Lecture 6

Variance and standard deviation
Cumulative distributions
The normal distribution

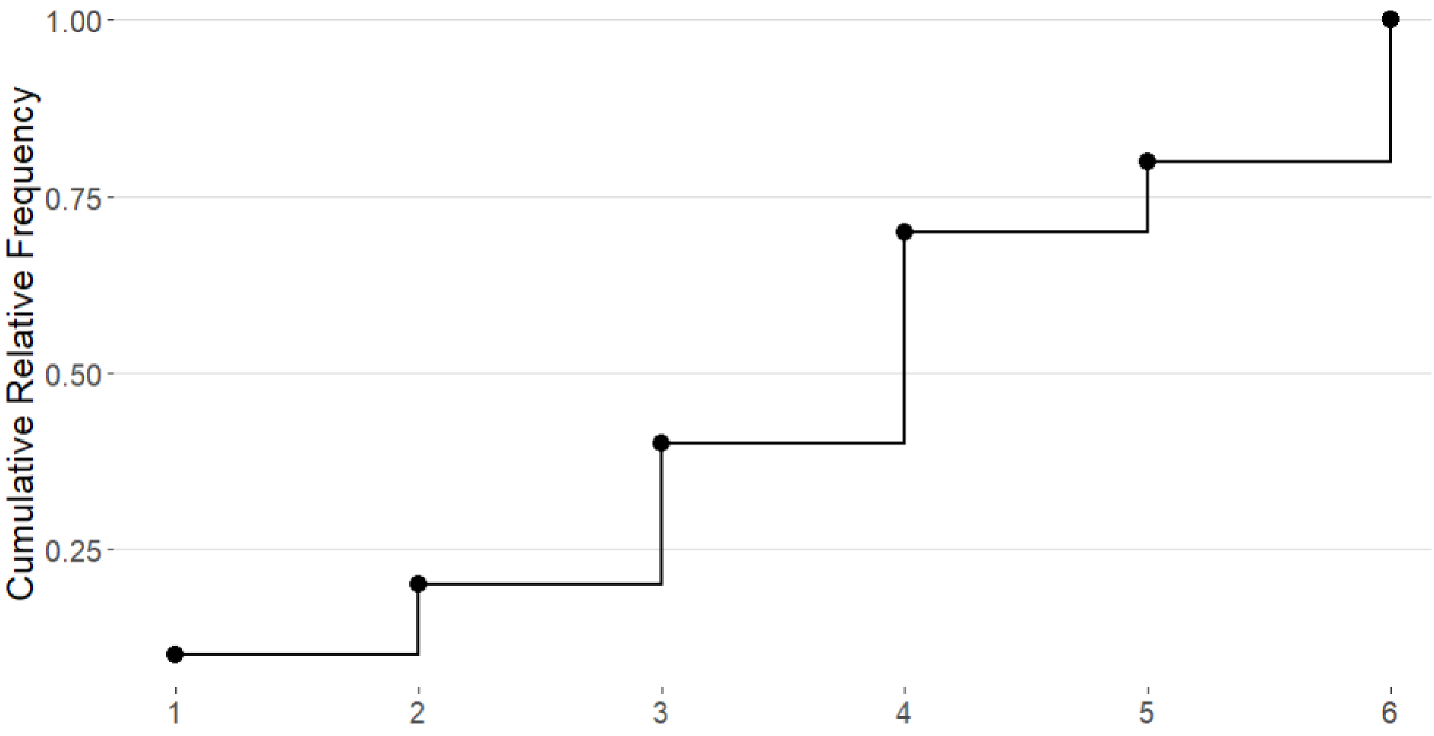
Review

- Percentiles
- Range and IQR
- The five number summary and boxplot

Cumulative Distributions

- A **cumulative distribution** shows the relationship between the value of a variable and the **cumulative relative frequency**
- We represent the cumulative distribution using a step function
- Data = 1,2,3,3,4,4,4,5,6,6

x	$F(x)$	$RF(x)$	$CRF(x)$
1	1	0.1	0.1
2	1	0.1	0.2
3	2	0.2	0.4
4	3	0.3	0.7
5	1	0.1	0.8
6	2	0.2	1.0



Cereal	Sodium	Sugar	Type
Frosted Mini Wheats	0	11	A
Raisin Bran	340	18	A
All Bran	70	5	A
Apple Jacks	140	14	C
Cap'n Crunch	200	12	C
Cheerios	180	1	C
Cinnamon Toast Crunch	210	10	C
Crackling Oat Bran	150	16	A
Fiber One	100	0	A
Frosted Flakes	130	12	C

Cereal	Sodium	Sugar	Type
Froot Loops	140	14	C
Honey Bunches of Oats	180	7	A
Honey Nut Cheerios	190	9	C
Life	160	6	C
Rice Krispies	290	3	C
Honey Smacks	50	15	A
Special K	220	4	A
Wheaties	180	4	A
Corn Flakes	200	3	A
Honeycomb	210	11	C

