ISTA 116: Statistical Foundations for the Information Age

Measures of Variability

7 September 2011

Outline

- Web Quiz 2 due Friday
- Lab Assignment 2 posted, due next week in lab
  - Can do first half now
  - Second half after today's/Monday's class
- New d2l discussion forum for R issues

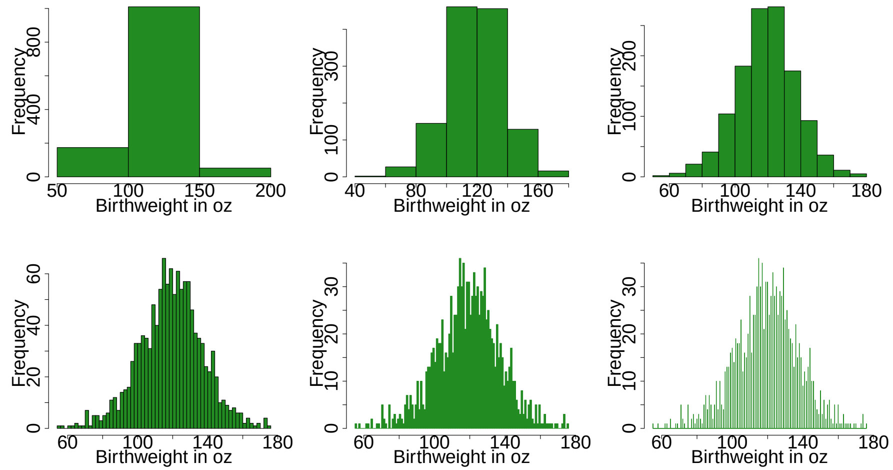


Figure: Histograms of Babies' Birth Weights

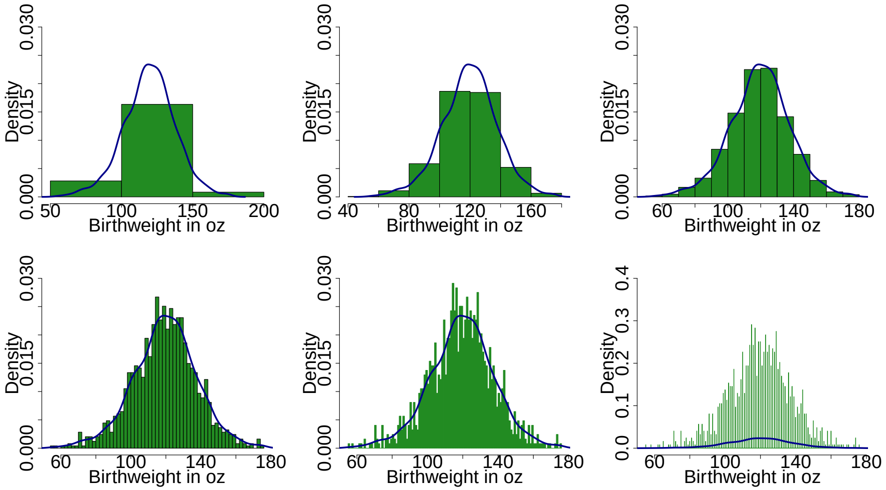


Figure: Densities of Babies' Birth Weights

Warmup

- Turn to the person next to you, and work together to produce your guess at:
  - A bar plot showing proportions of left-handed, ambidextrous and right-handed people.
  - A (rough) histogram with density estimate curve showing the quantitative measure of handedness (range is -1 for pure left to $+1$ for pure right).
  - On your histogram, show the mean, median and mode with vertical lines.

The Results...

