STAT F300
Statistics

Kaleb Burris

Lecture 1

(§1.2) Descriptive statistics; (§1.3) Measures of location

Common naming conventions:

- Population size: N
- Sample size: n
- Sample from two different populations: n, m, or n_1, n_2
- Data: $x_1, x_2, x_3, \ldots x_n$

Stem-and-leaf displays

```
> x sample(1:50, size=20, replace=TRUE)
> sort(x)
 [1]  2  2  2  3  9 14 18 19 20 21 21 22 22 29 30 32 32
[18] 33 44 47
> stem(x)
```

The sample function generates numbers in the range provided as the first argument, with a size equal to the second argument. sort(x) sorts the values stored in x, and stem(x) does the following:

Each "stem" refers to the highest digits and each "leaf" is the latter digits. This is the stem-and-leaf display for the dataset stored in x:

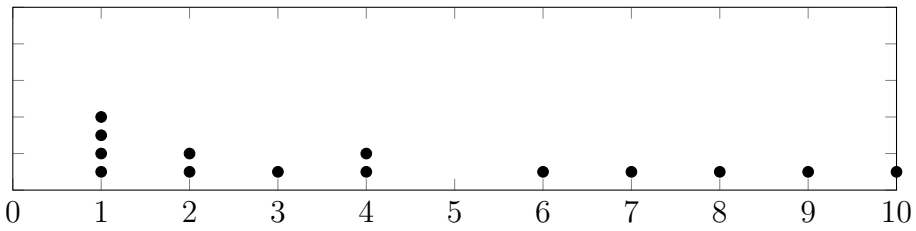
Stem	Leaves
0	2 2 2 3 9
1	4 8 9
2	0 1 1 2 2 9
3	0 2 2 3
4	4 7

Dot plots

```
> x <- sample(1:10, size=15, replace=TRUE)
> x
 [1]  7  8  1  3  4 10  1  2  2  1  1  4  4  9  6
> stripchart(x, method="stack", offset=0.5, at=0.15, pch=20)
```

`stripchart()` constructs the dot plot. The `offset` and `at` values affect the appearance. `pch` means `p` rinting `ch` aracter. Having it set to 20 makes the dot solid.

Dot Plot



Histograms

- For quantitative data.
- May differ depending on whether a variable is discrete or continuous.

Discrete data is either infinite or countably infinite: orderable, typically integers. Includes ration numbers.

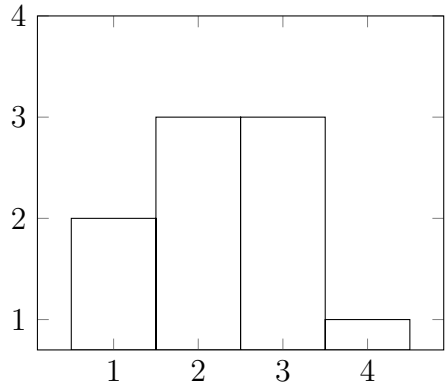
- $x_1 = \#$ of tuberculosis cases in the 50 largest U.S. cities in 2013. ← Might be 0, 1, 2, ... but $\neq \infty$.
- x_i # of "successes" when 100 patients treated. ← Count data in which we know a hard upper limit (in this case 100).
- x_i # of rooms in a rental unit. ← 1, 2, ..., 100. At least 1 guaranteed as a hard minimum limit.

Histograms, when data is discrete, use vertical bars with height proportional to the number of times each value occurs.

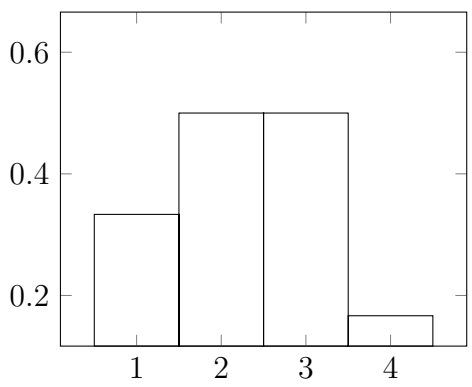
Frequency Histogram vs Relative Frequency Histogram

For the data: $\{1, 1, 2, 2, 2, 3\}$

Frequency Histogram



Relative Frequency Histogram



Relative Frequency Histograms have a total height of 1.

