ISTA 116: Statistical Foundations for the Information Age

Shape of a Distribution

12 September 2011

Outline

Announcements/Reminders

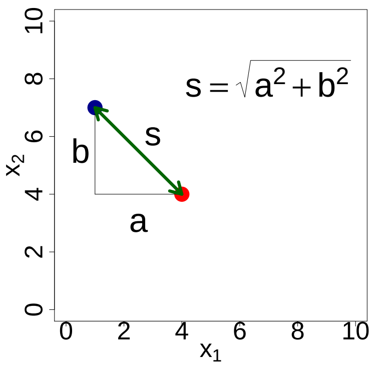
- Lab 2 due this week (see how far we get today)
  - Question 1f says "use the table from part (e)". Should say "use whatever table you used for the pie chart", as you don't have to implement your suggestion in e (though you're encouraged to do so!)

Measures of Variability

- Want a way to differentiate distributions based on how "spread out" they are, not just by their centers
- The **range**: difference between minimum and maximum values
  - Problems: only uses two values, and these are often the most unstable
  - Different distributions end up with similar ranges; similar distributions end up with different ranges
- The **variance** is based on the **squared deviations from the mean**, and uses all the data
- The **standard deviation** is the square root of the variance: takes it back to the original units

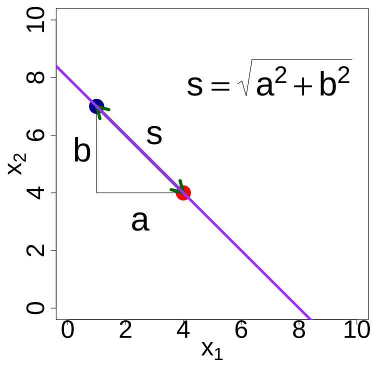
Geometric Analogy

- The standard deviation is the distance between these two points.

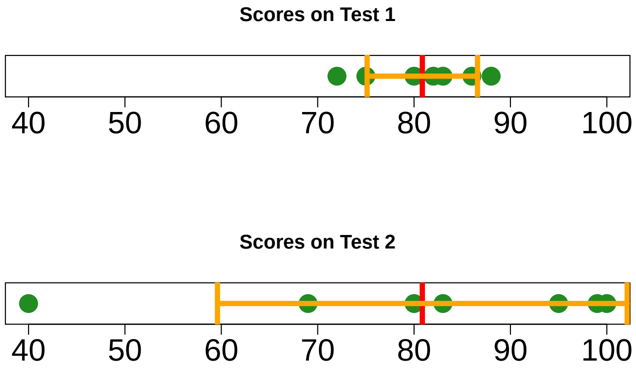


Geometric Analogy

- Notice that, since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ always, all possible data sets of two points with $\bar{x} = 4$ lie on a line.
- This is the reason for the $n - 1$ in the denominator: 2 data points, 1 dimension of deviations.



Same $\bar{x}$, different s



z-scores

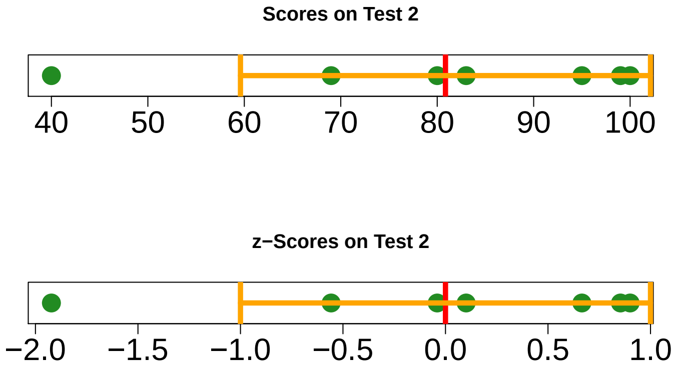
- A common application of standard deviation is as a way to measure how far a data point is from the mean, on a scale that is *independent of units*.
- By dividing each individual deviation score by the standard deviation, we obtain a **z-score** for that data point.

$$z_i = \frac{(x_i - \bar{x})}{s}$$

- Interpretation: "How many standard deviation units above the mean is that observation?" (negative = below the mean)

