Statistical Methods for Plant Biology

PBIO 3150/5150

Anirudh V. S. Ruhil

January 21, 2016

The Voinovich School of Leadership and Public Affairs

Table of Contents

1. Measuring Central Tendency
2. Median
3. Measuring Variability
4. Proportions
5. Comparing Measures of Location
6. Choice Rules of Thumb
7. Some Useful Plots

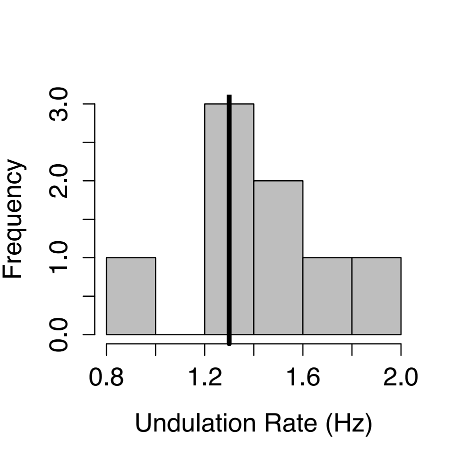
Descriptive Statistics

- We now turn to descriptive statistics that tell us something about what is “typical” of a given distribution and how much observations tend to “differ” from one another
- What is “typical” (i.e., what would you expect to see, on average) is measured via
    1. Mean
    2. Median
    3. Mode
- How observations “differ” is measured via
    1. Range
    2. Interquartile Range and the Semi-Interquartile Range
    3. Variance and the Standard Deviation

Measuring Central Tendency

Gliding Snakes

- Paradise tree snakes glide in the air as they travel
- Socha (2002) measured undulation rates of 8 snakes
- One might then ask: *What is the typical undulation rate of these snakes?*
- What you are really asking is: *If you observed, at random, ONE paradise tree snake launching from a height of 10-m, what undulation rate would you expect to see?*



Calculating the Arithmetic Mean

The Population Mean

$$\mu = \frac{\sum_{i=1}^{N} Y_i}{N}$$

where Y_i is the value of the variable Y for the i^{th} observation, $N =$ population size; $i = 1, 2, 3, \ldots, N$ are the observations making up the population, and $\sum_{i=1}^{N} Y_i$ essentially says add up every observation in the population

The Sample Mean

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

where Y_i is the value of the variable Y for the i^{th} observation, $n =$ sample size; $i = 1, 2, 3, \ldots, n$ are the observations making up the sample, and $\sum_{i=1}^{n} Y_i$ essentially says add up every observation in the sample

Mean Undulation Rate

$$\begin{aligned}
\bar{Y} &= \frac{\sum\limits_{i=1}^{n} Y_i}{n} \\
Y_i &= 0.9, 1.4, 1.2, 1.2, 1.3, 2.0, 1.4, 1.6 \\
\sum_{i=1}^{n} Y_i &= 0.9 + 1.4 + \ldots + 1.6 = 11 \\
n &= 8 \\
\therefore \bar{Y} &= \frac{11}{8} = 1.375
\end{aligned}$$

Average undulation rate (in Hertz) is 1.375 $approx = 1.37$

Note: For non-technical audiences you should round or truncate estimates to the nearest two decimal places but for technical audiences you should stay with three/four decimal places. Emulate the practice your field/sub-field tends to follow.

Another Example ...

Example

ID	Salary ($)	ID	Salary ($)
1	2,850	7	2,890
2	2,950	8	3,130
3	3,050	9	2,940
4	2,880	10	3,325
5	2,755	11	2,920
6	2,710	12	2,880

$$\bar{Y} = \frac{\Sigma Y_i}{n}$$

$$= \frac{Y_1 + Y_2 + \cdots + Y_{12}}{n}$$

$$= \frac{2{,}850 + 2{,}950 + \cdots + 2{,}880}{12}$$

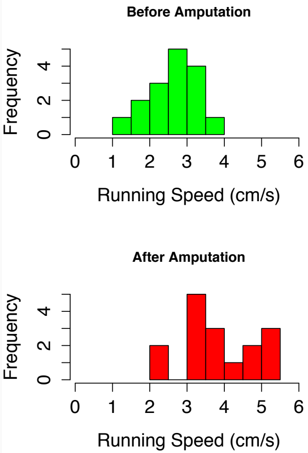
$$= \frac{35{,}280}{12}$$

$$= \$2,940$$

Example Using the Spider data

Male red Tidarren spiders amputate one of 2 external sex organs to move fast, win a mate.

