Numerical Summaries

Recall:
A variable has values and frequencies.
A variable's distribution describes the pattern of frequencies.

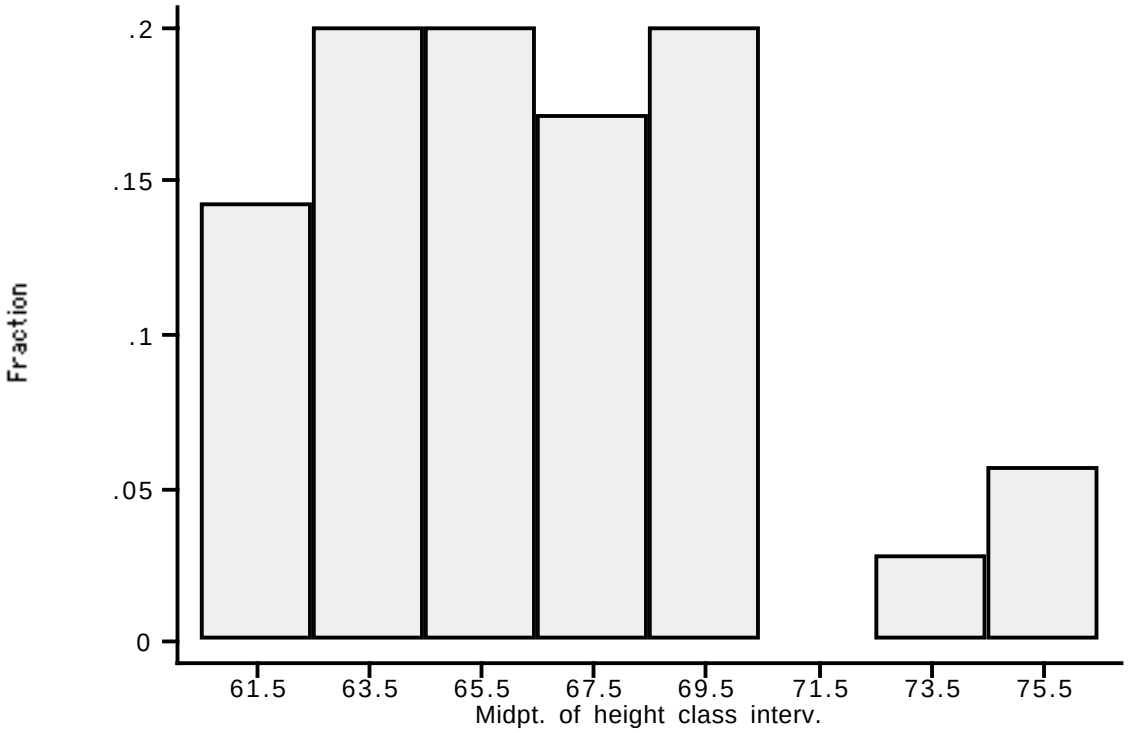
Roughly speaking, there are two aspects of a distribution we wish to summarize: the center and the spread. We'll talk first about center. There are other aspects, of course, but in many situations these are enough.

I. Measures of Center

What's a "typical" height of a UCLA college student? What's a "typical" height of the driver of a Ford Explorer? What's the typical weight of someone who rides an elevator? If you invite 100 people to your party, how many can you expect to come? How many glasses of beer can you expect to serve?

These are all questions about a variable. We could, in theory at least, collect data to begin to understand what values and frequencies we should see for these variables. For example, we might, over time, have many parties, and invite 100 people each time, and record how many come to the party. Or maybe we can include other party-throwers who invited 100 people and see how many came to their party.

These questions all ask, more or less, not about the shape of the distribution of the values, but about the "typical" value. This is a vague word, and as we shall see there are many ways of answering. Below is a histogram of the heights in this class. What is a "typical" height?


Fraction
.2
.15
.1
.05
0
61.5
63.5
65.5
67.5
69.5
71.5
73.5
75.5
Midpt. of height class interv.

You have several choices in answering such a question. One common approach is to take the number that occurs most frequently. This is called the **mode**. However, this doesn't work here because three bins are of equal height. This is a common problem with the mode.

Another popular approach is to take a number somewhere in the center of the distribution. There are two popular choices for the "center".

The average

To find the average, add all the items in your list and divide by the number of items. The average height, for these data, is 66.63 = (74 + 70 + 64....)/35 (only 35 people turned in a card.)

The symbol for the average is xbar (a lower-case x with a bar over it.)

The average has an interesting physical/geometric property. If you made the histogram of heights out of wooden blocks and tried to balance them on a see-saw, the balancing point would be at the average. This means that for symmetric distributions, the average is in the center.