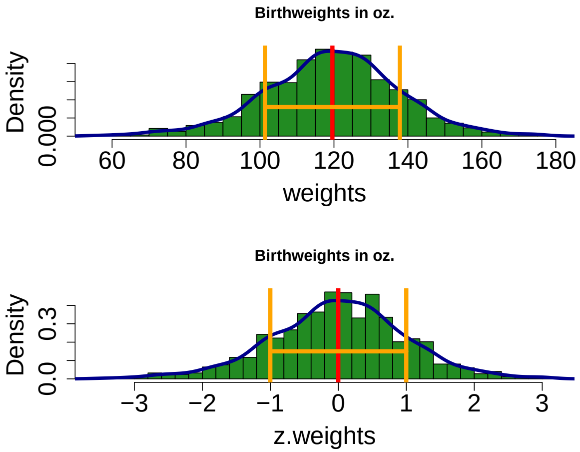
z-scores

- We can compute z-scores for the whole data set, and see their distribution.



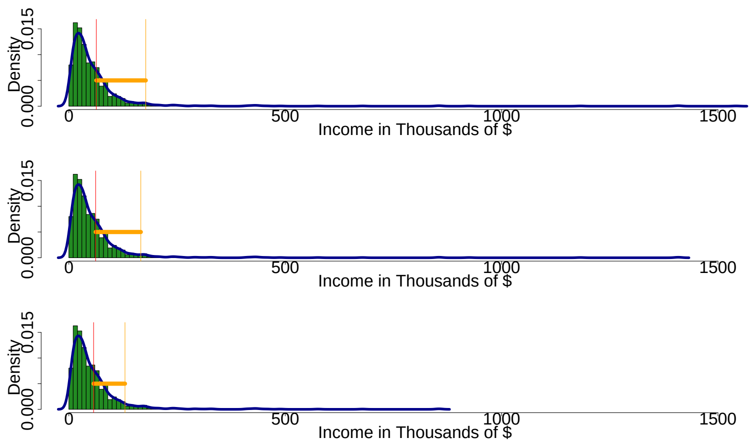
z-scores

- We can compute z-scores for the whole data set, and see their distribution.



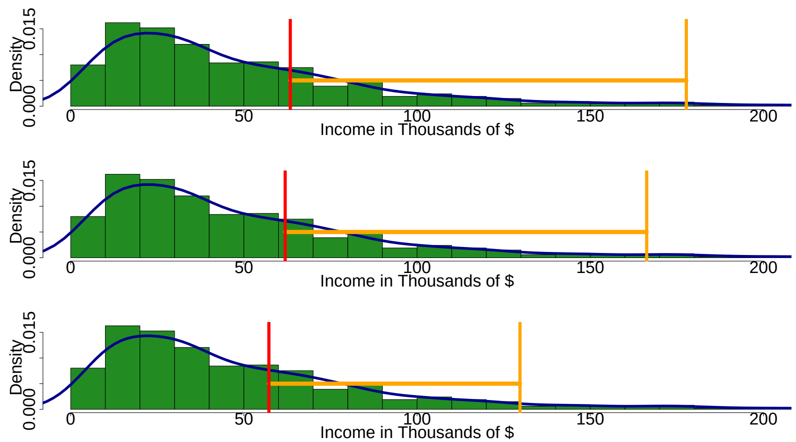
Problems with s and s^2

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



Problems with s and s^2

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



Robust Measures of Variability

- We'd like a more **robust** measure of variability, for cases like the above.
- Analogous to the median: describe what the "middle" part of the data is doing.
- The idea: describe the range of the "middle half" of the data.
- That is, exclude the lowest 25% and the highest 25%, and take the range of what remains.

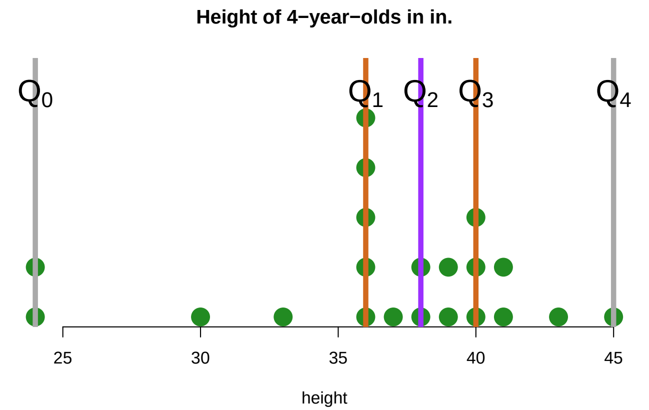
Quantiles

- Recall: the **median** is the point that one half, or 50%, of the data is below.
- Generalize this idea to define **percentiles**.
- The median is the ______ *percentile*.
- A similar idea, expressed with proportions rather than percentages, is that of the p^{th} **quantile**: same as the $100p^{\text{th}}$ percentile.
- The median is the ______ *quantile*.