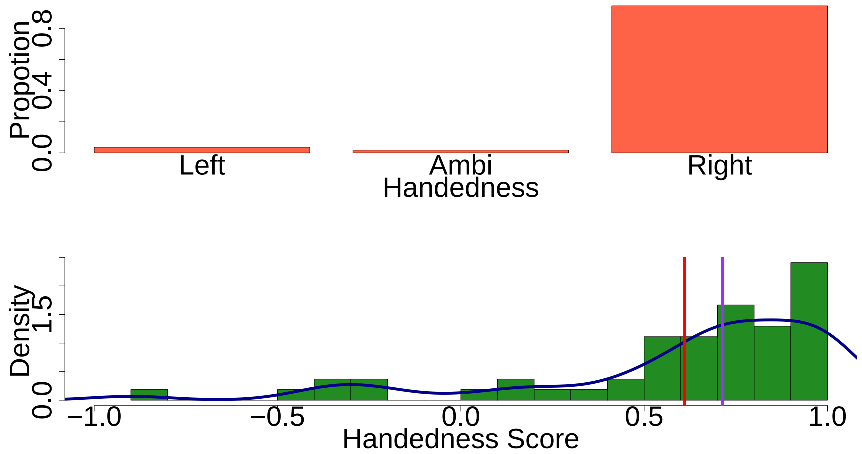


Figure: Handedness Distribution for Our Class

Measures of Central Tendency

- Most common value (the **mode**)
  - Advantages
    - Works for categorical data
    - Is a possible value
  - Disadvantages
    - Doesn't use much data
    - Often far from the center
- Value separating the data into halves (the **median**)
  - Advantages
    - In the middle of the data
    - **Robust** to extreme values
  - Disadvantages
    - Only sensitive to ranks
    - Discontinuous

Measures of Central Tendency

- Value at the "balance point" (the **mean**)
  - Advantages
    - Uses all the data
    - Intuitive for numeric data
  - Disadvantages
    - Not necessarily a possible value (discrete case)
    - Sensitive to extreme values
- Halfway between highest and lowest values (the **midrange**)
  - Advantages
    - Very easy to see on a graph
  - Disadvantages
    - Sensitive to *only* extreme values

A Simple Comparison

1^{st} Test	Stem	2^{nd} Test
	4	0
	5	7
	6	
7 5	7	0
8 7 5 2 0	8	2 8
	9	6
	10	0

Q: What's the difference between these two sets of exam scores?
A: Similar *centers*, but very different *spreads*.

Quantifying "Spread"

- How might we quantify the "spread" in a data set?

Quantifying "Spread"

- How might we quantify the "spread" in a data set?
- Some intuitions:
  - Difference between smallest and largest values (the **range**)
  - Average distance from the center (which center?)
  - Range of the central "bulk" of the data.

The Range

- Easy to compute

1st Test	Stem	2nd Test
	4	0
	5	7
	6	
7 5	7	0
8 7 5 2 0	8	2 8
	9	6
	10	0

The Range

- But very different distributions can have similar ranges.

1st Test	Stem	2nd Test
0	4	0
	5	7
	6	
7 5	7	0
8 7 5 2 0	8	2 8
	9	6
0	10	0

The Range

- And similar distributions can have very different ranges.

1st Test	Stem	2nd Test
0	4	
	5	
	6	0
7 5	7	0 5 7
8 7 5 2 0	8	0 2 5 7
	9	
0	10	0

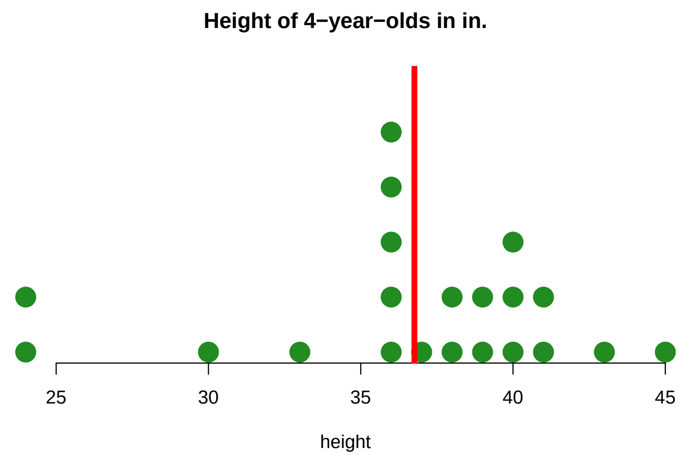
- This is because the range (like the midrange) is only sensitive to extreme values.