However, if the distribution is skewed, then the average is somewhat off-center.

This highlights an important feature of the average (although it is a nuisance); it is very affected by outliers. Recall that outliers are extreme values (either extremely big or extremely small compared to the rest of the values.) Think in terms of the balancing point. Where is the balancing point of this list:

2, 2, 4, 4
Right in the center: 3.

Now how about:
2,2,4,100

The balancing point is now at 27. (108/4).

The moral of the story: if the data have outliers, or are skewed, the average may not be a good representation of the "typical" value.

Thus, it is very important to make a picture of the distribution BEFORE computing an average and claiming that it is a good representation.

Skewed Distributions:

Certain variables typically have skewed distributions. For example, right-skewed distributions typically have a lower bound and no upper bound. Left-skewed are the other way around: there's an upper-bound but no lower bound.

For example, an income cannot be less than 0, but there is no upper limit. For this reason, distributions of incomes are often skewed right. Usually, therefore, the average will not be a good summary since it will strike most people as being "too" large.

Try to guess what the general shape of these distributions is most likely to be:

- Number of hot-dogs a person can eat in 1 minute.
- Height of U.S. women
- Lifetime of U.S. residents

The Median

The median is another popular measure of center. Simply put, it is the "middle" value. Here's how to find it:
1) Sort the numbers in your list from smallest to largest.
2) If the list has an odd number of items, the median is the value in the exact middle of the list.
3) If the list has an even number of items, the median is the average of the two middle values.

Ex1: 2,3,3,6,7 There are 5 items, the middle one is 5, so the median is 5.

Ex2: 2,3,3,6,7,7 There are 6 items, the middle two are 3,6 and the average of these is $(3+6)/2 = 4.5$

The median divides the distribution in half. Precisely half of the values are below, half above.

In a relative frequency histogram, this means that the area to the left of the median (and to the right) is 50%.

The median height is 66. This falls somewhere in the 4^{th} bin. Note that the total area of the first four bins is about 54%. The reason that it is bigger than 50% is that there are some values in the fourth bin bigger than 66.

Resistance

Here's why the median is sometimes a useful measure of center. Let's see what happens when we change the values on a list as we did for the average.

2, 2, 4, 4
xbar = 3
median = 3 ((2+4)/2 = 3)

2,2,4,100
xbar = 27
median = 3

The technical term is that the median is resistant to spread. This means that the ends of a list can be made more extreme (more spread out) and the median will not be affected. This also means that the median is not affected by skewed distributions.

Comparison between average and median

Note that if the distribution is symmetric, these two are equal. The median is in the exact center because it always is. The average is in the exact center because it is the balancing point.

Now stretch out the right tail of the distribution. To maintain the "balancing point", you must move the average to the right. But the median stays put; it is unaffected. The lesson:

For a right-skewed distribution, the average is greater than the median.

For the same reason, the average is less than the median for a left-skewed distribution.

Moral: If the distribution is not symmetric, the median is a better measure of the center than the average.

So why use the average at all? First, for mathematical reasons, it turns out that the average is "easier" in many contexts. It turns out, as you'll see later, it has some very nice properties. (These mostly come into play when we want to make inferential conclusions.)

Also, a surprising number of real-life situations do have symmetric distributions. Not all, but many. And since the average works as well as the median but is also easier to use for inferential purposes, the average is sometimes blindly used.