Quartiles

- Notice that percentiles divide the data into 100^{th}s. We could just as easily divide the data into tenths ("deciles"), fifths ("quintiles"), etc.
- After percentiles, the most common division is into quarters. The k^{th} **quartile** (written Q_k) is the point below which k *quarters* of the data lies.
- So, the median is ______, the minimum is ______, the maximum is ______.
- We can re-express the range as ______.

Quartiles



The Inter-Quartile Range (IQR)

- We define the **Inter-Quartile Range** (or **IQR**) as the distance between the first and third quartiles:

$$IQR = Q_3 - Q_1$$

- Easily computed in R (use the `IQR()` function), but complicated to do by hand (different rules for quartiles depending on whether n is divisible by 4, by 2 but not by 4, is one more, or one less, than a multiple of 4, etc.)

ISTA 116: Statistical Foundations for the Information Age
Measures of Variability
The IQR and H-Spread

The Inter-Quartile Range (IQR)

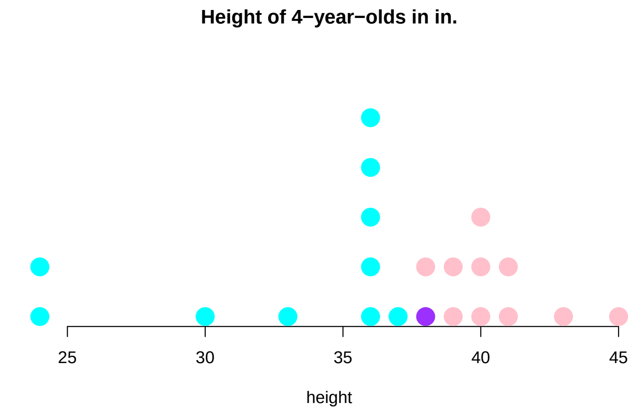
ISTA 116: Statistical Foundations for the Information Age
Measures of Variability
The IQR and H-Spread

The Hinges

- A closely related notion to quartiles is that of **hinges**: easier to compute by hand.
- Arguably obsolete due to computers, but still used for historical reasons.
- The **lower hinge** (or H_1) is defined by looking at the data *at or below the median*. It is the median of this subset.
- The **upper hinge** is the same idea, using the data *at or above the median* (or H_3)

ISTA 116: Statistical Foundations for the Information Age
Measures of Variability
The IQR and H-Spread

The Hinges



The H-spread

- The **H-spread** is defined the same way as the IQR, but with hinges rather than quartiles:

$$\text{H-spread} = H_3 - H_1$$

- Sometimes identical, almost always very close, to the IQR.

ISTA 116: Statistical Foundations for the Information Age
Measures of Variability
The IQR and H-Spread

The H-spread

- The **H-spread** is defined the same way as the IQR, but with hinges rather than quartiles:

$$\text{H-spread} = H_3 - H_1$$

- Sometimes identical, almost always very close, to the IQR.

The H-spread

- Let's find the H-spread of our rat survival time data:

40	62	77	88
94	109	128	136
137	152	152	160

The Five-Number Summary

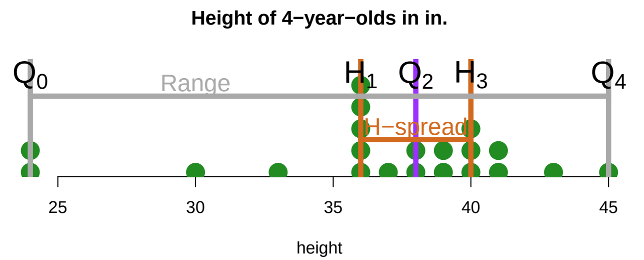
- The median and the hinges are very natural to report together to describe the center and spread of a distribution.
- Together with the minimum and maximum, they form the **five-number summary** of a univariate numeric distribution.

