Math 3339

Section 27204
MWF 10-11:00am AAAud 2

Bekki George
bekki@math.uh.edu
639 PGH

Office Hours:
M & Th noon – 1:00 pm & T 1:00 – 2:00 pm
and by appointment

Popper 02

1. Which of the following is quantitative data?

   a. Hair color

   b. Letter grade for a class

   c. Rating of movie on a scale of 1 to 5

   d. Numerical grade on a test

   ~~e. None of these~~

2. In #1, the quantitative data is

   a. Discrete

   b. Continuous

3. The command in R to find the mean is:

   a. mean

   b. average

   c. avg

   d. sum

   e. none of these

4. Suppose we were looking at salaries for a small company. Most employees make the same amount per year but the CEO makes 10 times that amount. Which is larger:

   a. mean

   b. median

Frequency Distributions

A frequency distribution is a tabular summary of data showing the frequency (or number) of items in each of several non-overlapping classes. The relative frequency of a class is the fraction or proportion of the total number of data items belonging to the class.

Table 1.1 Frequency Distribution for Hits in Nine-Inning Games

Hits/Game	Number of Games	Relative Frequency	Hits/Game	Number of Games	Relative Frequency
0	20	.0010	14	569	.0294
1	72	.0037	15	393	.0203
2	209	.0108	16	253	.0131
3	527	.0272	17	171	.0088
4	1048	.0541	18	97	.0050
5	1457	.0752	19	53	.0027
6	1988	.1026	20	31	.0016
7	2256	.1164	21	19	.0010
8	2403	.1240	22	13	.0007
9	2256	.1164	23	5	.0003
10	1967	.1015	24	1	.0001
11	1509	.0779	25	0	.0000
12	1230	.0635	26	1	.0001
13	834	.0430	27	1	.0001
				19,383	1.0005

