Kaleb Burris
STAT F300

(§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:

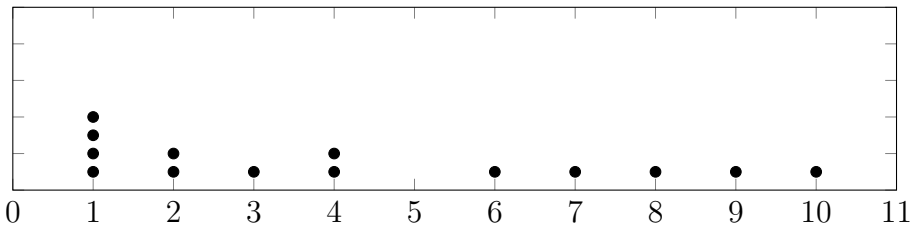
Stem	Leaves
0	2 2 2 3 9
1	4 8 9
2	0 1 1 2 2 9
3	0 2 2 3
4	4 7

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.

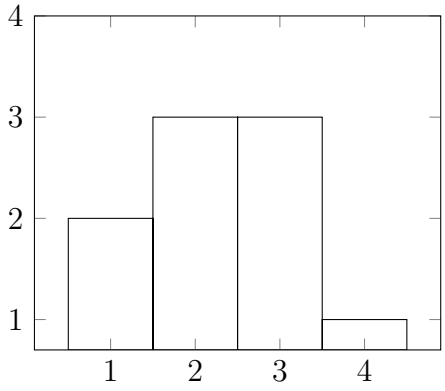
- $x_1 = \#$ of tuberculosis cases in the 50 largest U.S. cities in 2013. ← Might be 0, 1, 2, ... but $\neq \infty$.
- x_i # of "successes" when 100 patients treated. ← Count data in which we know a hard upper limit (in this case 100).
- x_i # of rooms in a rental unit. ← 1, 2, ..., 100. At least 1 guaranteed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

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

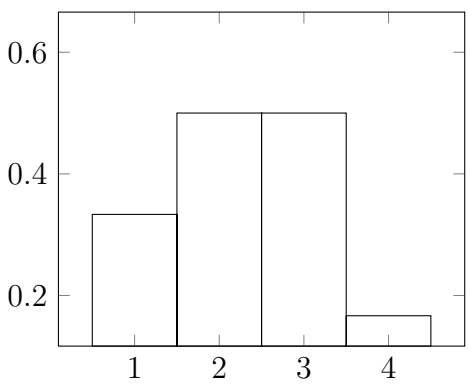
Frequency Histogram vs Relative Frequency Histogram

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: $\tilde{\mu}$ = 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.