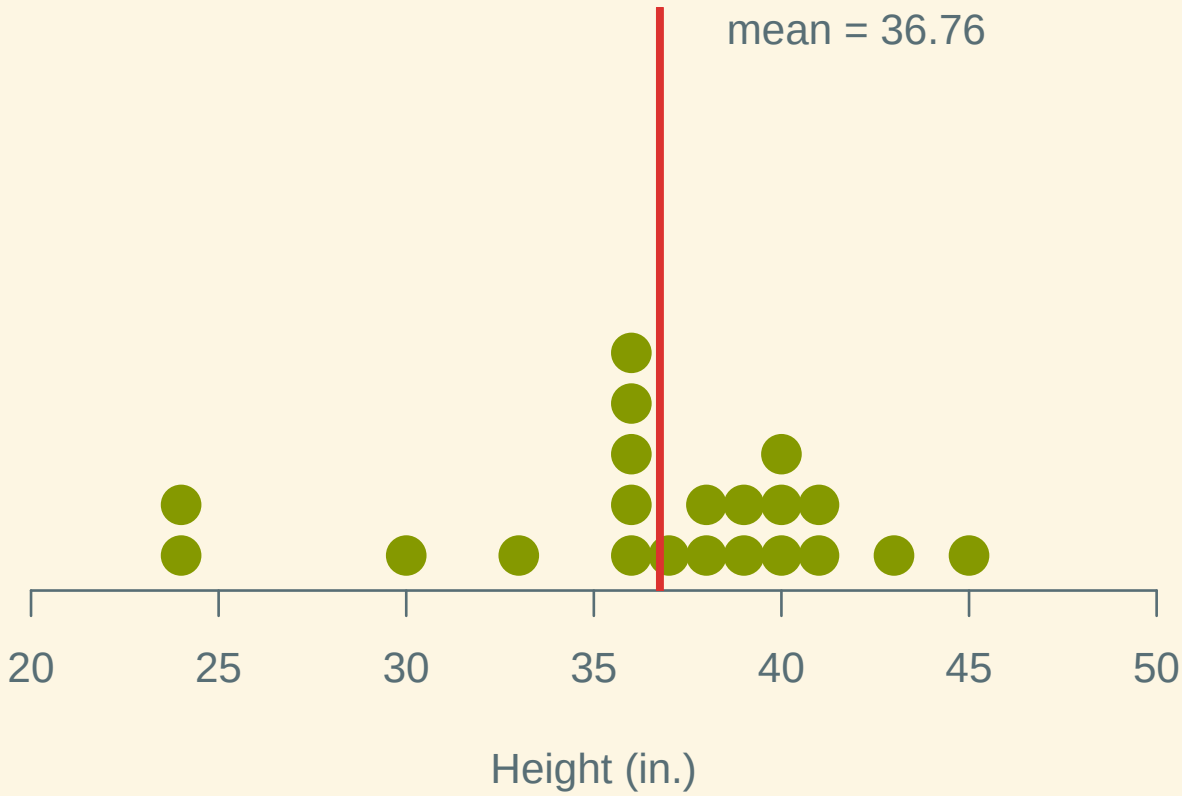
Rather than simply measuring the distance between extremes, we can develop measures based on distance from "center".

Deviation Scores

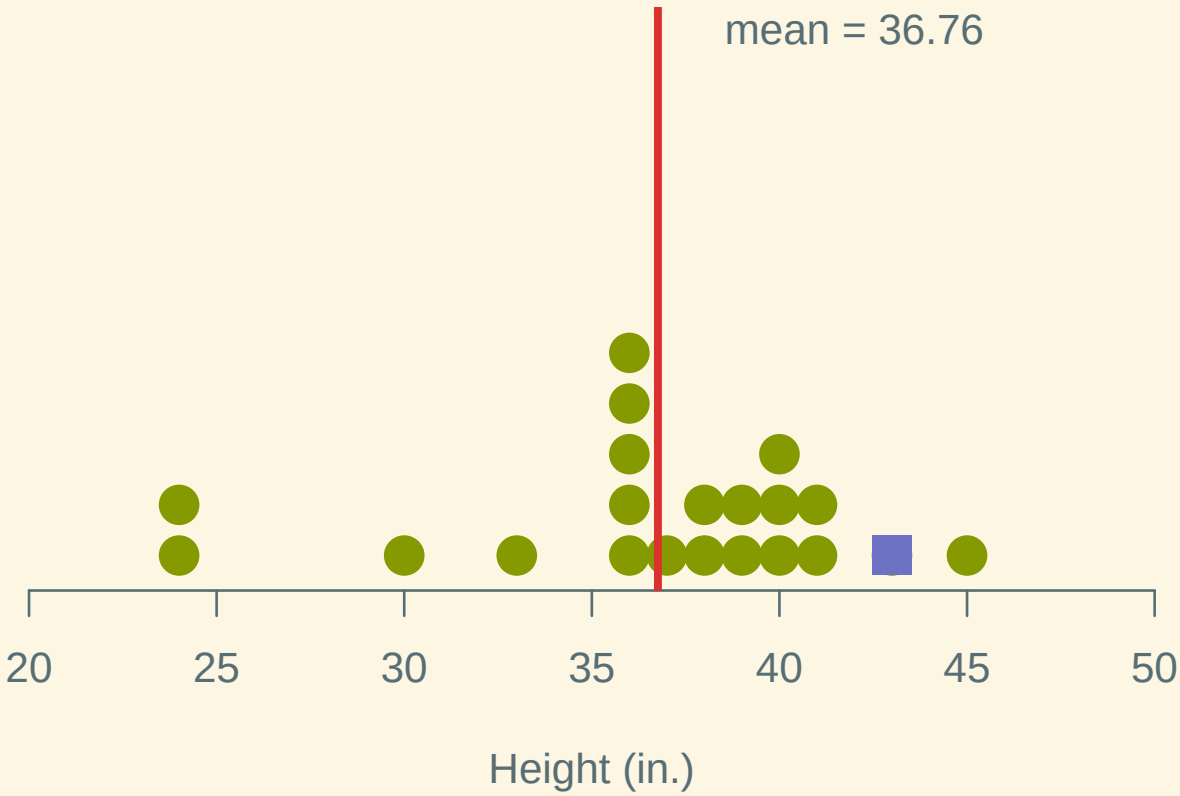
For *each* data point, its **deviation score** is its "distance" from the mean.

