Measures of Spread

Statistics overview

- From a given random process, we can make **probability** statements about what we expect to happen.
  - The random process can be:
    - A random sample (of size n) is taken from a much larger population
    - A random number generator is rolled n times.
- From a given sample, we can **infer** what the population/spinner looks like.

Measures of spread

- Range
- Inter-quartile range (IQR)
- Mean absolute deviation (MAD, average absolute deviation, AAD)
- Standard deviation
- (Bessel corrected) sample standard deviation.

Range

- Range is the difference between maximum and minimum.

$$\text{Range} = \max - \min$$

- Example:

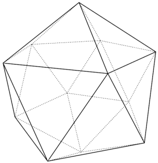
$$\text{sample} = 8,5,20,6,5,4,19$$

$$\text{range} = 20 - 4 = 16$$

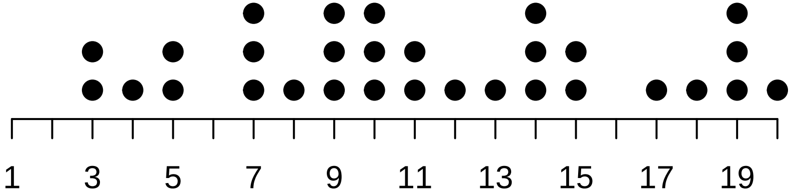
The problem with range...

- A sample's range often underestimates the population's range.
- A sample's range often underestimates a spinner's range.
- When we use a sample statistic (like sample's range) to estimate a population parameter (like population's range), we call that sample statistic an "estimator".
- Range is a **biased estimator**.

Example: icosahedron



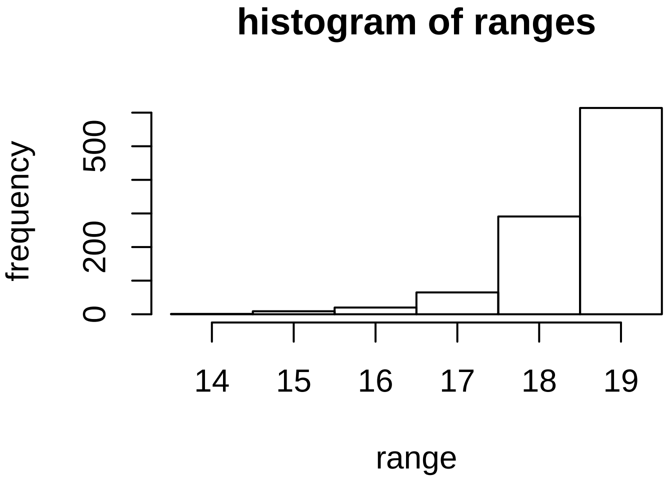
- Take a 20-sided die. It should be equally likely to land on any integer between 1 and 20. (Discrete Uniform Distribution)
- We say the population has a range of 19, because 20-1=19.
- I rolled 30 20-sided dice:



- The sample range is 17.
- Is it possible for a sample range to be larger than the population range?

Repeat the 30 rolls many times.

I "rolled" a 20-sided die 30 times, 1000 times. For each sample of 30 rolls, the sample's range was determined. Notice the sample range never was larger than the population range.



IQR

- We will come back to inter-quartile range another day.
- IQR is the difference between the 75th percentile and the 25th percentile.
- The 75th percentile is the smallest value larger (or equal) to 75% of the other values.

Mean absolute deviation

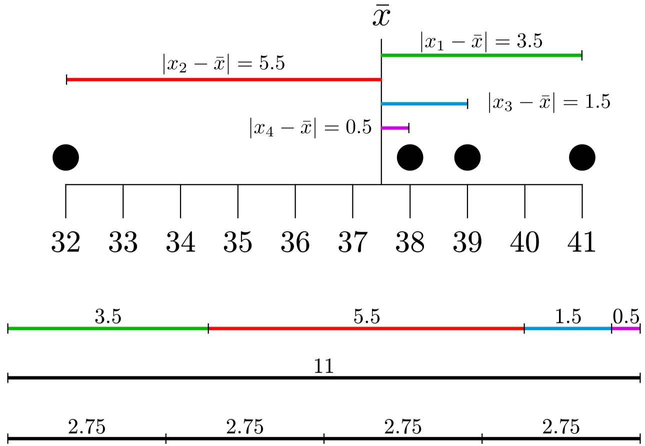
$$\text{MAD} = \frac{\sum |x - \bar{x}|}{n}$$

The mean absolute deviation represents how far from center the values are on average.

