Intro to Introductory Statistics

Chad Worley

9/13/2021

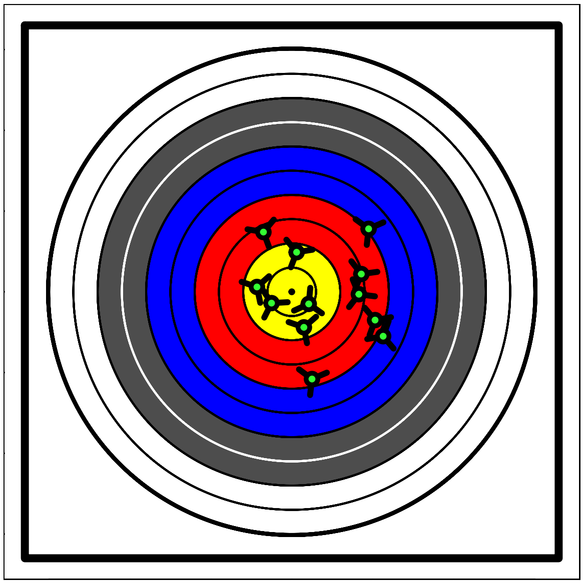
Definition of Statistics

- A "statistic" is a number that summarizes data.
- An average is an example of a statistic.
- In Statistics, we think deeply about:
  - collection of data
  - summarization of data (with statistics)
  - conclusions we can draw from data.
- Statistics is the language, tools, and logic of research.
- Statistics is quantitative epistemology - the study of knowledge itself.
- How should someone update their beliefs when provided new information?
- How should someone use data to make predictions?

Inference, soup

- In Statistics, our final goal is inference (unit 4).
- From a small **sample**, we infer about a larger **population**.
- A chef tastes soup with a spoon:
  - The spoonful is a sample.
  - The statistic = "delicious".
  - The chef infers the whole soup tastes delicious.
  - Hopefully the sample (spoon) was **representative** of the population (pot).

Archery example



- sample = 12 shots
- population = infinite potential shots under these conditions
- inference = should this archer adjust her aim?
- statistics:

$$\bar{x} = 37 \text{ mm}$$

$$s_x = 56.6 \text{ mm}$$

Basketball freethrows example

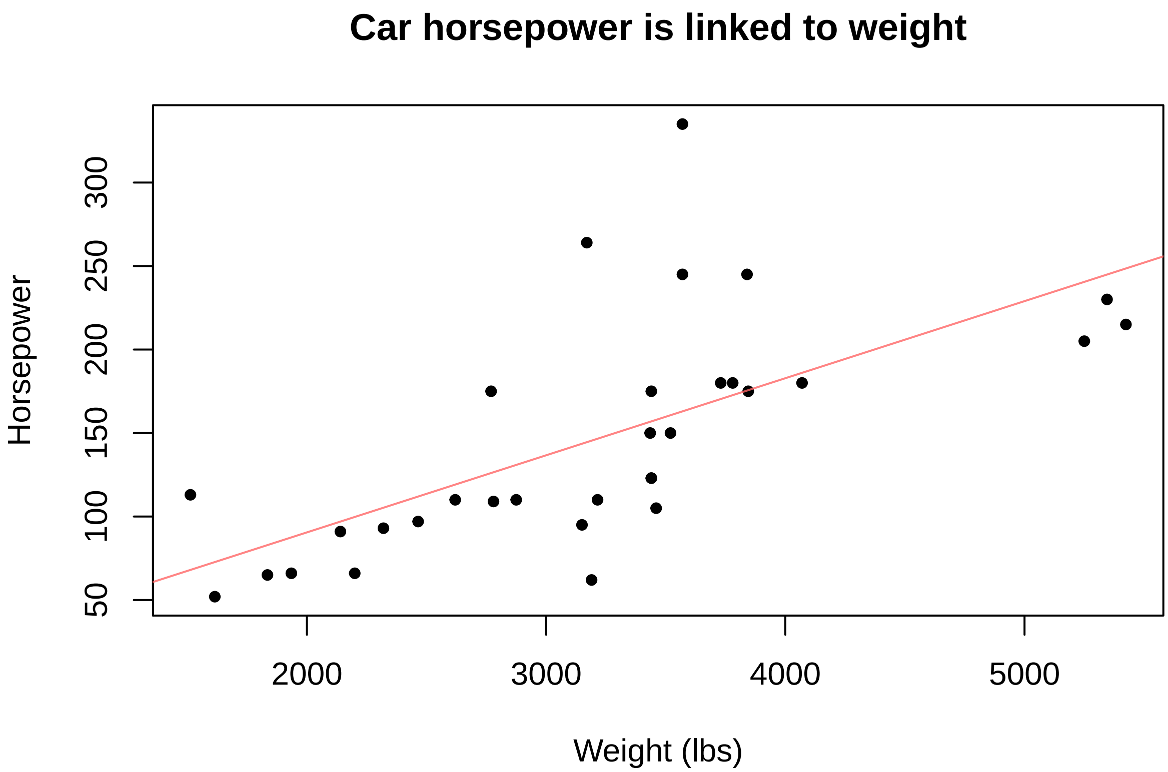
- In 2020-21 regular season, Chris Paul attempted 181 freethrows and made 169 of them (93.4%).
- Damian Lillard attempted 483 and made 448 of them (92.8%).
- Can we conclusively say Chris Paul is a better freethrow shooter?
- 2 samples: the 181 attempts and the 483 attempts
- 2 populations: the infinite potential attempts.

```r
prop.test(c(169,448),c(181,483))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(169, 448) out of c(181, 483)
## X-squared = 0.011223, df = 1, p-value = 0.9156
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.04062479  0.05295564
## sample estimates:
##    prop 1    prop 2
## 0.9337017 0.9275362
```