Finding Percentiles from Cumulative Distributions

lower half middle upper half

- Data = 0, 1, 3, 3, 4, 4, 5, 6, 7, 9, 10,11,11,12,12,14,14,15,16,18

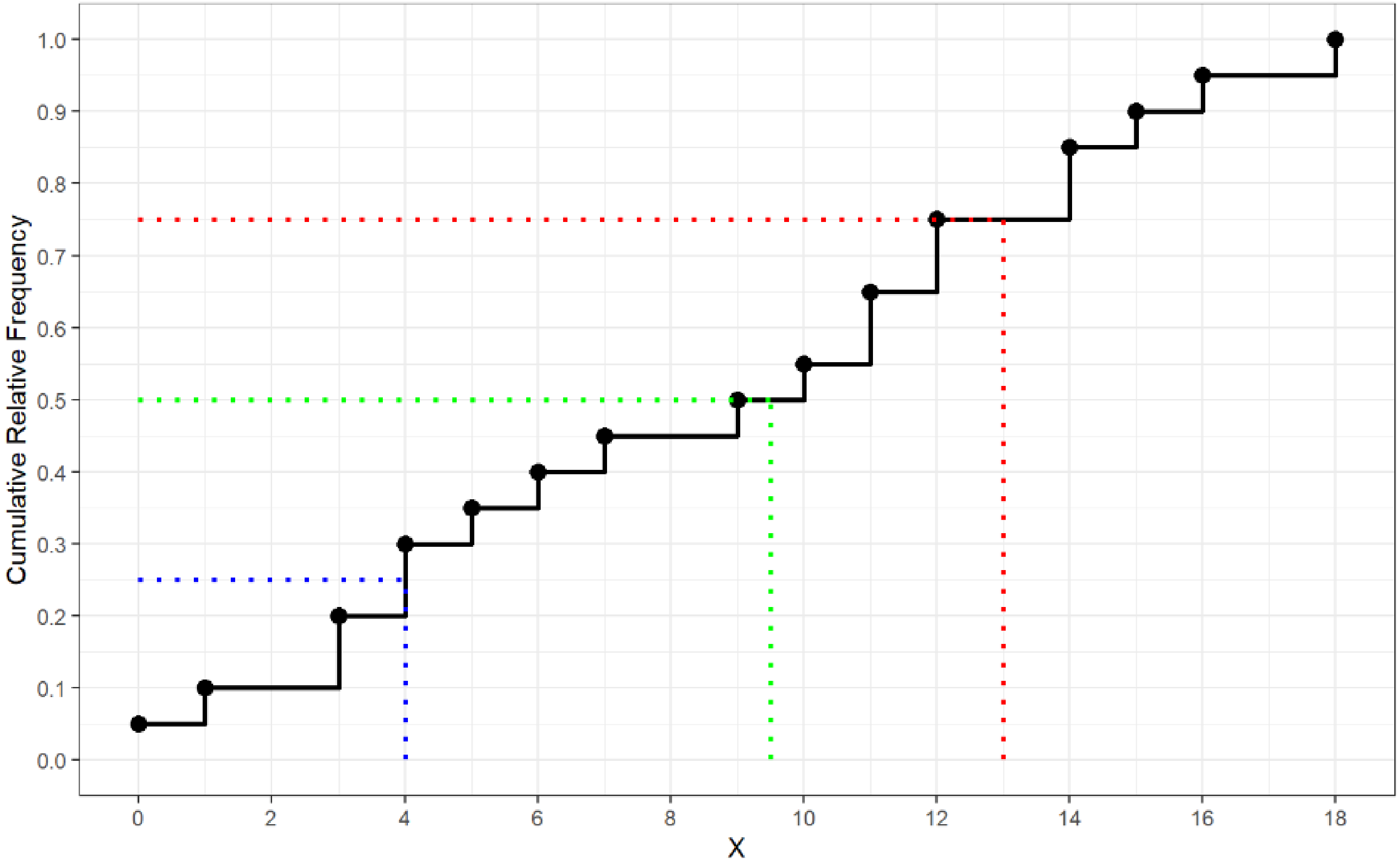
$Q1 = \frac{4+4}{2} = 4$

$Q2 = \frac{9+10}{2} = 9.5$

$Q3 = \frac{12+14}{2} = 13$

What is the IQR?