R Code for Histograms

```
> # Use R to simulate 100 values from {1, 2, 3, ..., 10}
> my_data <- sample(1:10, size=100, replace=TRUE)
> # Here are the first 20 values:
> my_data[1:20]
 [1]  4  3  5  6  5  4 10  2  6  2  5  8  7 10  4 10  7  6
[19]  5  7
> # Create a frequency histogram:
> hist(my_data. main="Plot 1 - discrete data")
> # Create a relative frequency histogram:
> hist(my_data, probability=TRUE, main="Plot 2")
```

Note: R uses # for comments. You're a CS major, you know what these functions and their arguments do.

R Code of a Histogram Using Continuous Data

```
> # Generate 100 rational values from 1:10.
> my_data <- runif(100, min=1, max=10)
> # Create a histogram using the same code for discrete data:
> hist(my_data, main="Plot 3 - Continuous Data")
> # And a relative frequency histogram:
> hist(my_data, probability=TRUE, main="Plot 4")
```

Here, `r` refers to random and `unif` refers to uniform in `runif`, it does not refer to `run if`.

(§1.3) Measures of Location

There are several ways to measure where the center of a data set s. Different measures are used in different scenarios because people use what they're accustomed to using. Here are the most frequently used ones:

1. Sample Mean: $\overline{x} = x_1 + x_2 + \cdots + x_n = \frac{1}{n}\sum_{i=1}^{n} x_i$

2. Sample Median: Sort values in increasing order, then:

$$\tilde{x} = \begin{cases} \text{Middle value} & \text{If n is odd} \\ \text{Average of two middle values} & \text{If n is even} \end{cases}$$

4. Population mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

5. Population median: $\tilde{\mu} = \text{median of} \{x_1, x_2, \ldots, x_n\}$

The median generally gives a good idea of the center of a sample of a population, whether the sample is symmetric or asymmetric, whether skewed or not. We use the mean because (A) people are used to it, and (B), the sample mean has very nice mathematical properties.

6. Trimmed mean: omit the largest and smallest value, then calculate the mean of the remaining values. This reduces the influence of the most extreme value on either end.

Lecture 2

(§1.4) Measures of Variability or Spread

Why do we care how spread out a data set is? Isn't it enough to just know something about a (typical) middle value?

Example: Suppose you hear that three different instructors assign an average grade of "B" (3.0) in a course you're considering. How might you decide which instructor's class to sign up for?

The three instructors give out grades thusly:

- All B's
- C → A evenly spread out
- Half the class gets C's, half get A's

There are several measures of variability:

1. Range = max - min. Easy to calculate, but isn't very good as a measure of spread, for example: $1, 5, 5, 5, 9 : (9 - 1) = 8$. Even though there are three 5s, the range is higher than all but one member of the sample.

2. Population Variance = σ^2 = average of the squared deviations from the population mean.

Example: Amounts that all 5 of your friends owe you (You only have 5 friends):

$$\{20, 30, -10, 20, 90\}$$

$N = 5$ = population or sample size

---

x_i	$x_i - \mu$	$(x_i - \mu)^2$
20	−10	100
30	0	0
−10	−40	1600
20	−10	100
90	60	3600
150	0	5400

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{5}(150) = 30$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{5}(5400) = 1080$$

Population standard deviation:

$$\sigma = \sqrt{\sigma^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{1080} \approx 32.86$$

2'. Sample Variance: S^2

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

S^2 = average of the squared deviations from the sample mean, $(x_i - \overline{x})$