$$\text{Deviation}_i = x_i - \bar{x}, \qquad \text{for each } i = 1, \ldots, n$$

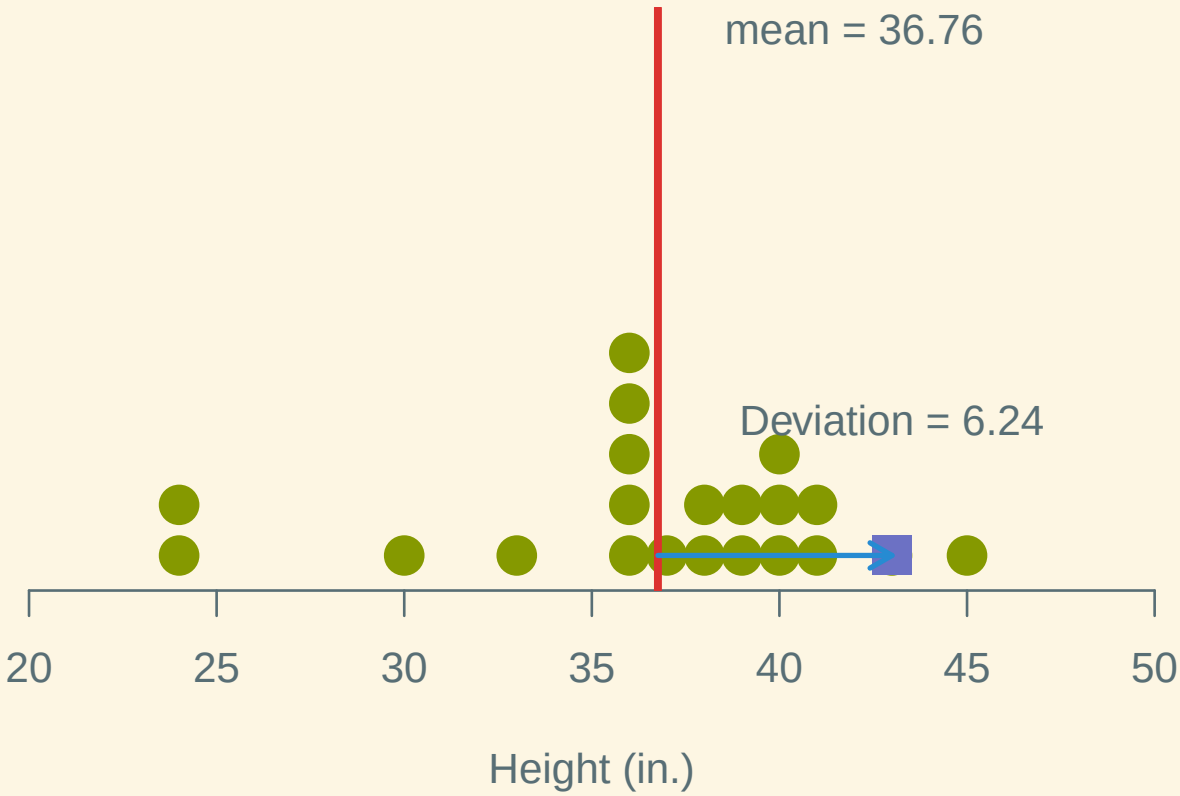
Deviations



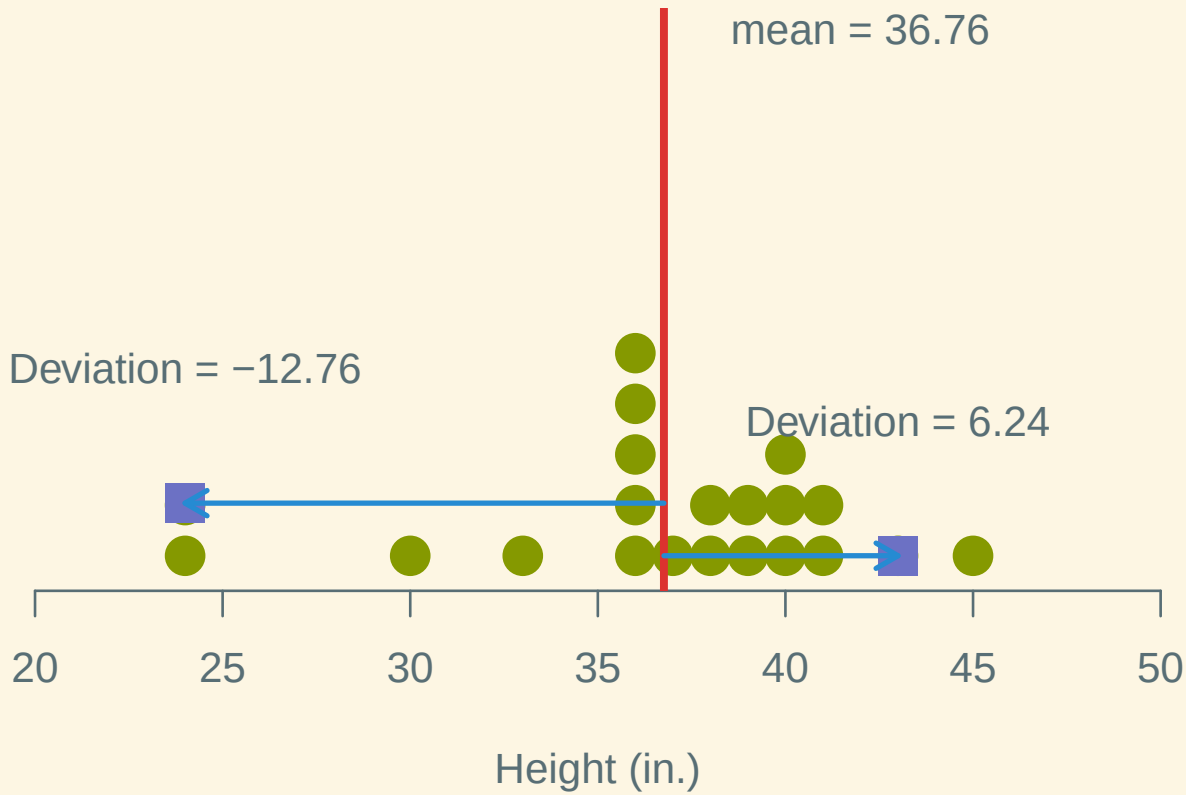
Deviations



Deviations



Deviations



How can we use these for an overall measure of spread?

Variance

- If we square all the deviations from the mean and average them, we get the **variance**.

Variance

The **variance**, written s^2, is the average of the squared deviations from the mean. That is,

$$s^2 = \frac{\sum_{i=1}^{n} \text{Deviation}_i^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

What's with that denominator?

- With an average, you're supposed to divide by the number of things, aren't you? Why $n-1$?
- Usually we are working with a sample, and are interested in *estimating* the population variability.
- We get no information about variability from the first observation, so there are only $n-1$ "degrees of freedom" in the sample.
- Interesting math side fact: Variance is equivalent to average squared distance between all distinct pairs of data points.

Standard Deviation

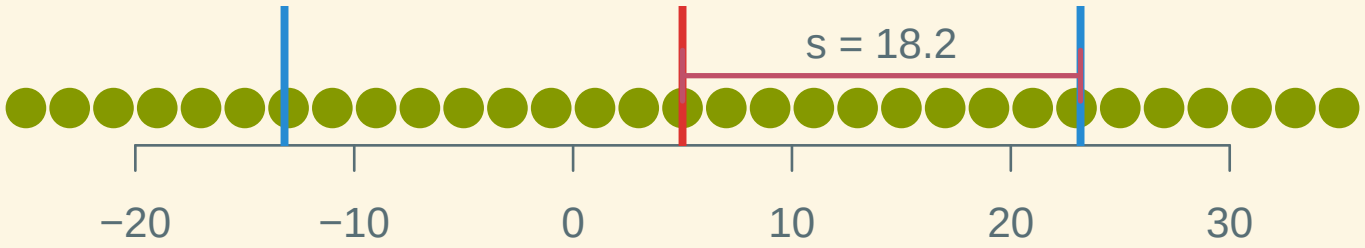
- Variance (s^2) is in squared units relative to the data.
- No problem: just take the square root.