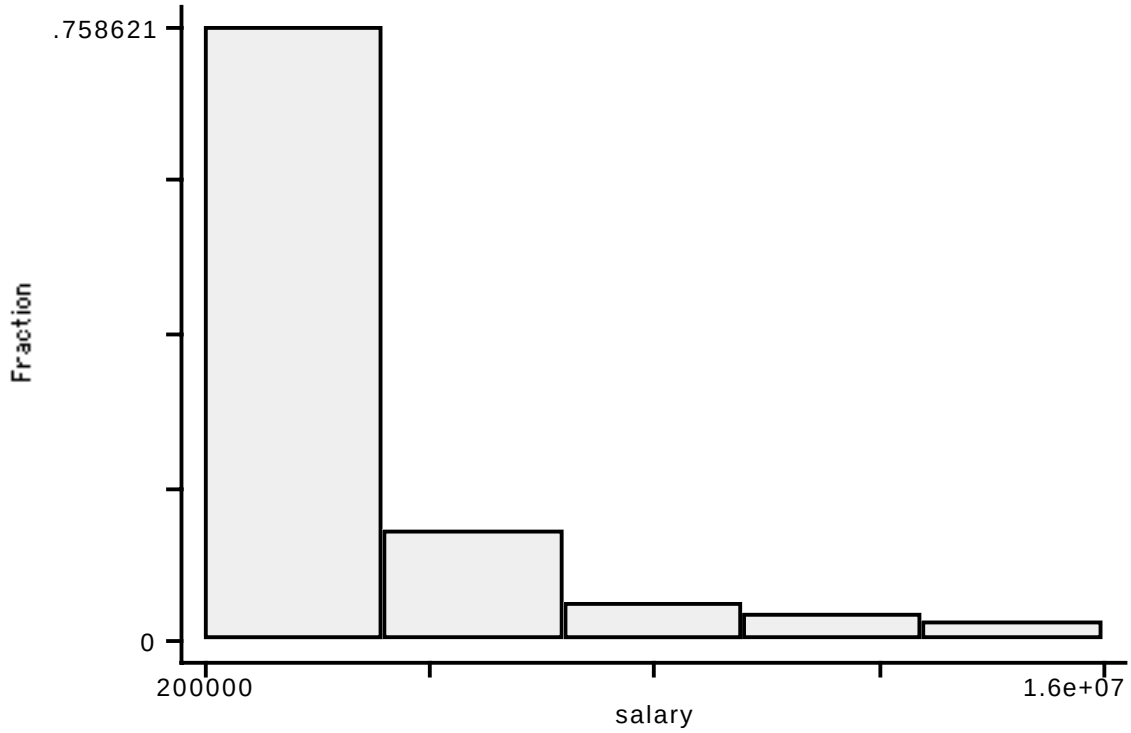
Warning: just because it is almost always used, does not mean the average should be used. Frequently, for example, you will see in labor disputes that the management side will cite the average income of its workers. Why? Because they know that the distribution is right-skewed and therefore the average income will be bigger than the median income. And this means that in truth a majority (more than 50%) of the workers make **less** than the average income.

Below are salaries of players on 3 southern California baseball teams. (Dodgers, Angels, Padres.)

dl-Kevin Brown	$15,714,286	D
Shawn Green	$12,166,667	D
Gary Sheffield	$9,916,667	D
Chan Ho Park	$9,900,000	D
Darren Dreifort	$9,400,000	D
dl-Carlos Perez	$7,833,333	D
Eric Karros	$7,375,000	D
Jeff Shaw	$6,383,333	D
Andy Ashby	$6,000,000	D
Marquis Grissom	$5,000,000	D
Mark Grudzielanek	$4,000,000	D
Tom Goodwin	$3,416,667	D
Terry Adams	$2,600,000	D
Gregg Olson	$1,750,000	D
Mike Fetters	$1,725,000	D
dl-Adrian Beltre	$1,250,000	D
Chad Kreuter	$900,000	D
dl-Dave Hansen	$625,000	D
Tim Bogar	$525,000	D
Jeff Reboulet	$450,000	D
Chris Donnels	$300,000	D
Matt Herges	$250,000	D
Alex Cora	$240,000	D
Paul Loduca	$230,000	D
Eric Gagne	$220,000	D
Angel Pena	$210,000	D
Hiram Bocachica	$200,000	D
Jose Nunez	$200,000	D
Ken Prokopec	$200,000	D
dl-Mo Vaughn	$13,166,667	A
Tim Salmon	$5,683,013	A
Garret Anderson	$4,500,000	A
Darin Erstad	$3,450,000	A
Troy Percival	$3,400,000	A
Ismael Valdes	$2,500,000	A
Pat Rapp	$2,000,000	A
Glenallen Hill	$1,500,000	A
Troy Glaus	$1,250,000	A
Shigetoshi Hasegawa	$1,150,000	A
Scott Spiezio	$1,125,000	A
Orlando Palmeiro	$900,000	A

Alan Levine	$715,000	A
Mike Holtz	$705,000	A
Jorge Fabregas	$500,000	A
Benji Gil	$350,000	A
Ben Molina	$350,000	A
dl-Gary DiSarcina	$320,000	A
dl-Adam Kennedy	$280,000	A
Scott Schoeneweis	$275,000	A
dl-Kimera Bartee	$270,000	A
dl-Jarrod Washburn	$270,000	A
Ramon Ortiz	$250,000	A
Jose Nieves	$232,000	A
Lou Pote	$215,000	A
Matthew Wise	$207,500	A
Ben Weber	$203,500	A
Shawn Wooten	$200,500	A
David Eckstein	$200,000	A
dl-Rendy Espina	$200,000	A
Wally Joyner	$200,000	A
Trevor Hoffman	$6,600,000	P
dl-Sterling Hitchcock	$6,000,000	P
Ryan Klesko	$5,750,000	P
Woody Williams	$5,083,333	P
Chris Gomez	$3,000,000	P
Mark Kotsay	$2,125,000	P
Phil Nevin	$1,625,000	P
Tony Gwynn	$1,600,284	P
Jay Witasick	$800,000	P
Bobby J. Jones	$625,000	P
Dave Magadan	$575,000	P
Alex Arias	$550,000	P
Kevin Jarvis	$550,000	P
Damian Jackson	$350,000	P
Bubba Trammell	$335,000	P
Rodney Myers	$290,000	P
Kevin Walker	$255,000	P
Wiki Gonzalez	$240,000	P
Ben Davis	$235,000	P
Adam Eaton	$235,000	P
Tom Davey	$230,000	P
Brian Tollberg	$230,000	P
dl-Carlton Loewer	$225,000	P
Mike Darr	$215,000	P
Santiago Perez	$207,000	P
David Maurer	$202,500	P
Donaldo Mendez	$200,000	P