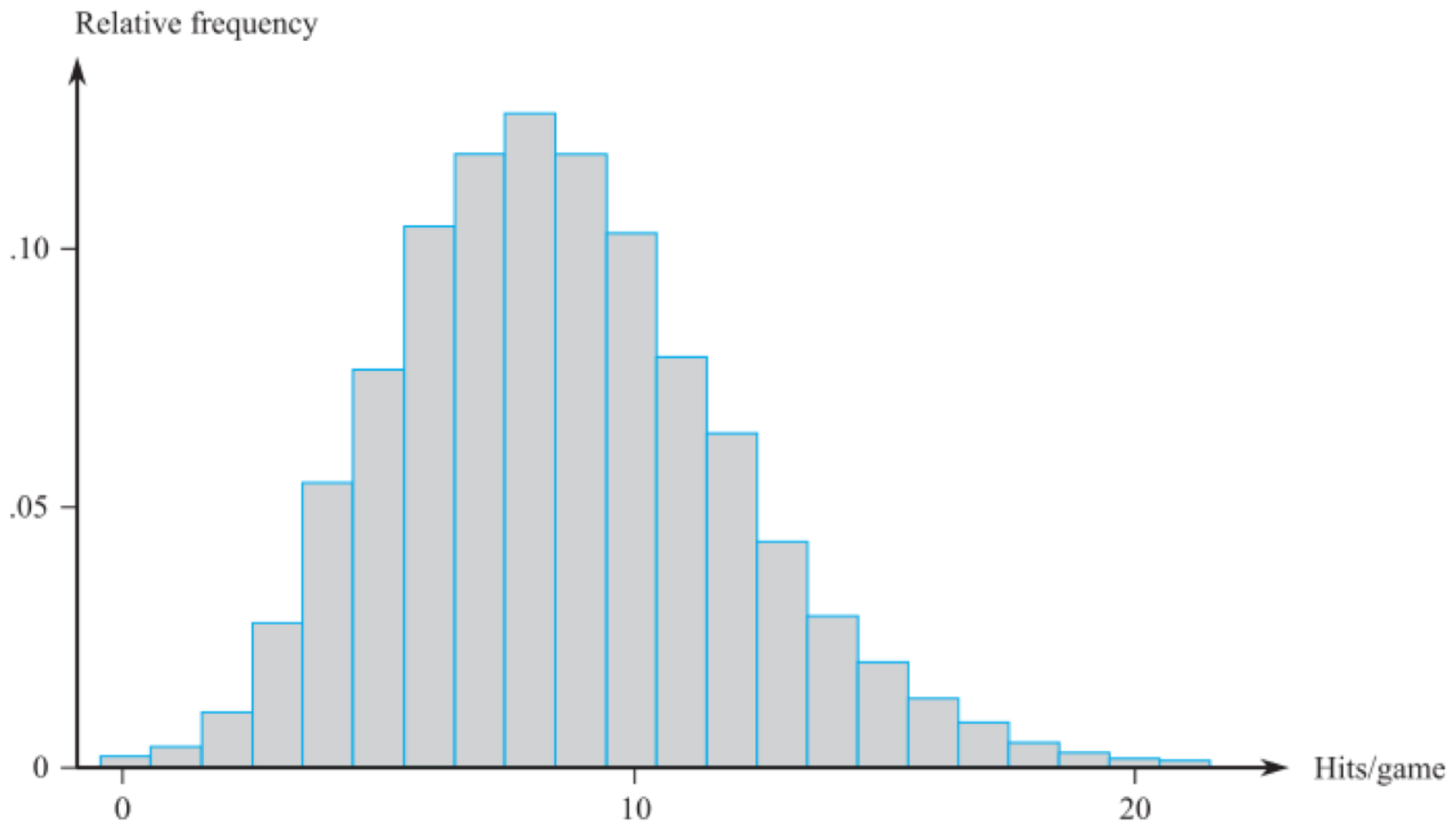


Figure 1.7 Histogram of number of hits per nine-inning game

Cumulative Frequency Histograms

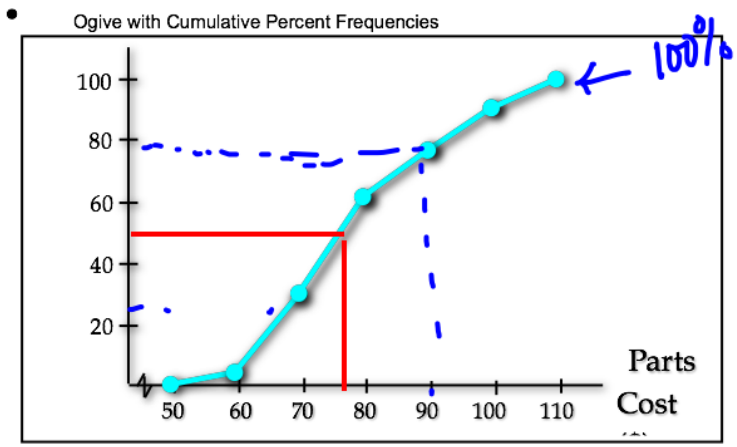
The cumulative frequency distribution shows the number of items with values less than or equal to the upper limit of each class.

The cumulative relative frequency distribution shows the proportion of items with values less than or equal to the upper limit of each class.

A **cumulative frequency plot** of the percentages (also called an **ogive**) can be used to view the total number of events that occurred up to a certain value.

Example: Here is an ogive for Hudson Auto Repair's cost of parts sold:

Example: Hudson Auto Repair



Where is the median of this data?

Sec 2.2 – Variability

Measures of Variability

Dispersion (*spread*)

1. The simplest way to measure dispersion is ***range***. This is the difference between the smallest and largest measurements. max − min
   Drawbacks of range: sensitivity to outliers

2. Another method is *interquartile range*, $IQR = Q_3 - Q_1$. This is not sensitive to outliers, but still has some drawbacks as a measure of dispersion. middle 50%

3. The most common measure is **sample standard deviation**. Roughly speaking, standard deviation is the average distance values fall from the mean (center of graph).
   The *sample variance* is defined as

$$s^2 = \frac{1}{n-1}\left[(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

   and the *sample standard deviation* is given by s, the square root of the sample variance.

(Note: this is different from the *population variance*)

pop. var. →
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu(x)^2,$$

Sample Var. →
$$s^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2).$$

4. The **coefficient of variation** measures relative variability.

$$cv(x) = \frac{sd(x)}{\mu(x)}$$

This is used for variables that have only positive values.

Let's compute the sample standard deviation of our measurements "height" data:

66, 68, 63, 71, 68, 69, 65, 70, 73, 67 $\quad \bar{x} = 68$

$$s^2 = \frac{1}{10-1}\left[(66-68)^2 + (68-68)^2 + (63-68)^2 + (71-68)^2 + \cdots + (67-68)^2\right]$$

avg. squared distance from $\bar{x}$

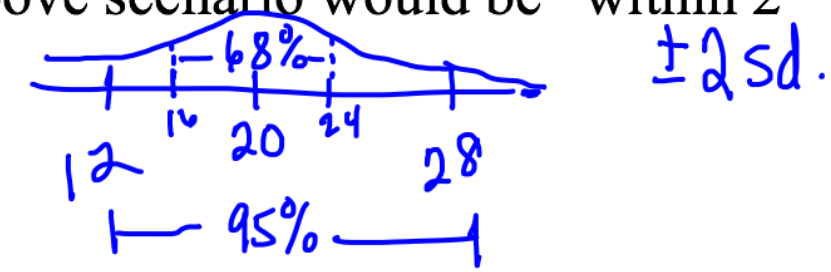
$$s = \sqrt{s^2} = 2.94392$$

> sd(height)

Distance from the mean is sometimes measured in standard deviations. For example, if $\bar{x} = 20$ and $s = 4$ then a measurement of 12 would be "2 standard deviations from the mean".