#	Speed Before	Speed After
1	1.25	2.40
2	2.94	3.50
3	2.38	4.49
4	3.09	3.17
5	3.41	5.26
6	3.00	3.22
7	2.31	2.32
8	2.93	3.31
9	2.98	3.70
10	3.55	4.70
11	2.84	4.94
12	1.64	5.06
13	3.22	3.22
14	2.87	3.52
15	2.37	5.45
16	1.91	3.40



Mean speed before $= 2.66$

Mean speed after = 3.85

Properties of the Mean

1. Changing the value of any observation changes the mean
2. Adding or subtracting a constant k from all observations is equivalent to adding or subtracting the constant k from the original mean
3. Multiplying or dividing a constant k from all observations is equivalent to multiplying or dividing the original mean by the constant k

Example

ID	Y	$(Y-2)$	$(Y \times 2)$	$\left(\frac{Y}{2}\right)$
1	6	4	12	3
2	3	1	6	1.5
3	5	3	10	2.5
4	3	1	6	1.5
5	4	2	8	2
6	5	3	10	2.5
Total	26	14	52	13

Median

The Median

The median halves the distribution ...

1. Sort the data (ascending or descending order)
2. If n is odd, median is the observation in the $\frac{n+1}{2}$ position

   Say we had n=7: $\left(0.9, 1.2, 1.2, \enclose{circle}{1.3}, 1.4, 1.4, 1.6\right)$

   Then middle observation is $\frac{n+1}{2} = 4^{th} observation =$ the Median value.
3. If n is even, median is the average of middle two obs $\frac{Y_{\frac{n}{2}} + Y_{\frac{n+1}{2}}}{2}$

   If we had n=8: $\left(0.9, 1.2, 1.2, \enclose{circle}{1.3}, \enclose{circle}{1.4}, 1.4, 1.6, 2.0\right)$

   then median $=$ Average of Middle 2 observations $= \left(\frac{1.3+1.4}{2}\right) = 1.35$

   i.e., $\left(0.9, 1.2, 1.2, 1.3\enclose{circle}{1.35}1.4, 1.4, 1.6, 2.0\right)$

Another Median Example (*n* is even)

Example

ID	Salary ($)	ID	Salary ($)
1	2,710	7	2,920
2	2,755	8	2,940
3	2,850	9	2,950
4	2,880	10	3,050
5	2,880	11	3,130
6	2,890	12	3,325

$$Md = \frac{2{,}890+2{,}920}{2}$$

$$Md = \frac{5{,}810}{2} = \$2,905$$

Another Median Example (*n* is odd)

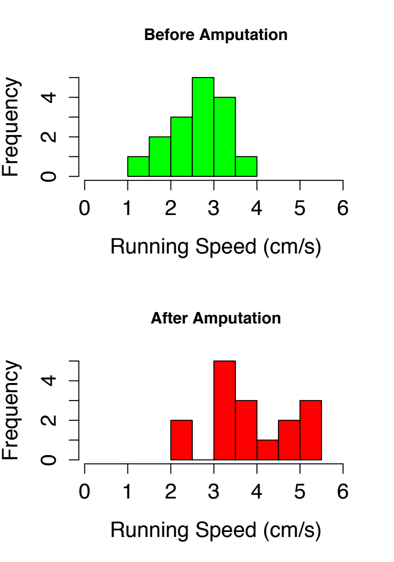
Example

ID	Salary ($)	ID	Salary ($)
1	2,710	7	2,920
2	2,755	8	2,940
3	2,850	9	2,950
4	2,880	10	3,050
5	2,880	11	3,130
6	2,890		

$$Md = \frac{n+1}{2} = 6^{th}$$

$$Md = \$2,890$$

Median with the Spider data



Md speed before = 2.90

Md speed after = 3.51

Quartiles

Definition

Quartiles divide the data into four parts and are denoted as Q_1, Q_2, Q_3
Q_1 is the first quartile or the $25^{th} percentile$
Q_2 is the second quartile or the $50^{th} percentile = Md$
Q_3 is the third quartile or the $75^{th} percentile$

- Q_1 and Q_3 of undulation rates are 1.200 and 1.450, respectively
- Q_1 and Q_3 of speed before are 2.355 and 3.022, respectively
- Q_1 and Q_3 of speed after are 3.510 and 4.760, respectively

Mode

Definition

The Mode is the value with the greatest frequency in the data set

Example

Drink	Freq.
Coke Classic	19
Diet Coke	8
Dr. Pepper	5
Pepsi-Cola	13
Sprite	5
Total	50