Standard Deviation

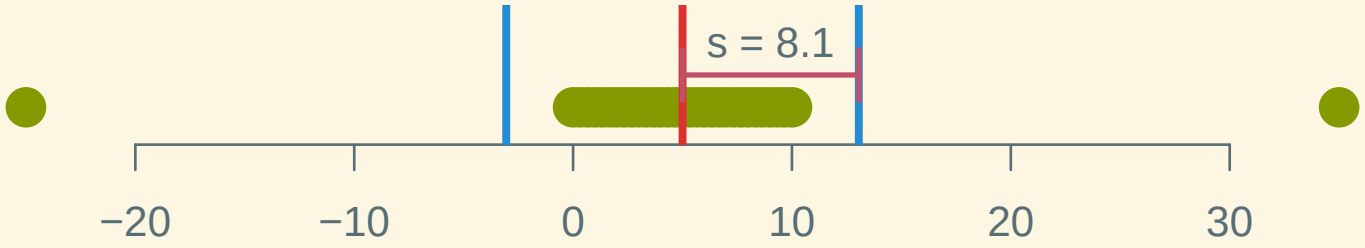
$s = \sqrt{s^2}$ is the **standard deviation**

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} \text{Deviation}_i^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Same range, different s



Fund C1



Fund C2

The standard deviation uses *all* the data.

Distribution of a Quantitative Variable

The distribution of a numeric variable is characterized by:

A. Shape (symmetric, skewed, bimodal, etc.)
B. Center (mean, median)
C. Spread (Interquartile Range, Standard Deviation)
D. **Outliers (if any)**

Outliers

- Skewness can be an important feature of a distribution.
- But sometimes a few unusual data points make an otherwise "well-behaved" distribution look skewed/multimodal.
- When not part of the overall pattern, these are called **outliers**.
    - Sometimes reflect measurement errors (e.g., misplaced decimal)
    - Sometimes represent genuinely unusual observations