A casual examination shows that the data are rather skewed. The graph bears this out:



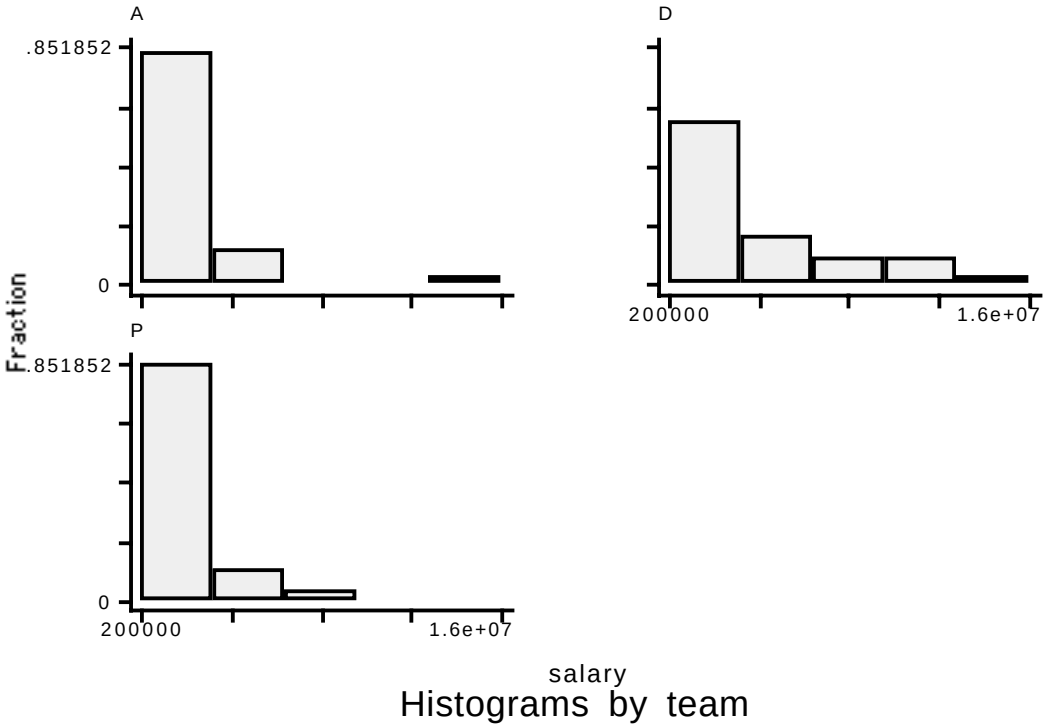
All three teams.

The data are very skewed, as you can see. This means that we expect the average to be bigger than the median, as in fact it is:

average: $2,228, 532
median: $550,000

We might use the average or the medians to compare the teams. Which team pays the most?

Team	**Average**	**Median**
Angels	$1,502,199	$350,000
Dodgers	$3,757,964	$1,725,000
Padres	$1,419,745	$350,000



II. Measure of Spread

The Dodgers clearly have a "typical" salary (no matter what you use as typical) that's higher than the other teams. But which team has the highest paid player? The lowest paid? Suppose we were to pick a player at random, how far away are they likely to be from the typical salary?

A slightly different problem. You want to choose which company to apply to jobs at. Both pay average/median salaries of $40,000 to starting employees. But one pays all new employees (with your expertise) $40K, the other varies between $30K and $50K. Obviously, even though both have the same "typical" salaries, the distribution of starting salaries is very different.

The point is that there is more to a distribution than its center. There is also (among other things) the spread. As it turns out, a measure of center together with a measure of spread tells us an awful lot about the distribution.

1. Range

The most obvious, and probably simplest, measure of spread is the range. The range is simply the biggest value minus the smallest. The range in heights for our class data was 14 inches. This means that the difference between the tallest and shortest person in the class was 14 inches.