$12 = 20 - 2(4)$

What interval of measurements from the above scenario would be "within 2 standard deviations from the mean"?



$(12, 28)$

Within 1.5 standard deviations?

$20 \pm 1.5(4)$

$20 - 6, \quad 20 + 6$

$(14, 26)$

Calculated Standard Deviation is a measure of Variation in data

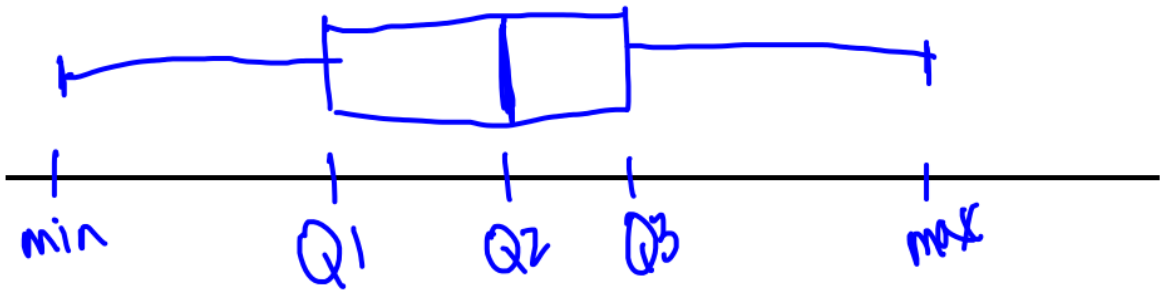
Sample Data Set	Mean	Standard Deviation
100, 100, 100, 100, 100	100	0
90, 90, 100, 110, 110	100	10
30, 90, 100, 110, 170	100	50
90, 90, 100, 110, 320	142	99.85

Box-and-Whisker Displays (Boxplots)

(median)
min Q1 Q2 Q3 max

Making a Boxplot from the Five Number Summary

1. Order the values in the data set in ascending order (least to greatest).
2. Find and label the median.
3. Of the lower half (less than the median—do not include), find and label Q1.
4. Of the upper half (greater than the median—do not include), find and label Q3.
5. Label the minimum and maximum.
6. Draw and label the scale on an axis.
7. Plot the five number summary.
8. Sketch a box starting at Q1 to Q3.
9. Sketch a segment within the box to represent the median.
10. Connect the min and max to the box with line segments.



Boxplot—5 Number Summary

Computersx1000

250

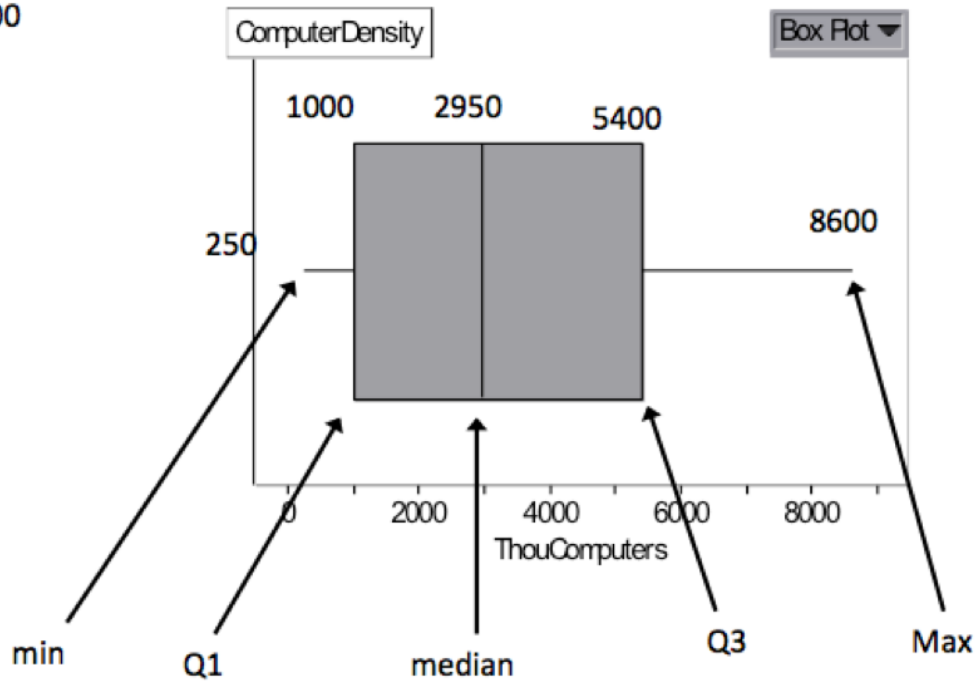
1000

2400

3500

5400

8600



IQR = Q3 – Q1 = 5400 – 1000 = 4400

$1.5(IQR) = 1.5(Q3-Q1)$

Calculating Outlier BOUNDARIES

Follow the formula (Q1 – 1.5 (IQR), Q3 + 1.5 (IQR)) Hint: you need to know what Q1 and Q3 are numerically.