- Example: sample = $\{41, 32, 39, 38\}$

x	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$
41	3.5	3.5
32	-5.5	5.5
39	1.5	1.5
38	0.5	0.5
========	========	========
$\sum x = 150$		$\sum \lvert x - \bar{x} \rvert = 11$
$\bar{x} = 37.5$		$\text{MAD} = \frac{11}{4} = 2.75$

Mean absolute deviation



$$\text{MAD} = \frac{\sum |x - \bar{x}|}{n} = \frac{3.5 + 5.5 + 1.5 + 0.5}{4} = \frac{11}{4} = 2.75$$

Standard deviation

- Standard deviation without Bessel correction:

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

- Standard deviation with Bessel correction (sample standard deviation):

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

- The Bessel correction makes the estimator less biased.

Standard deviation example

- Example: sample = $\{41, 32, 39, 38\}$

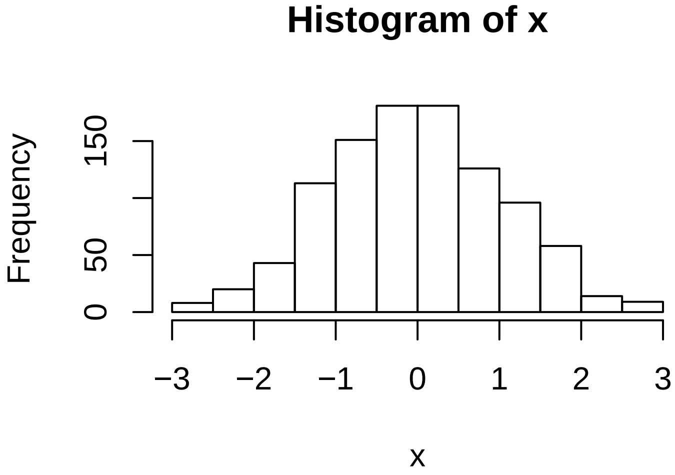
x	$x - \bar{x}$	$(x - \bar{x})^2$
41	3.5	12.25
32	-5.5	30.25
39	1.5	2.25
38	0.5	0.25
=========	=========	=========
$\sum x = 150$		$\sum (x - \bar{x})^2 = 45$
$\bar{x} = 37.5$		

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{45}{4 - 1}} = \sqrt{15} \approx 3.87$$

Estimating standard deviation from a histogram

Shape	Estimated s
Bell	$s \approx \frac{\text{range}}{6}$
Uniform	$s \approx \frac{\text{range}}{4}$
Bimodal	$s \approx \frac{\text{range}}{2}$

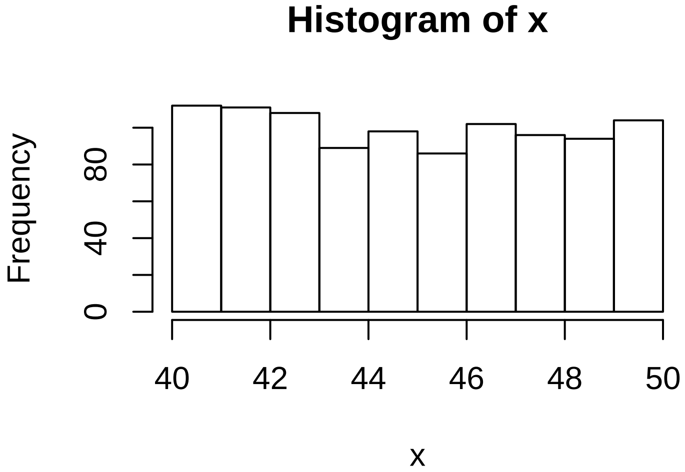
Estimating standard deviation from a histogram



$$s \approx \frac{6}{6} = 1$$

The actual value is 1.0449493

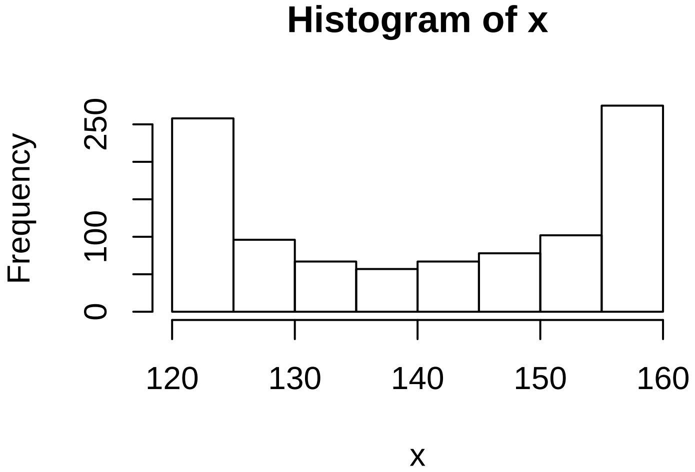
Estimating standard deviation from a histogram



$$s \approx \frac{10}{4} = 2.5$$

The actual value is 2.948021

Estimating standard deviation from a histogram



$$s \approx \frac{40}{2} = 20$$

The actual value is 15.0979203