On-Base Percentage

A common statistic for batters in baseball is *On-Base Percentage*

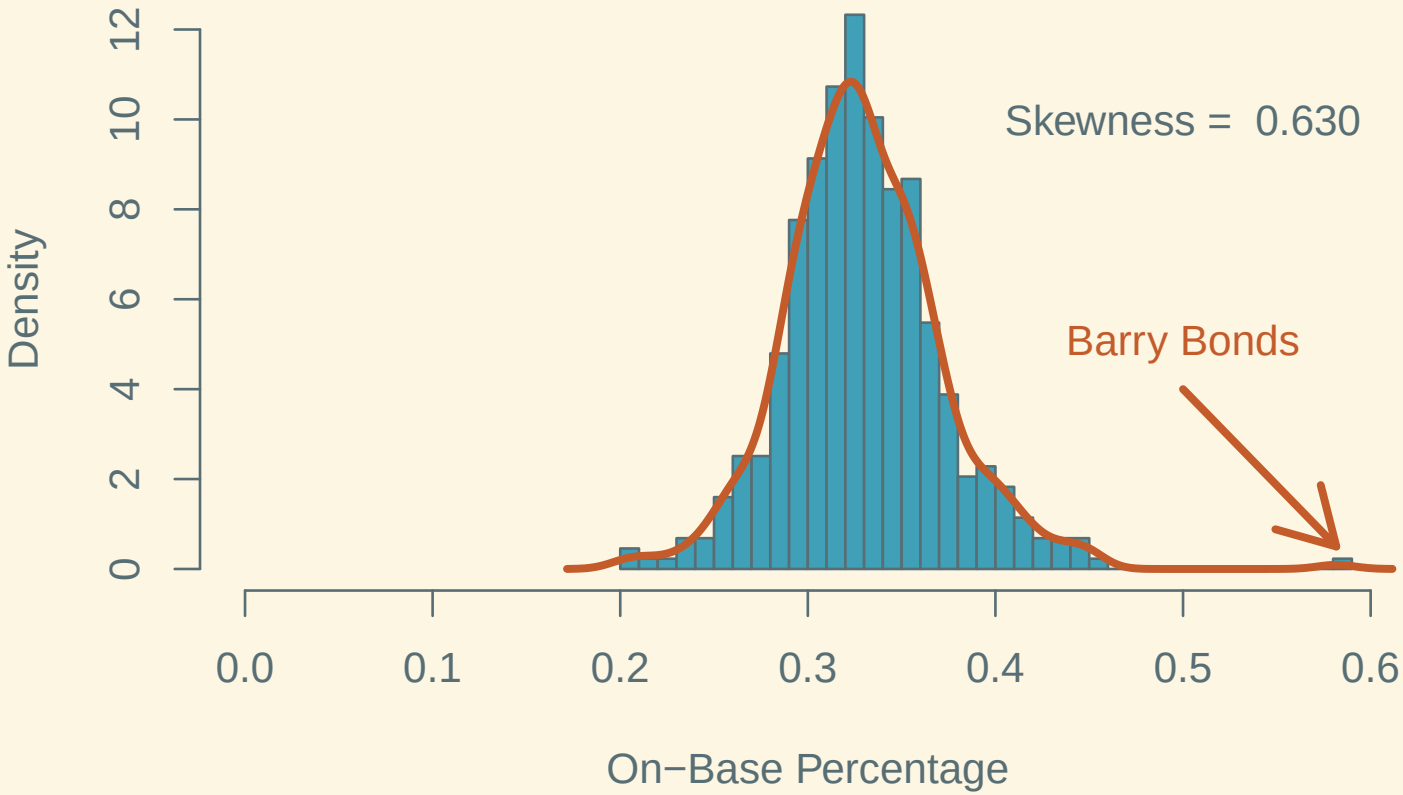
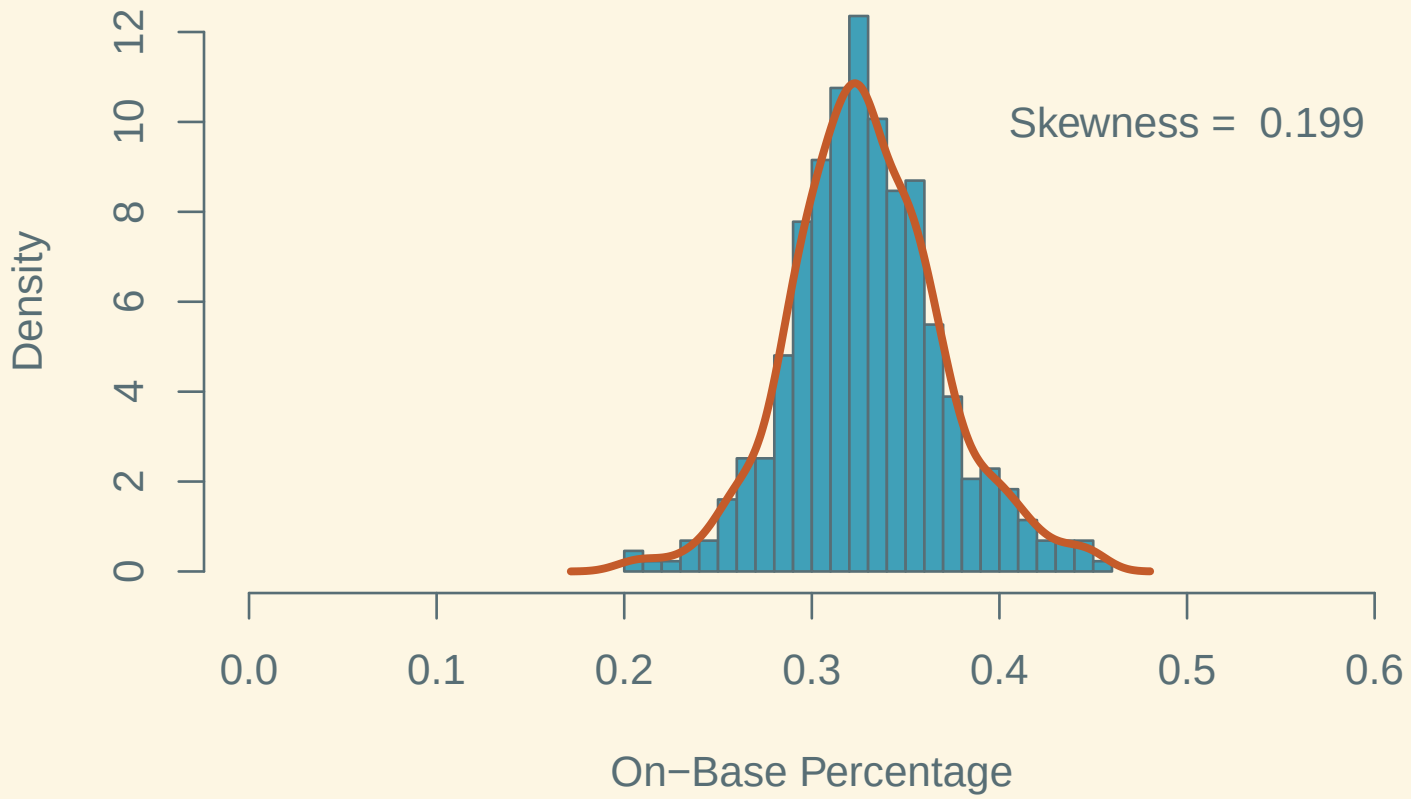