Mode = Coke Classic

Measuring Variability

Range, IQR, and S-IQR[1]

- Range is a crude measures of variability: $Y_{max} - Y_{min}$
- Median halves distribution (i.e., 50% below, 50% above)
- Quartiles quarter the distribution (i.e., 25%, 25%, 25%, 25%)
    1. Data (n forced to be odd): $\left(0.9, \boxed{1.2}, 1.2, \boxed{1.3}, 1.4, \boxed{1.4}, 1.6\right)$
    2. $Q_1 = 1.2$; $Q_2 = 1.3$ (the median); $Q_3 = 1.4$
- Interquartile Range (IQR) is the middle 50% of the distribution
    1. $IQR = Q_3 - Q_1 = 1.4 - 1.2 = 0.2$
- Semi-Interquartile Range (S-IQR) is the middle 25% of the distribution
    1. $S - IQR = \left(\frac{Q_3 - Q_1}{2}\right) = \left(\frac{1.4 - 1.2}{2}\right) = \frac{0.2}{2} = 0.1$

Using R ...

1. Snakes: $Range = 2.000 - 0.900 = 1.100; IQR = 1.450 - 1.200 = 0.250$
2. Spiders (before): $Range = 3.550 - 1.250 = 2.300; IQR = 3.022 - 2.355 = 0.6675$
3. Spiders (after): $Range = 5.450 - 2.320 = 3.130; IQR = 4.760 - 3.510 = 1.540$

[1] Software defaults to one of 9 methods for calculating IQR; don't be alarmed

Variance & Standard Deviation

Population Variance

$$\sigma^2 = \frac{\sum (Y_i - \mu)^2}{N}$$

Population Standard Deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (Y_i - \mu)^2}{N}}$$

Sample Variance

$$s^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$$

Sample Standard Deviation

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n-1}}$$

Note: Sum of Squares $= \sum (Y_i - \bar{Y})^2$

Note also: For samples we divide by $n-1$; we'll try to understand why we do this in a few slides

The Calculations ...

i (Snake ID)	Y	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$
1	0.900000	-0.475000	0.225625
2	1.400000	0.025000	0.000625
3	1.200000	-0.175000	0.030625
4	1.200000	-0.175000	0.030625
5	1.300000	-0.075000	0.005625
6	2.000000	0.625000	0.390625
7	1.400000	0.025000	0.000625
8	1.600000	0.225000	0.050625
$n = 8$	$\sum Y_i = 11$	0.000000	0.735000

What would $\frac{\sum(Y_i - \bar{Y})}{n}$ equal??

Another Example ...

Graduate	Y	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
1	2850	-90	8100
2	2950	10	100
3	3050	110	12100
4	2880	-60	3600
5	2755	-185	34225
6	2710	-230	52900
7	2890	-50	2500
8	3130	190	36100
9	2940	0	0
10	3325	385	148225
11	2920	-20	400
12	2880	-60	3600

$$\bar{Y} = 2940$$

$$\Sigma(Y_i - \bar{Y}) = 0$$

$$\Sigma(Y_i - \bar{Y})^2 = 301850$$

$$s^2 = \frac{301850}{(12-1)} = \$27440.91$$

$$s = \sqrt{27440.91} = \$165.63$$

Why $n-1$?

Assume population is: 0, 2, and 4 and $\mu = 2$ while $\sigma^2 = \frac{8}{3} = 2.6667$

In the sample we would want an estimate of $s^2 = \sigma^2$

What happens if we draw all possible random samples (say with $n = 2$) from this population and calculate s^2 ... (a) without using $(n-1)$ or (b) using $(n-1)$?

Table 1: Without $(n-1)$

Sample	$\bar{Y}$	s^2
(0, 0)	0	0
(0, 2)	1	1
(0, 4)	2	4
(2, 0)	1	1
(2, 2)	2	0
(2, 4)	3	1
(4, 0)	2	4
(4, 2)	3	1
(4, 4)	4	0

Table 2: With $(n-1)$

Sample	$\bar{Y}$	s^2
(0, 0)	0	0
(0, 2)	1	2
(0, 4)	2	8
(2, 0)	1	2
(2, 2)	2	0
(2, 4)	3	2
(4, 0)	2	8
(4, 2)	3	2
(4, 4)	4	0

Which method yields average sample variance $= \sigma^2$?

