Introductory Statistics Lectures

Measures of Center

Descriptive Statistics II

Anthony Tanbakuchi
Department of Mathematics
Pima Community College

(Compile date: Tue May 19 14:48:21 2009)

Contents

1 Measures of Center

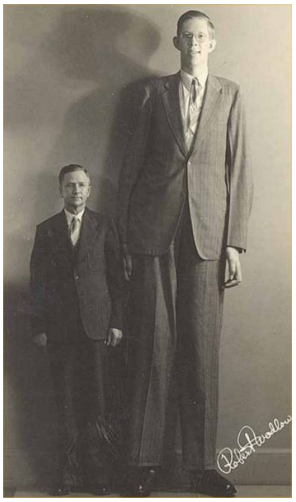
R tip
Tab completion: type the first few letters of a variable or function's name and R will complete it.

1.1 Introduction

Measures of center

Robert Pershing Wadlow (February 22, 1918 - July 15, 1940) is the tallest person in medical history for whom there is irrefutable evidence. He is often known as the "Alton Giant" because of his Alton, Illinois hometown.

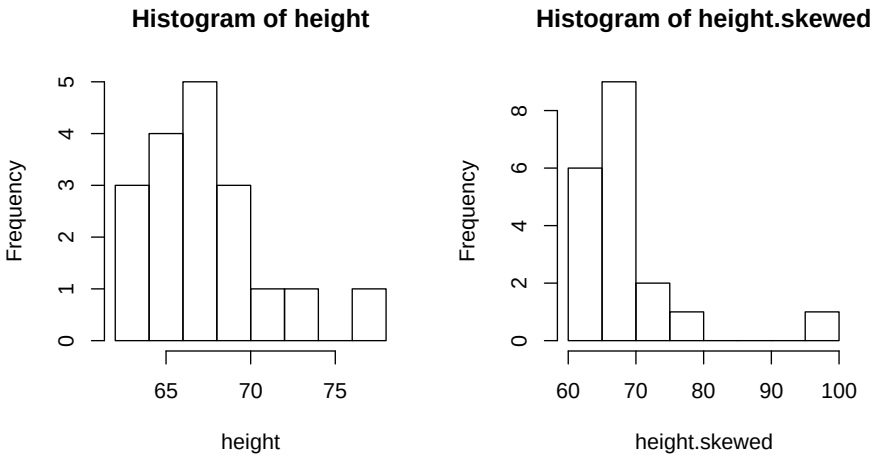
Wadlow reached an unprecedented 8 feet 11.09 inches (2.72 m) in height and weighed 440 pounds (199 kg) at his death. His great size and his continued growth in adulthood was due to hypertrophy of his pituitary gland which results in an abnormally high level of human growth hormone. He showed no indication of an end to his growth even at the time of his death.



Robert Wadlow compared to his father, Harold Franklin Wadlow.

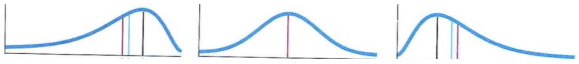
Make a new variable `heights.skewed` where Wadlow (97 inches) is added to our class.

```
R: load("ClassData.RData")
R: height = class.data$height
R: height.skewed = c(height, 97)
R: par(mfrow = c(1, 2))
R: hist(height)
R: hist(height.skewed)
```



SKEWED DISTRIBUTION. DEFINITION 1.1

has left and right tails that are not symmetrical.
positive longer right tail.
negative longer left tail.



negative (left) skewed, symmetrical, positive (right) skewed

NOTATION

Summation operator

$\sum$ operator
A compact notation for summation:

$$\text{total} = \sum_{i=1}^{n} x_i \quad (1)$$

$$= x_1 + x_2 + \cdots + x_n \quad (2)$$

SUMMATION: R COMMAND
`sum(x)`
Where `x` is a vector.

Mathematical Notation

	population	sample
data set	$x = \{x_1, x_2, \ldots, x_N\}$	$x = \{x_1, x_2, \ldots, x_n\}$
size	N	n
sum of data set	$\sum_{i=1}^{N} x_i$	$\sum_{i=1}^{n} x_i$
freq dist (k classes)		$f_i = \{f_1, f_2, \ldots, f_k\}$
prop freq dist		$n = \sum_{i=1}^{k} f_i$

R Notation

	population	sample
data set	`x=c(x1, x2, ...)`	`x=c(x1, x2, ...)`
size	`N=length(x)`	`n=length(x)`
sum of data set	`sum(x)`	`sum(x)`
freq dist (k classes)		`f=c(f1, f2, ...)`
prop freq dist		`n=sum(f)`

Measures of center and effect of outliers

How are each of the measures of center effected by outliers?

1.2 Three key measures of center

MEAN

DEFINITION 1.2

MEAN: μ, $\bar{x}$.