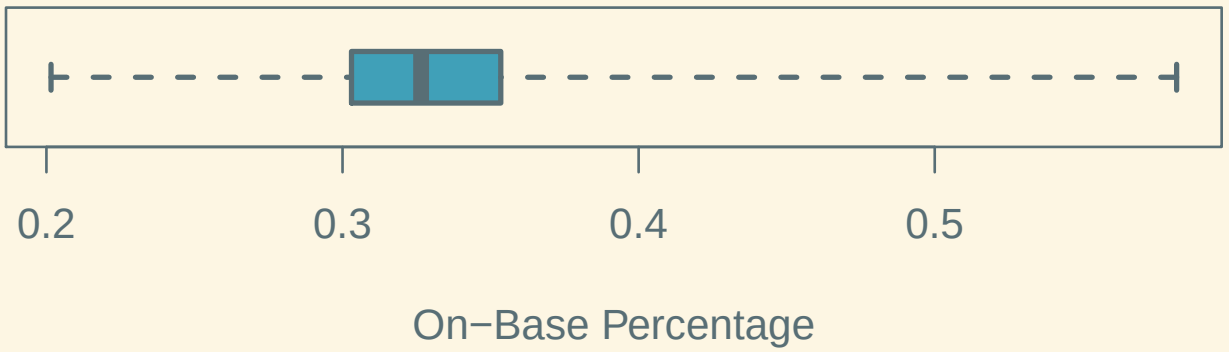
Figure: Distribution of major-league hitters with at least 100 Plate Appearances in 2002.