Intuitively: Drift between samples and populations; degrees of freedom

Arithmetic Mean with Grouped Data

No.Convictions	No.Boys	Convictions*Boys	For $\sum(Y_i - \bar{Y})^2$
Y_i	Freq (f_i)	$Y_i \times f_i$	$f_i \times (Y_i - \bar{Y})$
0	265	0	265$\times(0-\bar{Y})$
1	49	49	49$\times(1-\bar{Y})$
2	21	42	21$\times(2-\bar{Y})$
3	19	57	19$\times(3-\bar{Y})$
4	10	40	10$\times(4-\bar{Y})$
5	10	50	10$\times(5-\bar{Y})$
6	2	12	2$\times(6-\bar{Y})$
7	2	14	2$\times(7-\bar{Y})$
8	4	32	4$\times(8-\bar{Y})$
9	2	18	2$\times(9-\bar{Y})$
10	1	10	1$\times(10-\bar{Y})$
11	4	44	4$\times(11-\bar{Y})$
12	3	36	3$\times(12-\bar{Y})$
13	1	13	1$\times(13-\bar{Y})$
14	2	28	2$\times(14-\bar{Y})$

Note that $n = 395$

Calculate $Y_i \times f_i$ (i.e., Convictions $\times$ Boys)

$\bar{Y} = 1.126582 \approx 1.12$

$s^2 = 2377.671; s = 2.4566 \approx 2.45$

Coefficient of Variation

Definition

The Coefficient of Variation is the standard deviation expressed as a percentage of the mean

Useful when comparing dimensions, attributes, etc. that are not on the same scale (for example, elephants' weights versus elephants' life spans) $CV = 100\% \left(\frac{s}{\bar{Y}}\right) \cdots$ standard deviation divided by the mean
For the gliding snakes data

$$\begin{aligned} CV &= 100\% \left(\frac{0.324037}{1.375}\right) \\ &= 23.566\% \approx 23.56\% \end{aligned}$$

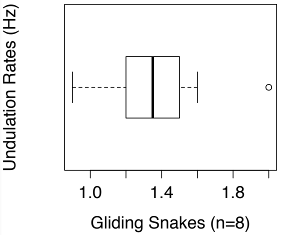
For the spider data we have ...

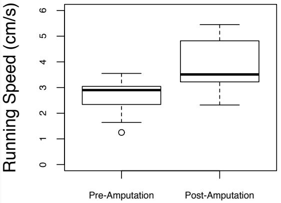
- Speed Before: CV = 24.04507 ≈ 24.04
- Speed After: CV = 25.75756 ≈ 25.75

... The higher the CV the more variability there is ...

Box Plots

- Very powerful for showing how distributions are shaped
- Utilize five numbers:
  Min; Q_1; Q_2; Q_3; and Max
- See two examples on the right
  - Gliding snakes data
  - Spider amputation data
  - Outliers
    1. Values $< Q_1 - 1.5 \times IQR$, or
    2. Values $> Q_3 + 1.5 \times IQR$





Proportions

Proportions

- For categorical variables proportions come in handy. These are nothing but the relative frequencies we saw in Chapter 2
- $\hat{p} = \frac{f_{Category}}{n}$
- Calculating $\hat{p}$ for MM, Mm, and mm yields ...

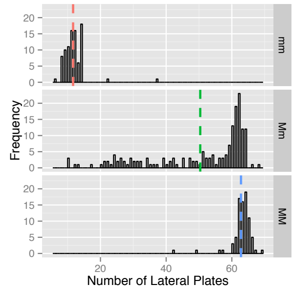
$$\hat{p}_{MM} = \frac{82}{344} = 0.23837$$
$$\hat{p}_{Mm} = \frac{174}{344} = 0.50581$$
$$\hat{p}_{mm} = \frac{88}{344} = 0.25581$$

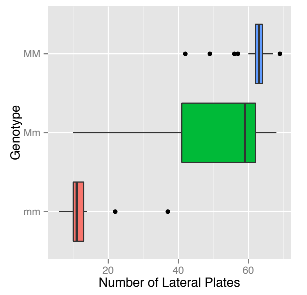
Comparing Measures of Location

Comparing Measures of Location

- Colosimo et al.'s (2004) study of crossed sticklebacks
- Threespine stickleback
  1. MM = 2 copies of gene from marine grandparent
  2. Mm = 1 copy of gene from marine + freshwater grandparent
  3. mm = 2 copies of gene from freshwater grandparent
- What can you discern?