The arithmetic mean is the sum of the data set divided by the number of values.

$$\boxed{\mu = \frac{\sum_{i=1}^{N} x_i}{N}} \quad \text{(parameter: population mean)}$$

$$\boxed{\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}} \quad \text{(statistic: sample mean)}$$

The balance point on a distribution.

How susceptible is the mean to outliers?

Question 1. Given that $x = \{9, 5, 6, 4\}$ find the mean of x.

R COMMAND

LENGTH OF A VECTOR:
`length(x)`

Returns the number of elements in the vector `x` .

Example 1. Find the mean of the student heights in R "manually".

```
R: n = length(height)
R: n
[1] 18
R: sum(height)/n
[1] 67.611
```

We can write this more compactly:

```
R: sum(height)/length(height)
[1] 67.611
```

Making your own mean function in R
Create a new function called `my.mean(x)` that takes one vector as an argument:

```
R: my.mean = function(x) {
+     sum(x)/length(x)
+ }
```

Now use `my.mean(x)` to find the mean of the student heights.

```
R: my.mean(height)
[1] 67.611
```

MEAN:
`mean(x, trim=0)`
Find the mean of the vector `x`. If you set the optional argument `trim=0.1` it will trim the top and bottom 10% of the data points before finding the mean.

R COMMAND

Example 2. Given that $x = \{9, 5, 6, 4\}$ find the mean of x. This is easy in R!

```
R: x = c(9, 5, 6, 4)
R: mean(x)
[1] 6
```

Example 3. Effect of skewed data:

```
R: mean(height)
[1] 67.611
R: mean(height.skewed)
[1] 69.158
```

Example 4. Trimmed mean

```
R: mean(height.skewed, trim = 0.1)
[1] 67.941
```

Mean from a frequency distribution
If don't have the original data but you have a frequency distribution table or histogram:

$$\bar{x} \approx \frac{\sum_{i=1}^{k} f_i \cdot \bar{x_i}}{n} \tag{3}$$

$$= \frac{f_1 \cdot \bar{x_1} + f_2 \cdot \bar{x_2} + \cdots + f_k \cdot \bar{x_k}}{n} \tag{4}$$

f_i class frequency (count), k classes
$\bar{x_i}$ class midpoint

Why is this only an approximation?
Given a frequency distribution table[1] of student height data:

	Class	Midpoints	Frequency
1	[62,64)	63.00	3
2	[64,66)	65.00	3
3	[66,68)	67.00	2
4	[68,70)	69.00	5
5	[70,72)	71.00	3
6	[72,74)	73.00	1
7	[74,76)	75.00	0

Example 5. Approximation of mean from frequency distribution:

```
R: f
[1] 3 3 2 5 3 1 0
R: midpoints
[1] 63 65 67 69 71 73 75
R: n
[1] 17
R: x.bar = sum(f * midpoints)/n
R: x.bar
[1] 67.588
```

Compare our approximate mean above with the true mean of 67.6111111111111.

MEDIAN

DEFINITION 1.3

MEDIAN $\tilde{x}$.

The middle value of a sorted data set:

1. Sort the values.
2. Median value is x_i where $i = \frac{n+1}{2}$. If i is not an integer, average neighboring two values.

Breaks frequency distribution into **two equal areas**.

R COMMAND

MEDIAN:
`median(x)`

Where `x` is a vector.

Example 6. Given that $x = \{9, 5, 6, 4\}$ find the median of x. This is easy in R!

```
R: x = c(9, 5, 6, 4)
R: median(x)
[1] 5.5
```

Question 2. Given that $x = \{2, 5, 6, 10, 11\}$ find the median of x.

[1] Recall $[a, b) = a \leq x < b$.

Question 3. Given that $x = \{2, 5, 6, 10, 11, \mathbf{14}\}$ find the median of x.

Example 7. Effect of skewed data on median:

```
R: median(height)
[1] 68
R: median(height.skewed)
[1] 68
```

Question 4. Is the median more or less resistant to outliers than the mean?

MODE

MODE: M. DEFINITION 1.4

The value that occurs most frequently. (Primarily for categorical data and discrete data.)

- Easily found with a frequency table.
- If two or more categories occur with the greatest frequency: **bimodal, multimodal**.
- If no class repeats, no mode.

Mode is the peak of a distribution.

Mode would not make sense for continuous data (ex. height, weight, ...).

Given the class data on type of transportation used:

```
R: table(class.data$transportation)
   CAR PUBLIC  TRUCK
    13      2      3
```

Question 5. What is the mode for the transportation data?

Given the following frequency table of randomly selected marble colors:

Red	Green	Blue	Black
4	8	3	8

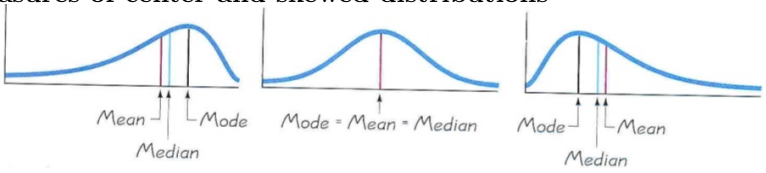
Question 6. What is the mode for the marble colors?