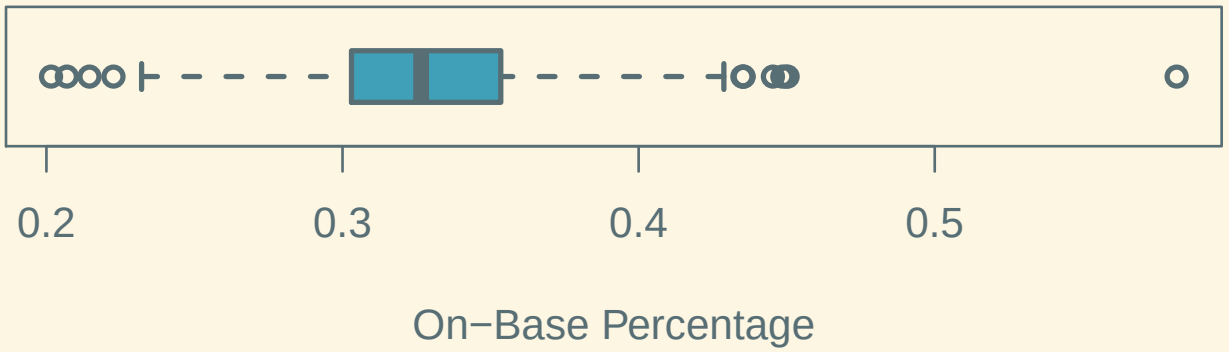
On-Base Percentage

Distribution without Bonds



Visualizing Outliers

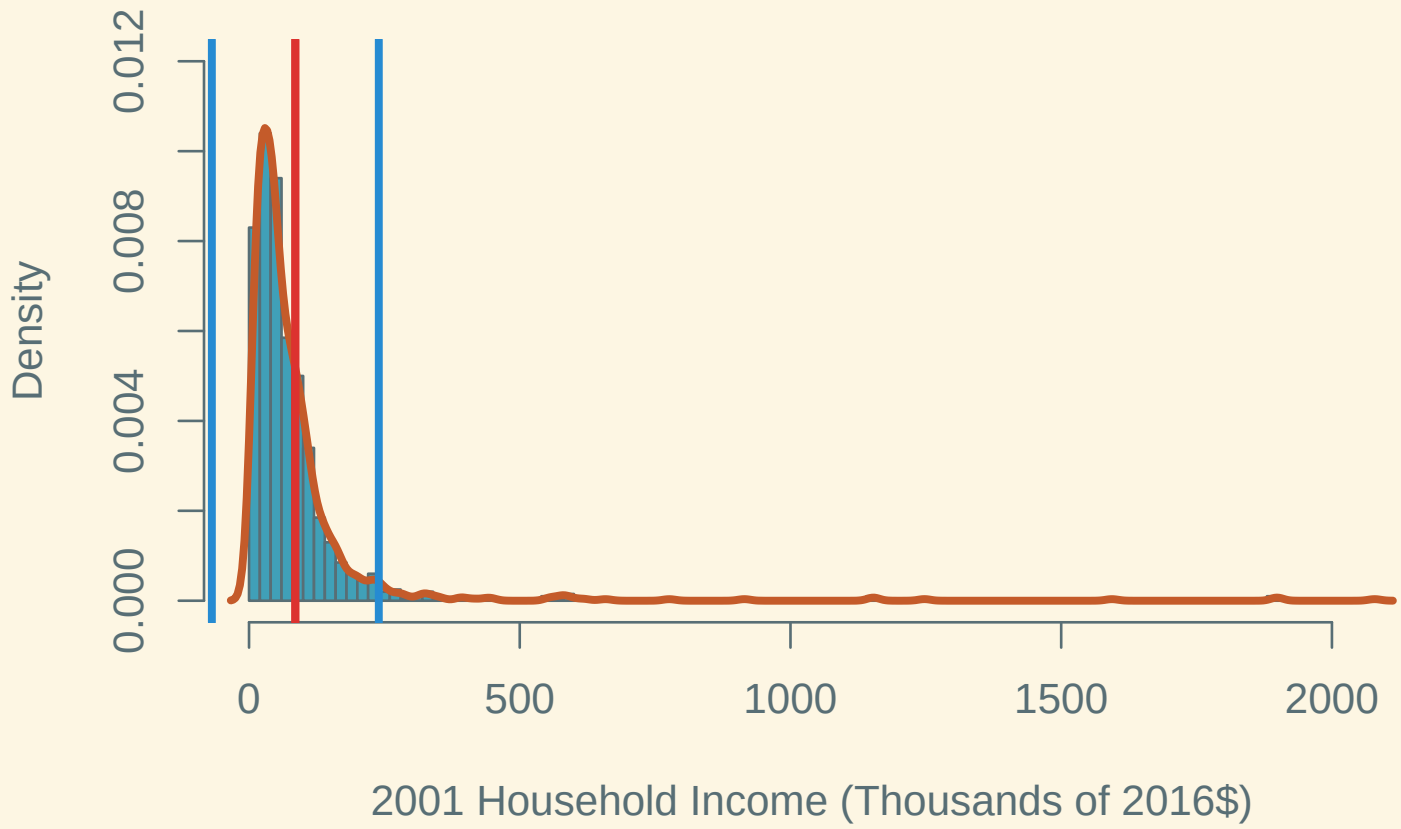




Outline