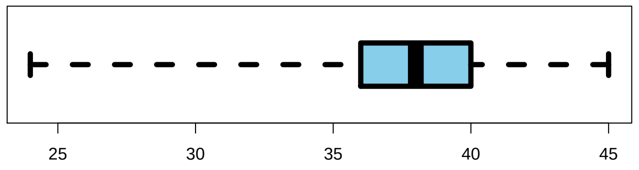
$$\begin{aligned}\text{Five Number Summary} &= (x_{(0)}, H_1, Q_2, H_3, x_{(n)}) \\ &= (Q_0, H_1, Q_2, H_3, Q_4) \\ &\approx (Q_0, Q_1, Q_2, Q_3, Q_4)\end{aligned}$$

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)
- Rules:
  1. Draw an axis
  2. Draw a rectangle (box) from H_1 to H_3
  3. Draw a line across the box at Q_2
  4. Draw lines (whiskers) extending outward from the box on both sides to $x_{(0)}$ and $x_{(n)}$.
- Note: R does something slightly more complicated with the whiskers.

Box-and-Whisker Plots



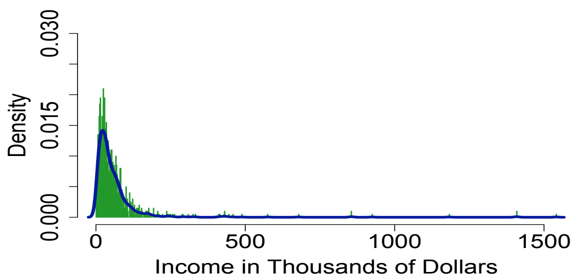


"Badly Behaved" Data

- We have seen that measures like mean and standard deviation are very sensitive to extreme values.
- In fact, these are only really representative for "well-behaved" distributions.
- What kinds of "bad behavior" are there?

Symmetry vs. Skew

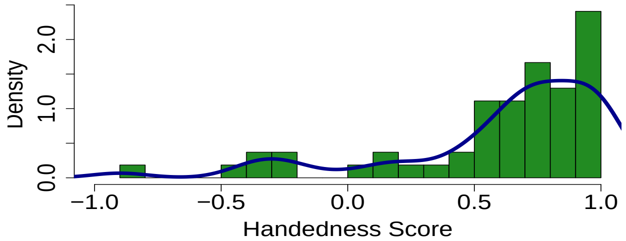
- We've already seen one example of a very badly behaved distribution.



- This distribution is characterized by extreme asymmetry, with a long **tail** going off to the right.
- We call this shape **right-skewed** (or sometimes **positively skewed**), after the side the tail is on.

Symmetry vs. Skew

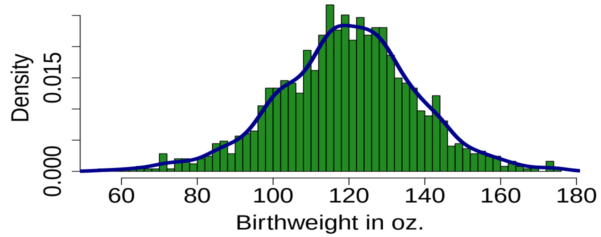
- Contrast this with our handedness data.



- Here, the tail goes off to the left, so we say the distribution is **left-skewed** (or **negatively skewed**).

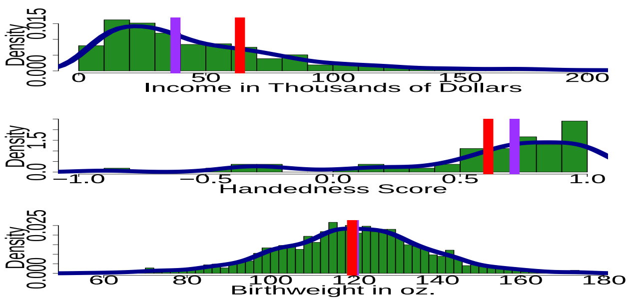
Symmetry vs. Skew

- In the absence of skew either direction, we just say the distribution is **symmetric**.



Symmetry vs. Skew

- The reason the skew is named for the tail-direction is because of what happens to the mean, relative to the median.