$\text{IQR} = 13 - 4 = 8$



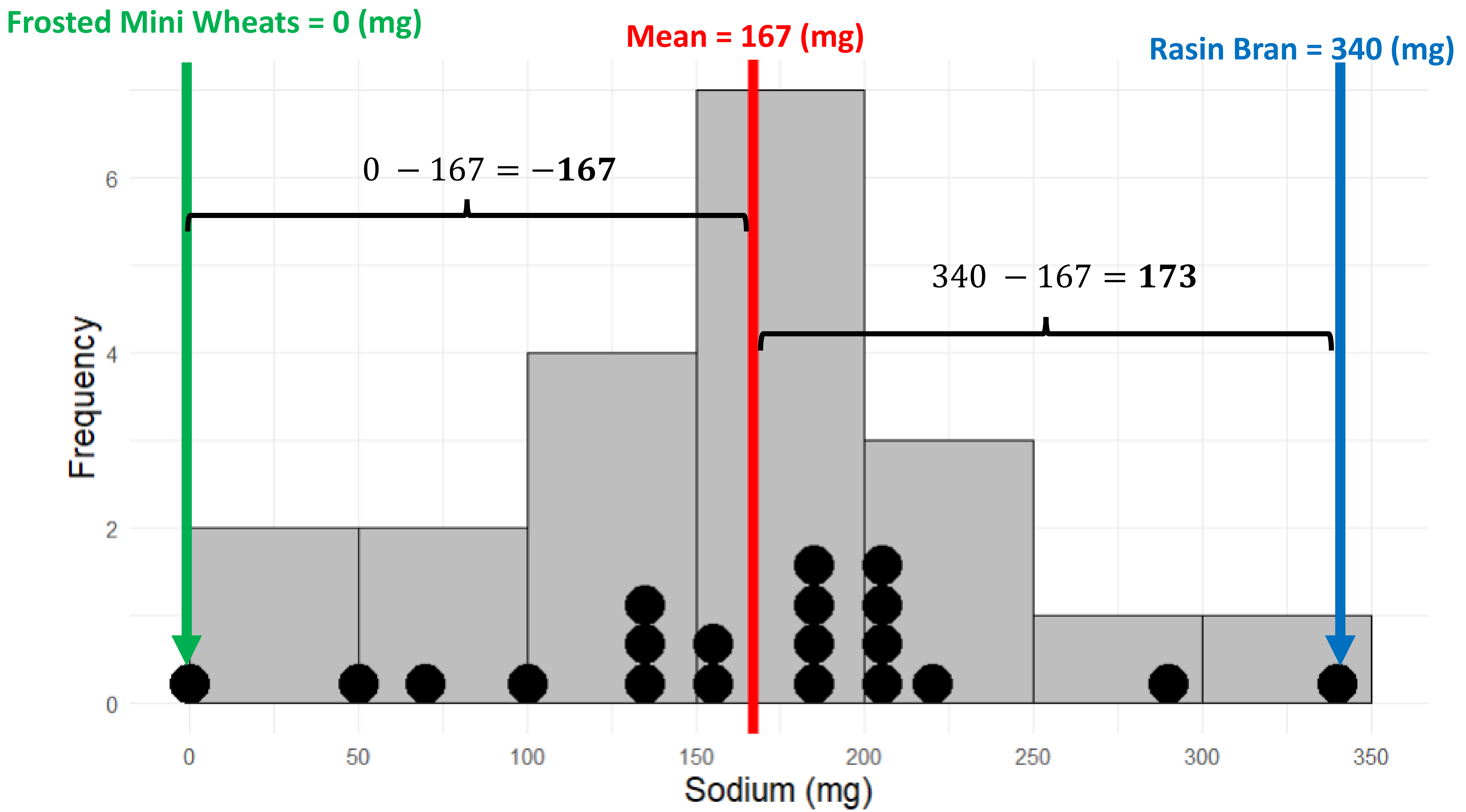
Measures of Spread: Deviation

- A better measure of variability that uses *all* the data is based on **deviations**

- **deviations** are the distances of each value from the mean of the data:

  Deviation of an observation $x_i = (x_i - \bar{x})$

- Every observation will have a deviation from the mean



Frosted Mini Wheats = 0 (mg)
Mean = 167 (mg)
Rasin Bran = 340 (mg)
0 − 167 = −167
340 − 167 = 173
Frequency
0
2
4
6
0
50
100
150
200
250
300
350
Sodium (mg)

Measures of Spread: Variance

- The sum of all deviations is zero. $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
- We typically use either the **squared deviations** or their **absolute value**

  Squared deviation of an observation $x_i = (x_i - \bar{x})^2$
- The **Variance** of a distribution is the <u>average</u> squared deviation from the mean

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- The sum $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is called the sum of squares

Measures of Spread: Standard Deviation

- Since the variance uses the squared deviation, we usually take its square root called the **standard deviation**

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

- The standard deviation represents (roughly) the average distance of an observation from the mean

- The greater s is the greater the variability in the data is

- We denote the population parameter for the variance and standard deviation using σ for s and σ^2 for s^2

Try it out: Computing s and s^2

Consider the following sample of 5 observations of the height of colleges students at the University of Idaho

- Data = 61,62,62,68,75
- Mean = 65.6

- What if we observe another student who has a height of $x = 92$ inches. How does including this observation change our estimate s

Why divide by $n - 1$?

- We divide by $n - 1$ because we have only $n - 1$ pieces of independent information for s^2

- Since the sum of the deviations must add to zero, then if we know the first $n - 1$ deviations we can always figure out the last one

- Ex.) suppose we have two data points and deviation of the first data point is $x - \bar{x} = -5$
  - Then the deviation of the second data point has to be 5 for the sum of deviations to be zero.

End of Material For Exam 1