The greatest strength of the range is it is extremely easy to compute, and therefore provides a quick understanding of how different distributions compare. The drawbacks, however, are many. First, much more so than the average, it is very sensitive to changes in the extremes. Second, a wide variety of very different distributions can have the same range. So the range (coupled with a measure of center) doesn't give us very precise information about the shape of the distribution.

Two alternatives exist:

2. The Inter-quartile Range

The IQR depends on two statistics: the first quartile and the third quartile. The first quartile cuts off the lower 1/4 of the data. The third cuts off the upper 1/4. (You've already met the 2^{nd} quartile; this is just another name for the median.) In effect, we are cutting the data into fourths.

Here's how to find the quartiles:

1) Find the median. You now have two lists: a left list (below the median) and a right list (above).
2) Find the median of the left list. This is Q1, the first quartile.

3) Find the median of the right list. This is Q3, the third quartile.
4) The IQR is Q3 -Q1

Note that 50% of the observations, in fact the middle 50%, lie between Q1 and Q3. So the IQR tells us how far about the middle part of the data are.

Ex: 2,3,3,4,5,6,6,7,8,9
The median is 5.5 (average of 5 and 6). This cuts the list into:
2,3,3,4,5 and 6,6,7,8,9
Q1 = 3, Q3 = 7, IQR = 7-3=4

Ex: For the height data:
Q1 = 64, Q3 = 69, IQR = 5"

50% of the class fall between 64 and 69 inches.

For the baseball data, the IQR is
Q1 = $235,000
Q3 = $3,000,000
IQR = $2,765,000

The IQR, like the median, is resistant to spread. It focuses solely on what the middle of the distribution is doing and ignores the tails.

3. The Standard Deviation

Get out your calculators. The Standard Deviation (SD), also called the sample standard deviation to emphasize that it comes from a sample of numbers, measures how spread out the numbers are from the average.

If a distribution has all of its numbers clustered tightly around the average value, this is a small number. If the numbers are very spread out, it is big.

We'll discuss
I. How to calculate
II. How to interpret

How to Calculate:
1) A "deviation" is the distance that a number is from the average: x - xbar. So the first step is to calculate the deviations for all of the numbers.

2) If we simply added up the deviations, we'd run into a problem: the numbers bigger than xbar have positive deviations, the number below have negative deviations. So if the distribution were perfectly symmetric, the deviations would sum to 0, even if the observations were very spread out. To fix this, we square the deviations. (An alternative

would be to take the absolute value, but this proves to be more difficult.) So for each observation, we calculate: (x - xbar)^2

3) Add up the squared deviations and divide by n-1 (n is the sample size -- the number of numbers in the list).

4) Take the square root. This is the standard deviation, s.

This is a complicated formula to type into a calculator. The book has an alternative formula that is somewhat easier for calculator use. Of course, on a computer this is even easier, since the computer does it for you.

Let's go through a simple example:
The list of numbers is 3,4,4,5,6. First we need to know xbar: 22/5 = 4.4

x	**deviation**	**squared deviation**
3	-1.4	1.96
4	-.4	.16
4	-.4	.16
5	.6	.36
6	1.6	2.56
sum		**5.2**

5.2/(n-1) = 5.2/4 = 1.3
sqrt(1.3) = 1.14018 = s

The standard deviation for the height data is best calculated with a computer. It is: 3.77.

The SD for the baseball data is: $3,303,380. Note that the SD is larger than the average. This is typical in skewed distributions.

But what does this mean?

<u>Interpretation.</u>

First, the SD is useful, as are all the other measures, as a means for comparing different distributions. If one list has a smaller SD than the other, then its distribution is more compact. The numbers are less spread out.

The SD is not resistant to spread, like the IQR. This means that if your data set has outliers it might be too big, but the blessing is that it gives a more sensitive understanding of the shape. For example:

x x
x x
x x
x x

Distribution A

x x
x x x
x x x

Disttribution B

Both distributions have the same range and IQR, but are the SDs the same?

In fact, B has a smaller SD because more of the points are closer to the average.

Empirical Rule

The true benefit of the SD and the average has to do with something called the Empirical Rule (also called the 65-99-99.7 Rule). This rule applies specifically to symmetric distributions, and to long lists of numbers, but as it turns out is a good rule-of-thumb for almost all occasions.

The rule has three parts:
1. About 65% of the observations in a list will lie within 1 SD of the average (xbar +- s)
2. About 95% will lie with 2 SDs of the average (xbar +- 2s)
3. About 99.7 (almost all) within 3 SDs of the average (xbar +- 3s)

Think back on the IQR. It is a range for which we know that 65% of the observations will be within it. This rule tells us something similar (although more approximate), about the SD. About 65% of the observations are within 1 SD of the average. And you can be pretty confident that you'll only rarely encounter observations more than 3 SDs from average. In fact, the great majority of the time, values of the variable will be within 2 SDs of average.