Deviations

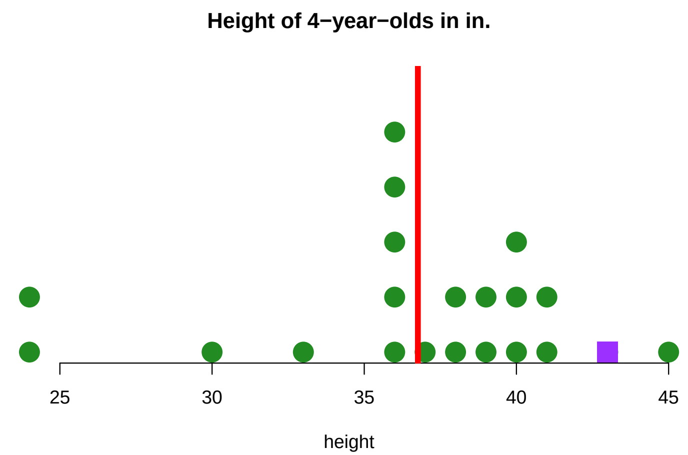
- Rather than simply measuring the distance between the extremes, we can develop measures based on distance from “center”.
- Which center? The mean is the logical choice, since it is the only one that uses specific numeric values.
- For *each* data point, its **deviation score** is its “distance” from the mean.

$$\text{Deviation}_i = x_i - \bar{x}$$

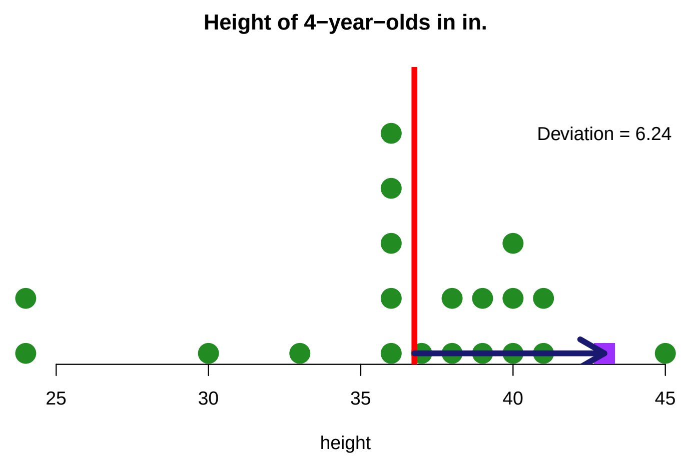
Deviations



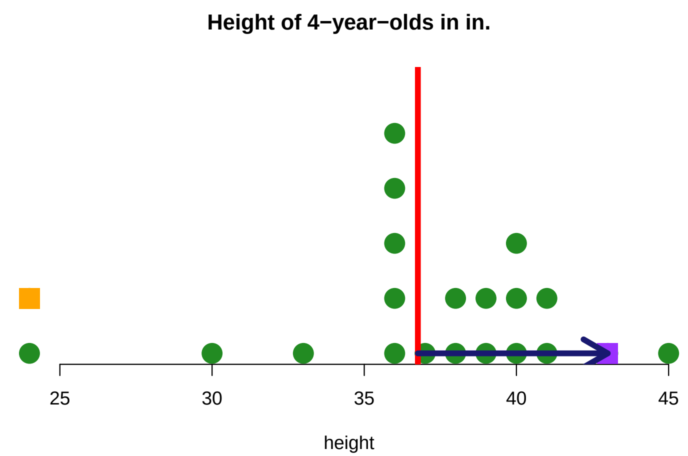
Deviations



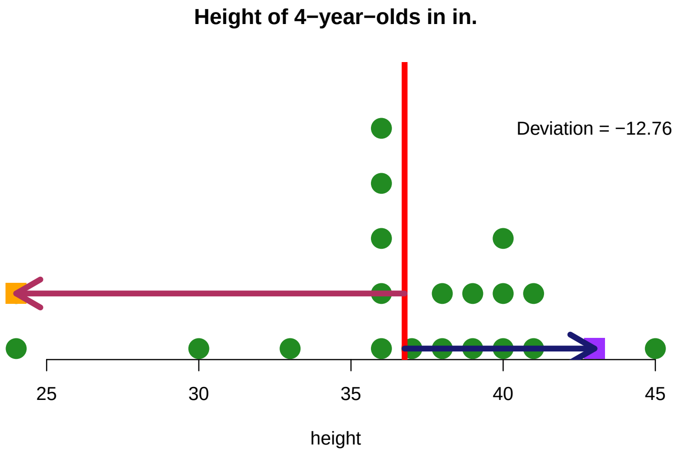
Deviations



Deviations



Deviations



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

Variance

- Problem with "average deviation": always zero, because of the definition of the mean.
- Could use "average absolute deviation", but absolute value has inconvenient mathematical properties.
- Instead, we use "average squared deviation", which is the **variance**.