Example: Amounts that five of your 528 Facebook friends owe you:

$$\{20, 30, -10, 20, 90\}$$
$N = 5$ = population or sample size

---

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
20	−10	100
30	0	0
−10	−40	1600
20	−10	100
90	60	3600
150	0	5400

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}(150) = 30$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{4}(5400) = 1350$$

Note: Sample variance gives a bigger value than what the population variance formula gave:

$$S^2 > \sigma \Rightarrow 1350 > 1080$$

An alternative formula for S^2, which is easier to compute by hand:

$$S^2 = \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2\right]$$

This formula gives the same result, it's simply that the algebra has been reformatted for human computational ease. It's sometimes referred to as the "computational formula".

We sometimes prefer to use σ rather than σ^2, and S rather than S^2 because:

- Units for S^2 square the units of observation such as square gallons, square people, etc. S takes $\sqrt{S^2}$, thus we end up with the same units we started with.

R code for sample variance σ

```
> xx <- c(20, 30, -10, 20, 90)
> var(xx)
[1] 1350
> sd(xx)
[1] 36.74235
```

Some properties of the sample variance formula:

How is the variance affected if I add a constant c to all numbers in the data set?

Example: $c = 3$

$$\{1,3,5,11\} \Rightarrow S_1^2 = 18.\overline{666}$$
$$\{1,3,5,11\} + 3 = \{4,7,8,14\} \Rightarrow S_2^2 = 18.\overline{666}$$

Adding a number to all values doesn't change the variance.

How is the variance affected if I multiply all the numbers in the data set by some constant k?

Example:

$$\{1,3,5,11\} \Rightarrow S_1^2 = 18.\overline{666}$$
$$\text{for} \quad k = 2 : \{1,3,5,11\} * 2 = \{2,6,10,22\} \Rightarrow S_2^2 = 74.\overline{666} = 4(S_1^2)$$
$$\text{for} \quad k = -3 : \{1,3,5,11\} * -3 = \{-3,-9,-15,-33\} \Rightarrow S_4^2 = 168 = 9(S_1^2)$$

Multiplying the dataset changes the variance by k^2.

If we're asked to calculate the variance of some list of numbers, when should we use the formula for σ^2 and when should we use S^2?

You can only tell by looking at the context. Look for keywords:

- 5 of my Facebook friends $\implies$ sample $\implies S^2$
- My 5 Facebook friends $\implies$ population $\implies \sigma^2$

Box Plots: Another way to plot a data set. Box plots indicate the center, spread, symmetry, and outliers in the data set.

Definitions:

- Q_1 = First Quartile = median of smallest half of values
- Q_3 = Second Quartile = median of largest half of values
- $IQR = f_s$ = Fourth Spread = $Q_3 - Q_1$

If n is odd, the middle value is used when calculating both Q_1 and Q_3.

Example: $\{-1,1,4,7,\vert 12,13,20,22\}$ $n = 8:$ even

$$Q_1 = \frac{1+4}{2} = 2.5$$
$$Q_3 = \frac{13+12}{2} = 16.5$$
$$IQR = Q_3 - Q_1 = 16.5 - 2.5 = 14$$

Example: $\{9,4,6,11,21,23,55\}$ $n = 7:$ odd

$$Q_1 = \frac{4+6}{2} = 5$$
$$Q_3 = \frac{21+23}{2} = 22$$
$$IQR = Q_3 - Q_1 = 22 - 5 = 17$$

Here's the recipe for constructing a boxplot ("cat and whisker plot"):

1. Draw a horizontal line that extends from the smallest to largest values in your data set.
2. Draw a rectangle with vertical lines at Q_1, Q_2, and Q_3. (Q_2 = median)
3. If $x_i Q_1 - 1.5 * IQR$ *or* $x_i > Q_3 + 1.5 * IQR$, then x_i is considered an outlier. Put a dot at the locations of outliers.
4. Draw whiskers that extend from the rectangle to the most extreme non-outlying observation.

Example: $\{1,5,7,18,20,22,50\}$