Type	n	$\bar{Y}$	Md	s	IQR
MM	82	62.8	63	3.4	2
Mm	174	50.4	59	15.1	21
mm	88	11.7	11	3.6	3





Choice Rules of Thumb

Choosing A Measure of Location

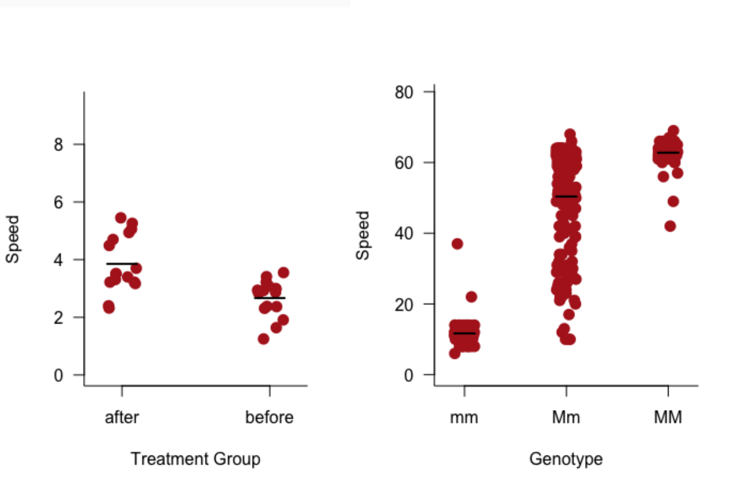
1. Mean usually preferred over Median and Mode because
    - uses all observations in the data
    - used in most statistical calculations
    - intuitive
    - However, asymmetric distributions skew the Mean
2. Mode is easy to calculate, and can be used with both qualitative and quantitative data so analysts often gravitate towards it
3. Median is usually preferred when data
    - have extreme scores
    - are open-ended (e.g., income with categories of $\leq 25,000$ and/or $\geq 200,000$)
    - have some undetermined values (e.g., time on task with some not completing task)

Rules of Thumb ...

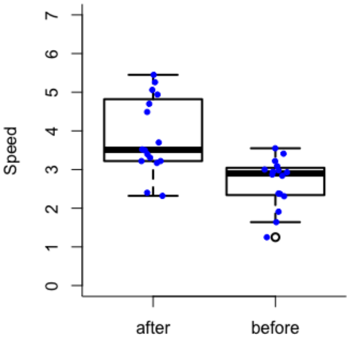
- **Numerical variables:** Mean and Standard Deviation usually preferred, unless
  1. Distribution is skewed ... use Median and IQR
  2. Distribution is open-ended ... use Median and IQR
  3. Distribution includes indeterminate values (for e.g., time on task studies) ... use Median and IQR
- **Categorical variables:** Mode is usually preferred
- With perfectly symmetric distributions: Mean = Median = Mode
- Note: With numerical variables we try to stick with the Mean as long as we can so even with some skewed distributions we will see ways to transform the data and make the distributions more symmetric. If these don’t work then the Median is the only option

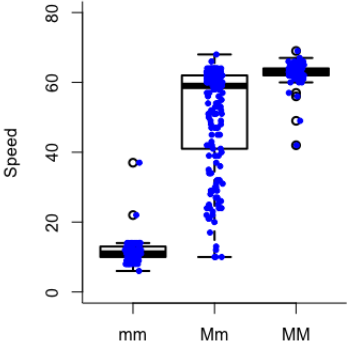
Some Useful Plots

Strip Charts

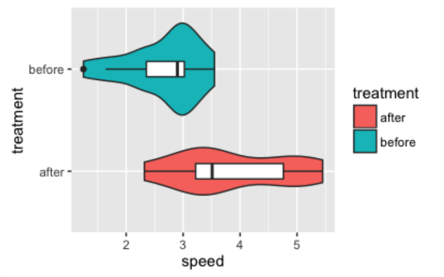


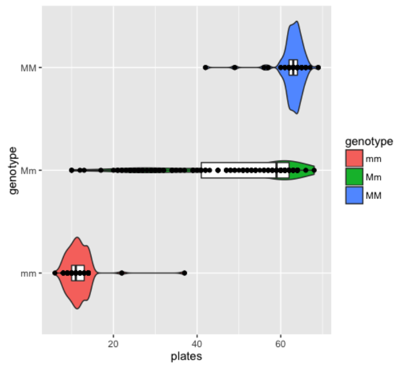
Box-Plots with Jittered Points





Violin Plots


treatment
before
after
speed
treatment
after
before


genotype
MM
Mm
mm
plates
genotype
mm
Mm
MM