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

Standard Deviation

- The problem with variance (s^2) as a measure of spread is that it's in squared units.
- No problem: just take the square root.
- $s = \sqrt{s^2}$ is the **standard deviation**

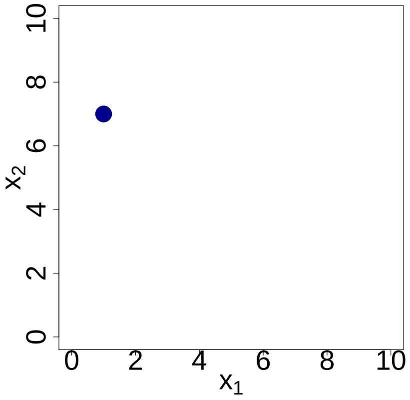
$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} \text{Deviation}_i^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Geometric Analogy

- Suppose we have just two data points: 1 and 7
- $\bar{x} = 4$

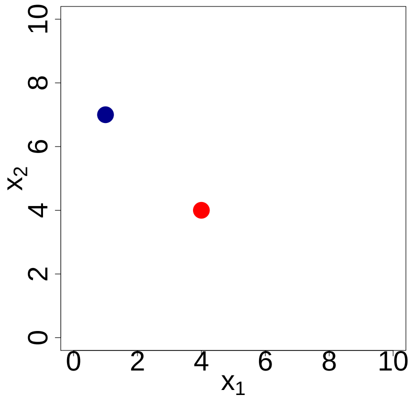
Geometric Analogy

- Imagine plotting one value on the x-axis, and one on the y-axis.



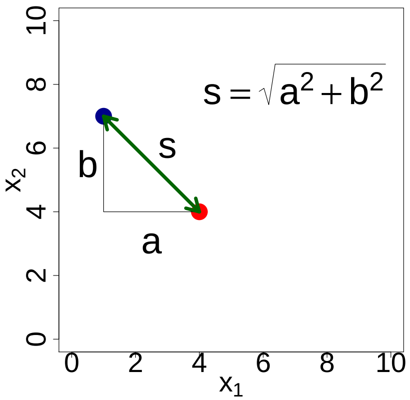
Geometric Analogy

- Now plot the data set with the same mean, but no variability (i.e. every value equal to the mean).



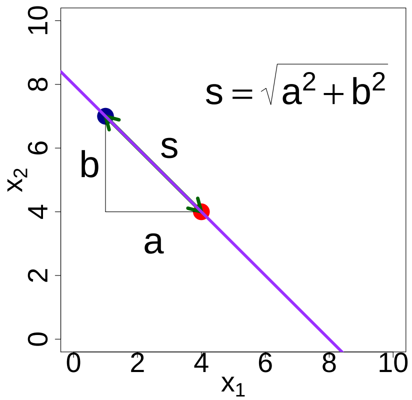
Geometric Analogy

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



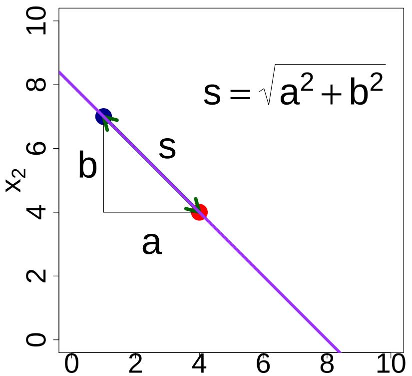
Geometric Analogy

- 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.