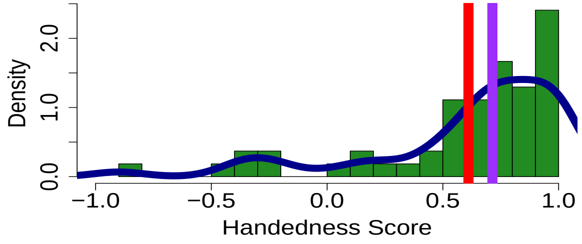


Symmetry vs. Skew

- Skew often arises when the underlying variable is structurally **bounded** (i.e., it has a minimum or maximum possible value), and there is data near the bound.
  - Example: Income bounded below by zero
  - Handedness scores can't go above $+1.0$
- Think of throwing some pudding at a wall: it piles up near the wall, and dribbles away forming a long, thin "tail"

Skew in Box Plots

- What will a box plot look like for a skewed distribution?

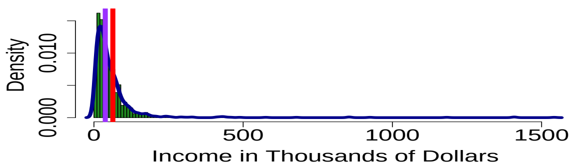


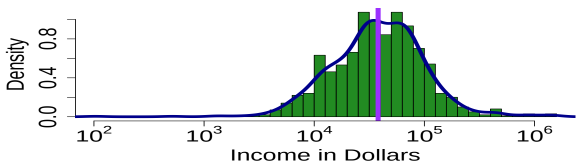
Variance-Stabilizing Transformations

- 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 $f(\cdot)$ 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:





Modality

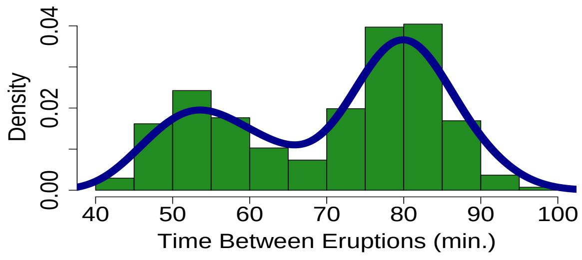
- We've defined the **mode** for discrete distributions: it is the value that appears most frequently.
- This definition breaks down for continuous variables, as there are no exactly repeated values (if there are, it's an artificial result of rounding).
- How might we generalize the concept?

Modality

- Most of the data we've looked at so far has a pretty unambiguous mode.
- Distributions like this, with a single peak, are called **unimodal**.
- Same naming conventions as with -*variate*: one, two or many:
  - **unimodal**: one mode
  - **bimodal**: two modes
  - **multimodal**: more than two modes

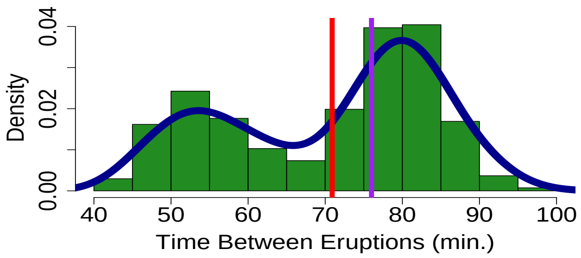
Modality

- Old Faithful isn't actually all that faithful.
- What is a "typical" waiting time?



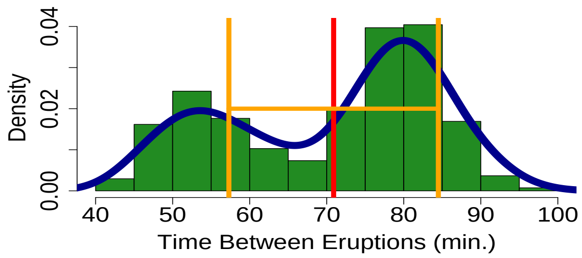
Modality

- What is a "typical" waiting time?



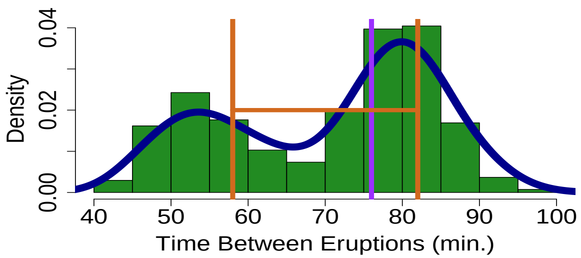
Modality

- How would we describe the spread?