Steps:

1) Find Q1 and Q3.
2) Calculate the Interquartile Range, where IQR = Q3 – Q1
3) Multiply IQR by 1.5.
4) Subtract 1.5 (IQR) from Q1, this is the lower bound.
5) Add 1.5 (IQR) to Q3, this is the upper bound.
6) Write outlier boundaries in interval notation, (lower bound, upper bound).

Popper 02

5. How do you find the IQR?

Inter quartile range

a. Q3-Q1

b. 1.5(Q3-Q1)

c. Q1-Q3

d. 1.5(Q1-Q3)

6. The values of the minimum, Q1, Q2, Q3 and the maximum make up what is called our

a. percent values

b. five number summary

c. quartiles

d. none of these

Example: 79, 81, 83, 86, 87, 88, 89, 90, 91, 95, 108, 111

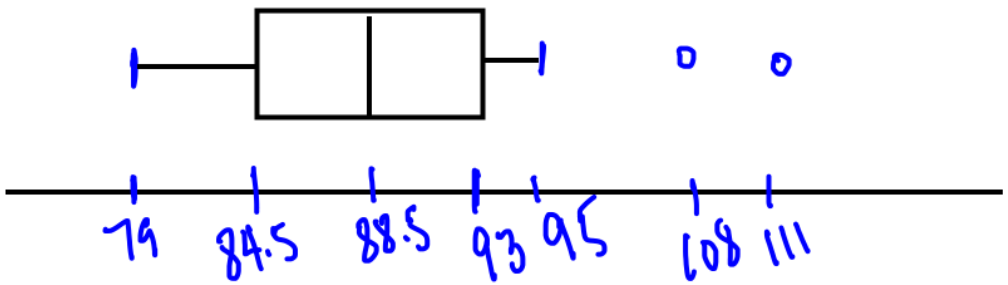
79	84.5	88.5	93	111
min	Q1	med	Q3	max

Steps:

1) Q1 = 84.5 and Q3 = 93.
2) Calculate the Interquartile Range $IQR = 93 - 84.5 = 8.5$
3) Multiply IQR by 1.5 $1.5(8.5) = 12.75$
4) Subtract 1.5 (IQR) from Q1. $84.5 - 12.75 = 71.75$
5) Add 1.5 (IQR) to Q3. $93 + 12.75 = 105.75$
6) Write outlier boundaries in interval notation, $(71.75, 105.75)$.

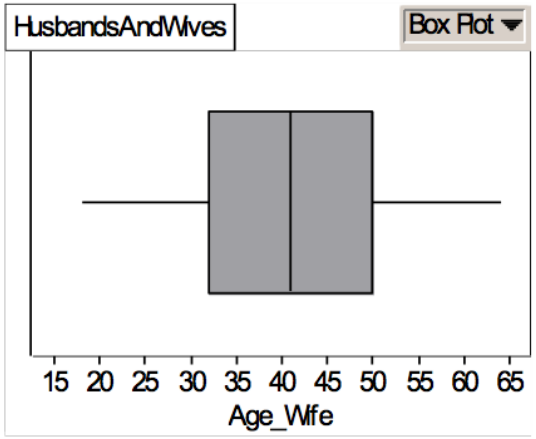
NOW…are there any data that falls OUTSIDE the boundary interval?

108 & 111 are outliers



Describing a distribution (CUDS – Center, Unusual Features, Dispersion, Shape)

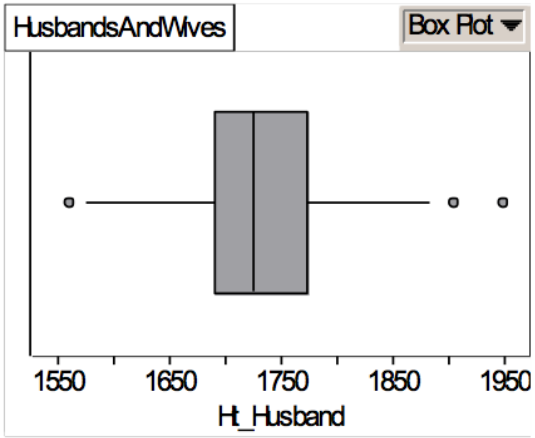
CENTER:



UNUSUAL FEATURES:

DISPERSION:

SHAPE:



CENTER:

UNUSUAL FEATURES:

DISPERSION:

SHAPE:

>boxplot(height)
>boxplot(height,horizontal=TRUE)