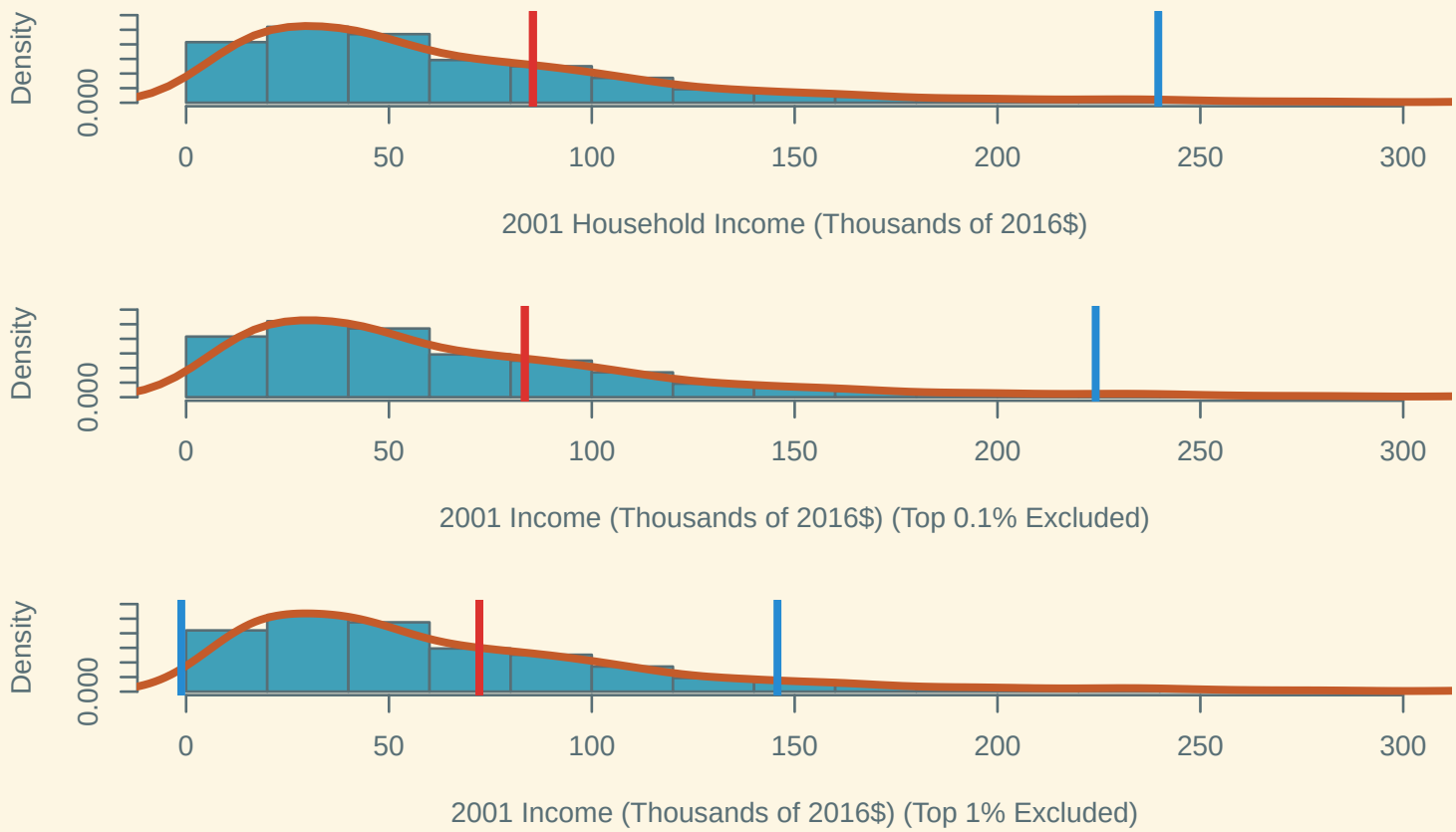
Problems with s and s^2

- These measures, even more than the mean itself, are heavily influenced by extreme values.