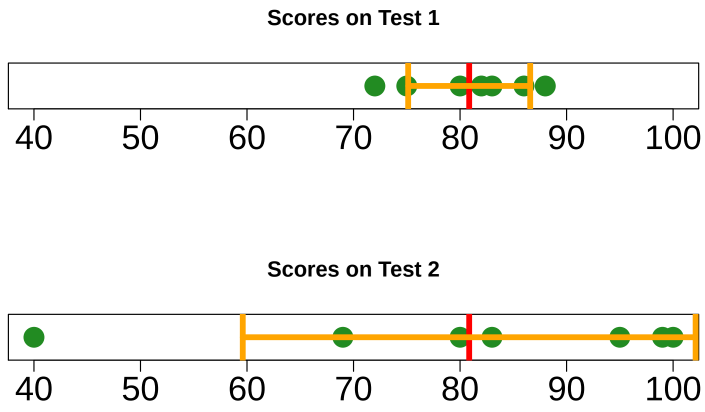


Geometric Analogy

- 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

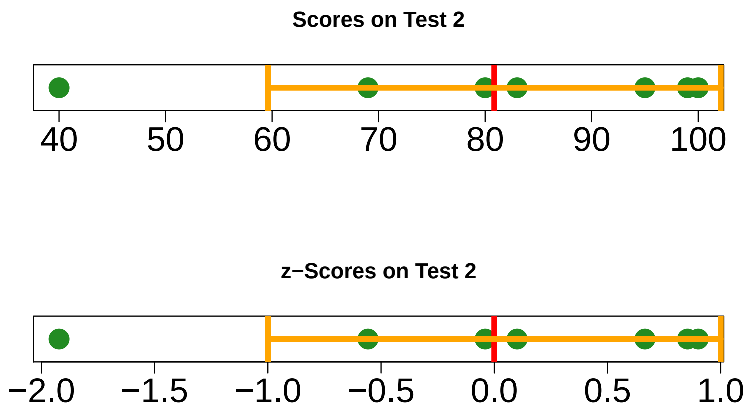
- A common application of standard deviation is as a way to measure how far a data point is from the mean, on a scale that is *independent of units*.
- By dividing each individual deviation score by the standard deviation, we obtain a **z-score** for that data point.

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

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

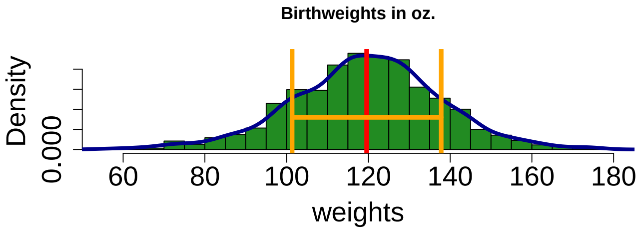
z-scores

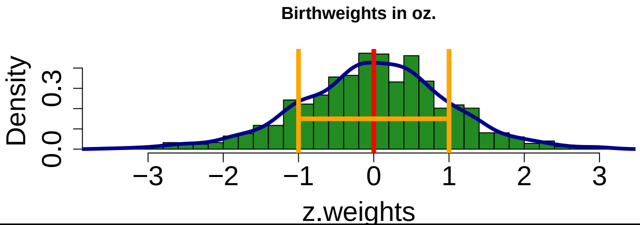
- We can compute z-scores for the whole data set, and see their distribution.



z-scores

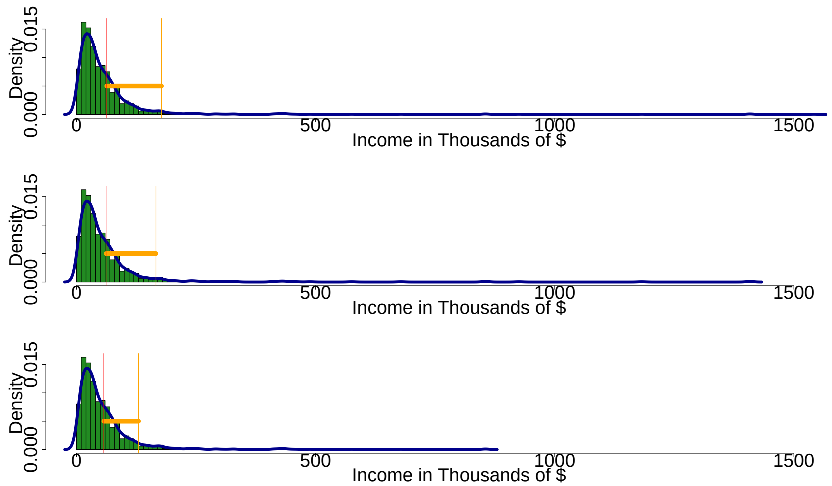
- We can compute z-scores for the whole data set, and see their distribution.