Let's try it out on the height data. Recall
xbar = 66.63 inches, s = 3.77 inches

	xbar +- 1s	**xbar+- 2s**	**xbar +- 3s**
limits	62.8, 70.4	59.1, 74.2	55.3, 77.9
# of obs	27	33	35
% of obs	77.1	94.3	100

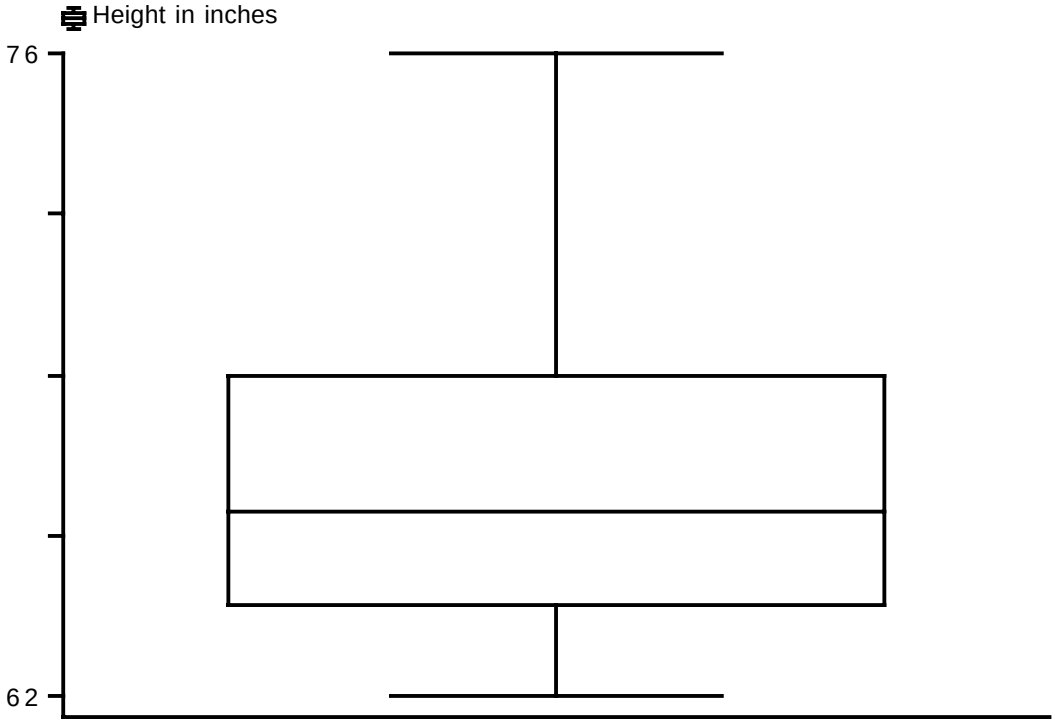
So it works, approximately, even though this is not a very symmetric histogram.

Boxplots

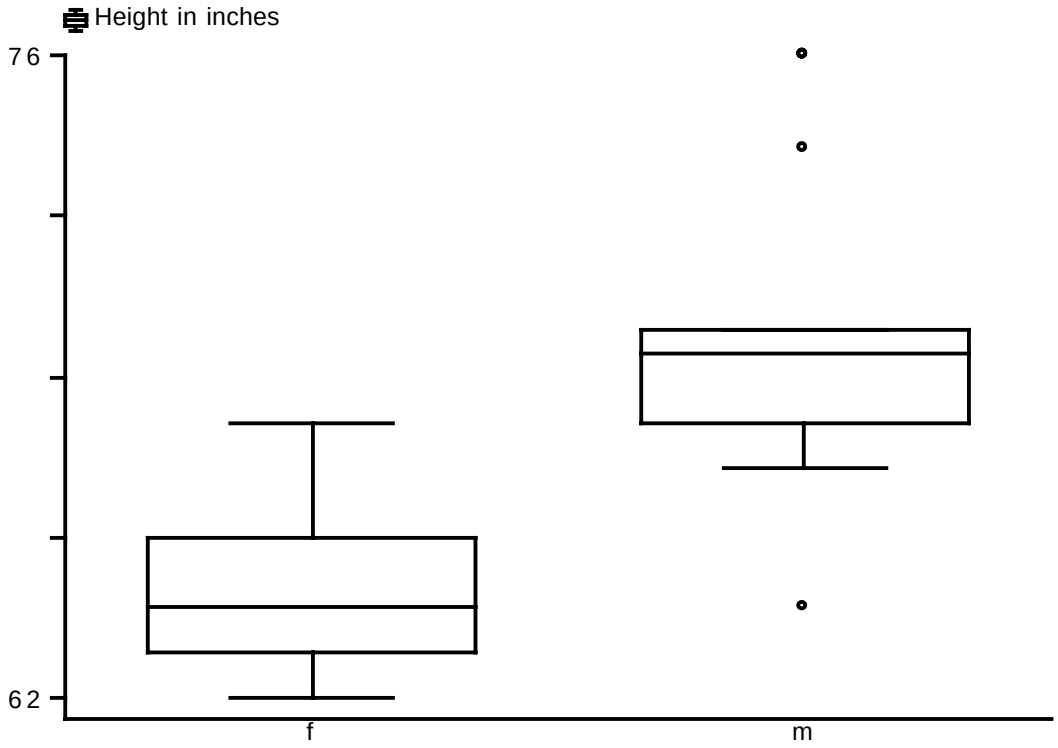
Boxplots are actually a graphical summary technique, but they require understanding of some of the numerical summaries. A boxplot draws, as the name suggests, a box. The box is drawn against a vertical axis so that:
a) the top of the box is at Q3
b) the bottom is at Q1
c) a horizontal line is drawn inside the box at Q2 (the median).