Problems with s and s^2

Problems with s and s^2

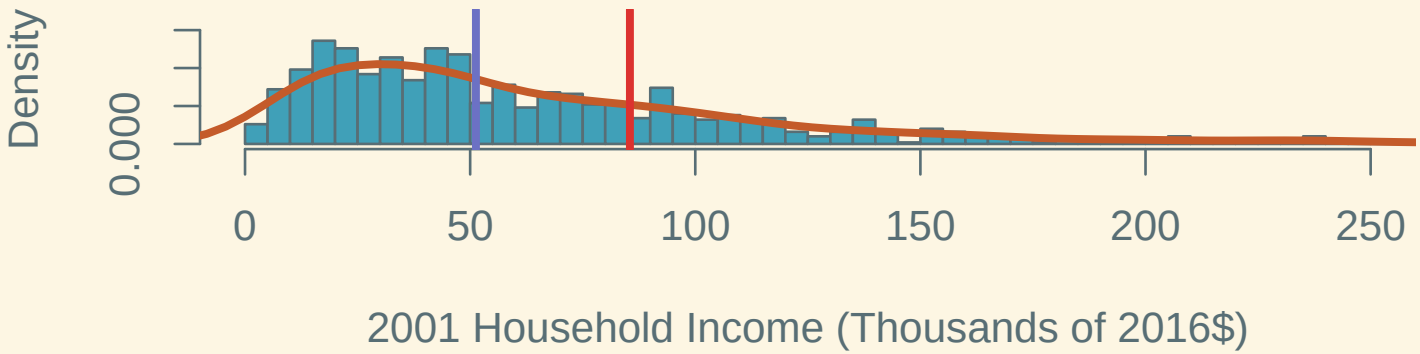


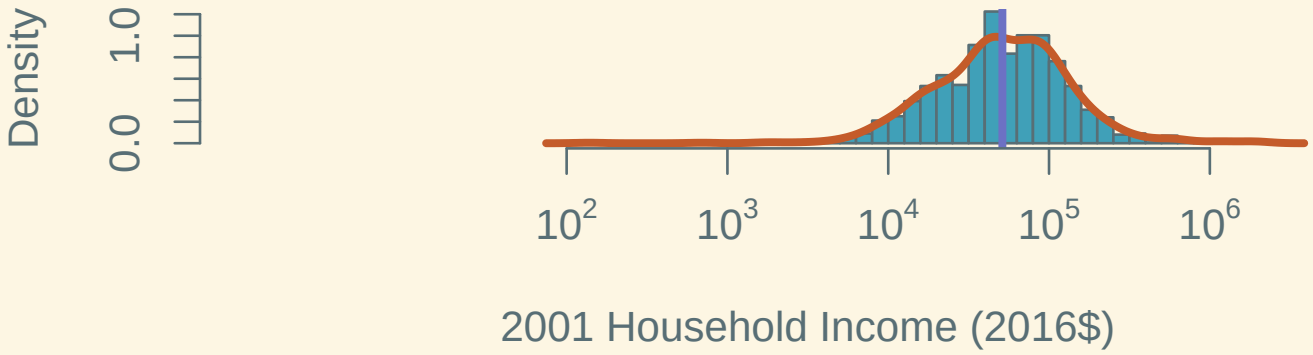
Variance-Stabilizing Transformations

- The mean and standard deviation are unstable in the presence of skew.
- However, they have such useful properties otherwise that it is often better to try to "remove" skew, rather than fall back on other measures.
- The most common way to remove skew is by a nonlinear **transformation** of the underlying scale.
  - Take the original variable, x, and define a new variable $Y = f(x)$, where f is a *one-to-one function*.
  - Most common case: right-skewed data with positive values
    - Logarithmic transform (take $y = \log(x)$)
    - Square Root (take $y = \sqrt{x}$)

Variance-Stabilizing Transformations

Original vs. Logarithmic Income Distribution:





Summary

Quantitative Data

Visualizing a quantitative variable

- Dot Plots
- Box-and-Whisker Plots
- Histograms
- Density curves

Describing the distribution of a numeric variable

- Shape (symmetry, skew, modes)
- Center (mean, median)
- Spread (IQR, standard deviation)
- Outliers (if any)

Summary

Shape and Center

- A distribution is **skewed** when the extreme values on one end are more extreme than on the other
- We say that it is skewed in the direction of the more extreme values (e.g., right-skewed if there are a few very large values)
- The mean is the "balance point of the data", written $\bar{x}$.
- Mean has nice math properties, but is affected by skew
- The median divides the cases in half
- It is **resistant** to outliers/skewness

Summary

Variability

- The range is unstable for a sample, and is *extremely* vulnerable to outliers/skew
- The Interquartile Range (IQR) is the range of the "middle half" of the data, and is "resistant" (like the median)
- The variance is the "average" of the squared deviations from each observation to the mean
- The standard deviation is the square root of the variance, in order to restore units to the original scale
- Nonlinear transformations (log, square root, etc.) can be used when appropriate to reduce skew and stabilize variance