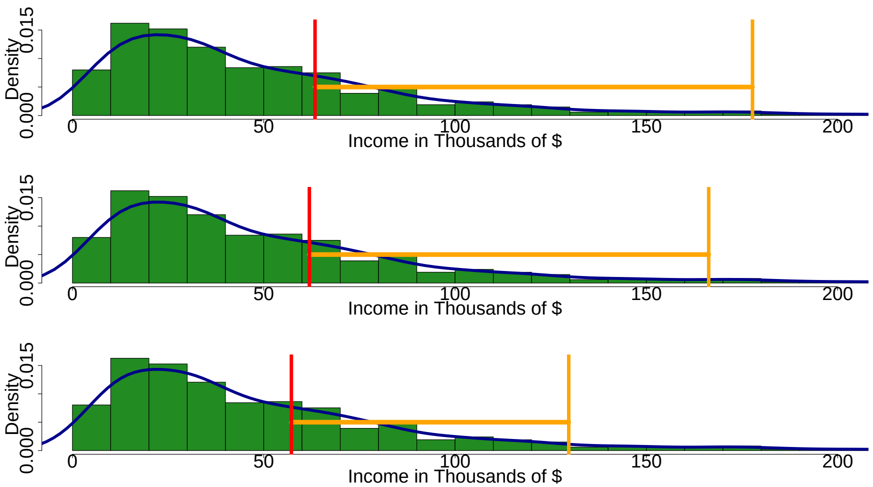
Problems with s and s^2

- 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.

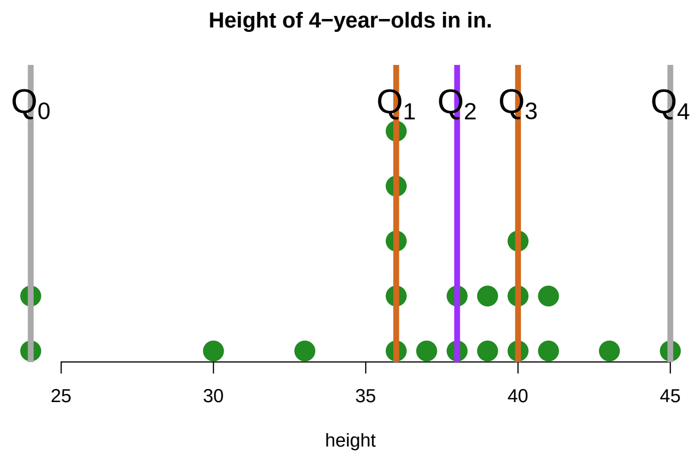
Quantiles

- 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*.

Quantiles

- 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.)

The Inter-Quartile Range (IQR)

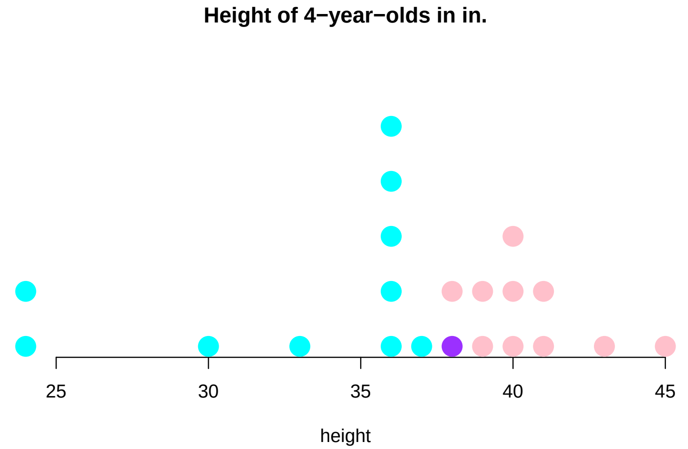
ISTA 116: Statistical Foundations for the Information Age
Measures of Variability
The IQR and H-Spread

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



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

- 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.