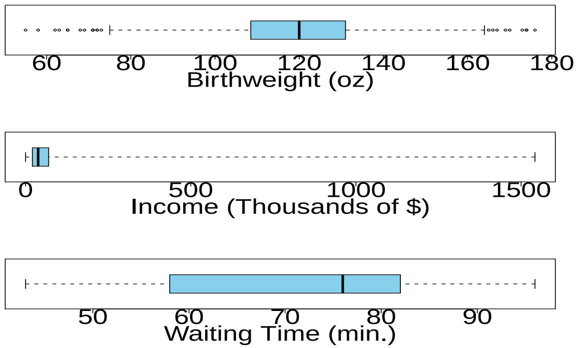
Modality

- How would we describe the spread?



Modality

- Compare the box plot to a "well-behaved" distribution like birth weight.
- Why is the box so much wider?

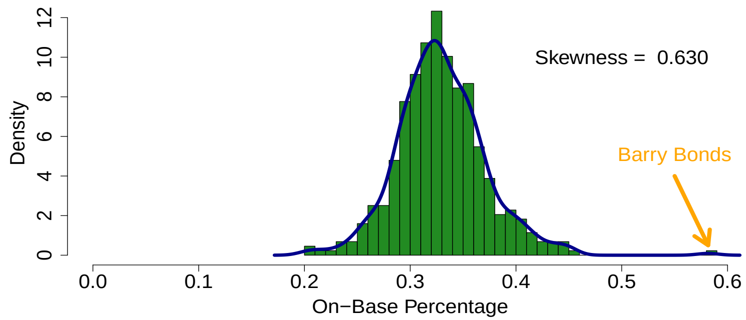


Outliers

- Skewness and bi-/multimodality can be important and meaningful features of a distribution.
  - E.g.: Inherent unevenness in income scale
  - E.g.: Bias in handedness, coupled with bounded scale
  - E.g.: Two discrete regimes in geyser behavior
- However, 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

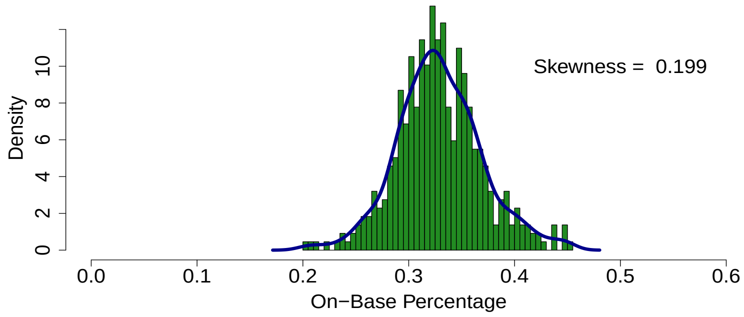
On-Base Percentage

- A common statistic for batters in baseball is *On-Base Percentage*
- Distribution of major-league hitters with at least 100 PA in 2002: (Why important to set PA cutoff?)



On-Base Percentage

- Distribution without Bonds

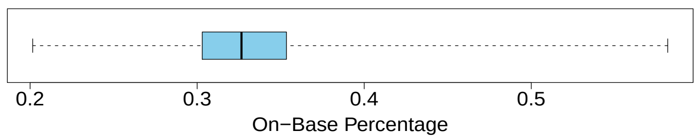


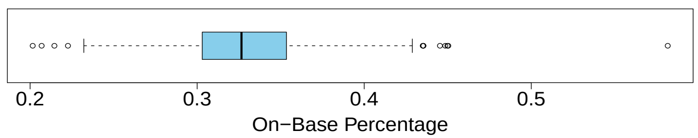
Visualizing Outliers

- R uses a modified procedure for box plots that displays potential outliers separately:
  1. Plot the box and median as normal (i.e., box from H_1 to H_3, with a line at Q_2)
  2. Calculate $W = 1.5 \times H$-spread
  3. Draw upper whisker from H_3 to the largest data point at or below $H_3 + W$
  4. Draw lower whisker from H_1 to the smallest data point at or above $H_1 - W$
  5. Plot any data points outside the whiskers individually

On-Base Percentage

- Compare the two procedures:





Next Time

- Begin discussing bi-/multivariate data