Given the following frequency table of pet types for 16 randomly selected people:

Dog	Cat	Reptile	None
4	4	4	4

Question 7. What is the mode for the pet type?

Measures of center and skewed distributions



1.3 Other measures of center

DEFINITION 1.5 AVERAGE.

The "average" of a set of data is **not specific.** Often people use average when they are referring to the mean, but not always! Be specific, use the proper mathematical name for the measure of center used.

- When someone gives you the average of a data set, ask what kind of average it was.
- When someone asks you to find the average of a data set, determine which measure of center would be appropriate!

DEFINITION 1.6 MIDRANGE.

value midway between max and min values

$$\text{midrange} = \frac{\max x + \min x}{2} \tag{5}$$

DEFINITION 1.7 WEIGHTED MEAN.

allows weighting data points by importance.

$$\text{weighted mean: } \bar{x} = \frac{\sum w_i \cdot x_i}{\sum w_i}$$

DEFINITION 1.8 HARMONIC MEAN.

used as a measure of central tendency for data consisting of rates of changes (ex. speed).

$$\text{harmonic mean} = \frac{n}{\sum \frac{1}{x_i}}$$

DEFINITION 1.9 GEOMETRIC MEAN.

for average rates of change, growth, or ratios. Popular in business and economics. Given n values (all positive)

$$\text{geometric mean} = \sqrt[n]{\prod x_i}, \quad x_i > 0$$

QUADRATIC MEAN OR ROOT MEAN SQUARE RMS. DEFINITION 1.10

used in physical applications.

$$\text{quadratic mean} = \sqrt{\frac{\sum x_i^2}{n}}$$

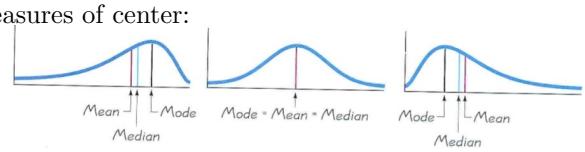
TRIMMED MEAN. DEFINITION 1.11

to reduce sensitivity to extreme values trim k% of max and min values. 10% trimmed mean cuts top and bottom 10% of values: `mean(x, trim=0.10)`

Question 8. Given that $x = \{1, 2, 3, 4, 5\}$ find the midrange of x.

1.4 Summary

Measures of center:



1. Quantitative data: listed in increasing resistance to outliers
   a) Midrange
   b) Mean: $\mu, \bar{x}$ (R: `mean(x)`)
   c) Trimmed mean (R: `mean(x), trim=0`)
   d) Median: $\tilde{x}$ (R: `median(x)`)
2. Qualitative data: mode: M (R: `table(x)`)

For each measure of center:

- Know how to use the equations!
- Know how to compute them in R!
- Know what they represent and which is most/least affected by outliers.

1.5 Additional examples

Example 8. Take a look at the built in data set `ChickWeight` in R. The following questions pertain to the column describing the chick weights in grams.

- Type `ChickWeight` and look at the table.
- Make a histogram of the chicken weights.
- Is the data symmetrically distributed, positively skewed, or negatively skewed?
- Which measure of center would you use to describe the data and why?
- Find the mean.
- Find the median.
- Should the mean equal the median? Why?
- Use the histogram of the data to estimate a mode.

1.6 Defining Functions In R: Optional Knowledge

To define your own function in R:

R Command

DEFINING FUNCTIONS:
`functionName = function(arg1, arg2, ...) { FunctionBody }`

functionName a name you pick for the function
arg1, arg2, ... arguments (inputs) for your function
FunctionBody the code for your function. Can be many lines. Just make sure it goes between { and } and the only available variables are the arguments. To return a value, use `return(value)` . If the end of a function is reached without calling `return` , the value of the last evaluated expression is returned.
To use your function:

`functionName(arg1, arg2, ...)`

where you replace `arg1, arg2, ...` with your variables that have data in them.

Example 9. Define a function that returns the range of values in a vector:

```
R: my.range = function(x) {
+     min.value = min(x)
+     max.value = max(x)
+     rng = c(min.value, max.value)
+     return(rng)
+ }
```

Now use it to find the range of values for the student heights:

```
R: my.range(height)
[1] 62 77
```

Example 10. Define a function that computes the log with optional base (default e). We use R's built in natural log function.

```
R: my.log = function(x, base = 2.7182) {
+     val = log(x)/log(base)
+     return(val)
+ }
```

Now use it to find the log of 256 with base e, 2, and 10:

```
R: my.log(256)
[1] 5.5453
R: my.log(256, base = 2)
[1] 8
R: my.log(256, base = 10)
[1] 2.4082
```