In addition, a box plot has "whiskers". Rules for exactly how to draw the whiskers varies, but in this class we'll use:
d) the upper whisker is put at either
the maximum value or
1.5 * IQR + Q3
whichever is shorter.
e) the lower whisker is put at either
the minimum value or
Q1 - 1.5*IQR
whichever is shorter.
f) If there are observations beyond the whiskers, they are marked with stars or points.

Boxplots are very useful for comparing distributions.

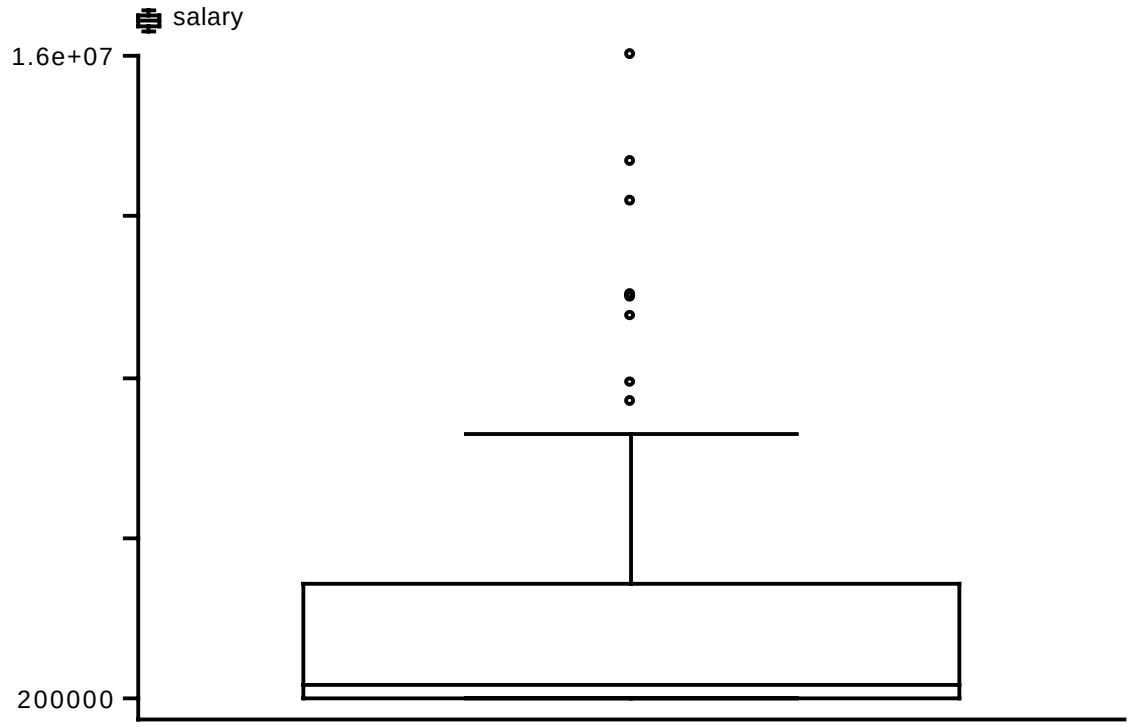


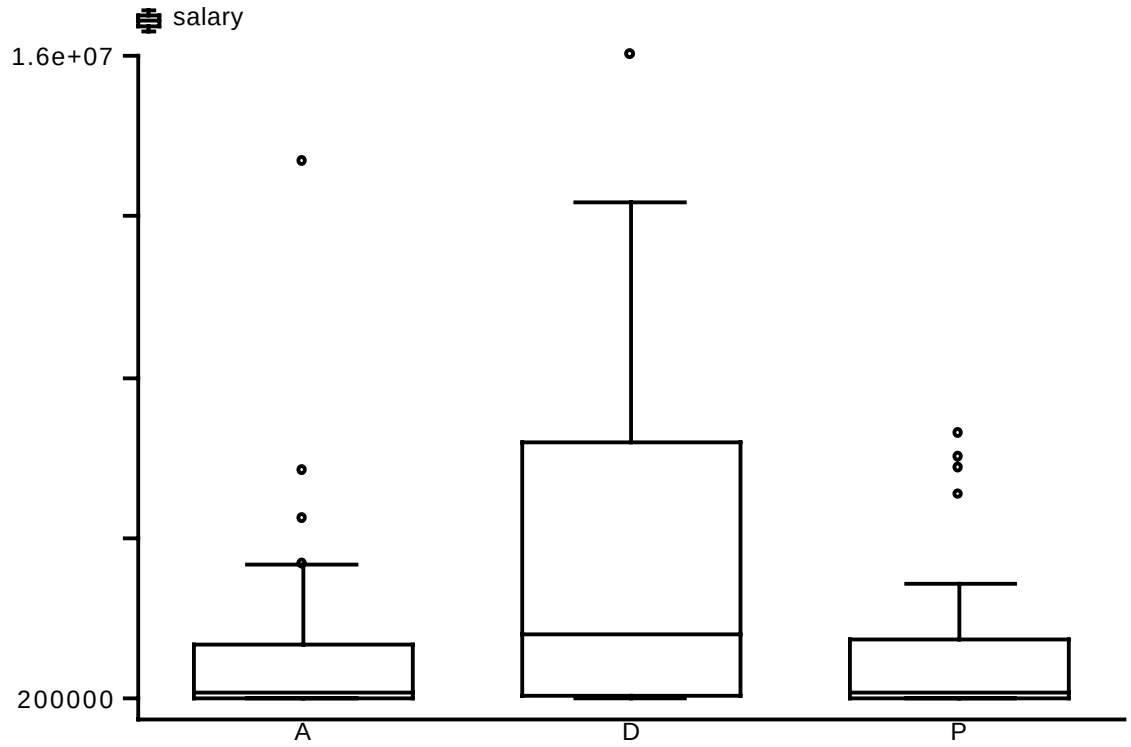
All heights in our class.



Heights by gender.

Baseball salaries, 3 teams





The Effect of Certain Transformation of the Data on the average and SD

Suppose we decide that baseball players make too much, and we pay every player 50K less.
1) What effect does this have on the average salary?
2) What effect does this have on the SD of the salary?

To answer these questions, think of what the histogram of before and after salaries would look like. The shape of the histogram would be the same, except that everything would be shifted to the left by 50,000. So the balancing point, instead of being at $2,228, 532 would now be at $2,228, 532 - $50,000 = $2,178,532.

The SD has not changed because the values are just as clustered about the average as they were before. This illustrates a general rule:

1) If you add/subtract a constant, K, to every item in the list, the average increases/decreases by K.
2) The SD stays the same.

Now suppose that this country passes a flat tax for income tax, and the tax is 35%. What will be the average tax paid by these players? What will be the SD of the amount of taxes paid?

To answer this, you need to know how to compute the tax. If your income is X, then your tax is 0.35 * x. So to calculate the tax for the ball players, we multiply every item on the list by 0.35.

What happens to the average and SD of a variable if every value is multiplied by a constant? This is harder to visualize, but the result is straight-forward:
• the new average is K * old average. So the average tax is 0.35 * $2,228, 532 = $779,986.
• the new SD is K*old SD. So the SD of taxes paid is 0.35*3,303,380 = 1,156,180

What would be the average and SD of the taxes paid on the "reduced" incomes?

Put the rules together. If X is the original income, then the taxes paid on the reduced income is Y = 0.35*(X - 50,000) = 0.35X - 17500. So this can be read as a two-fold operation: multiple each item on the list by .35, then subtract 17500 from each of the "new" items.

The new average is, therefore
new average = 0.35 * (old average) - 17500
For the baseball data: average tax paid (0.35 * 2228532) - 17500 = $762486

The new SD is
new SD = 0.35*(old SD)

(The 17500 doesn't affect the SD because it just shifts the distribution over; it doesn't change the shape of the distribution.)
SD of tax paid: $1,156,180