- The inference: it is very plausible that chance accounts for the difference.

Correlation is not causation

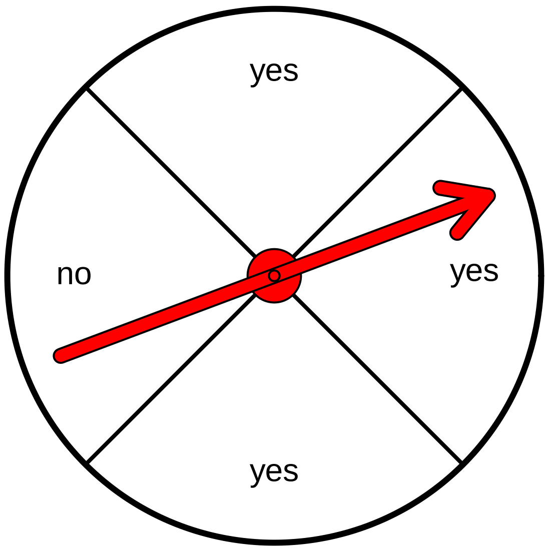


- Sample = 32 cars from 1974 Motor Trend magazine
- Population = all cars (from 1974)
- Inference = weight and hp are positively correlated
- Cars that weigh more tend to use higher horsepower.
- However, you won't increase the horsepower by filling a car with bricks.

Types of data

- **Binary**
  - 2 options: yes/no, success/fail, 0/1, orange/not orange
- **Discrete**
  - limited numerical options: dice rolls, # stairs between floors
  - count noun, "how many"
- **Continuous**
  - infinite numerical options: exact timing, exact distance
  - mass noun, "how much"
- **Categorical** (not examined much in introductory stats)
  - 2 or more nonnumeric options: favorite color

Binary spinner



- Population = the infinite potential spins
- Population parameters:

$$\text{population proportion} = p = 0.75$$

- As an example, imagine spinning 10 times.
- Sample: 1, 0, 1, 0, 0, 1, 1, 0, 1, 1
- Sample statistics:

$$\text{sample size} = n = 10$$

$$\text{sample success count} = 6$$

$$\text{sample proportion} = \hat{p} = 0.6$$

```
x = c(1,0,1,0,0,1,1,0,1,1)
n = length(x)
success_count = sum(x)
phat = success_count/n
print(n)
```

```
## [1] 10
```

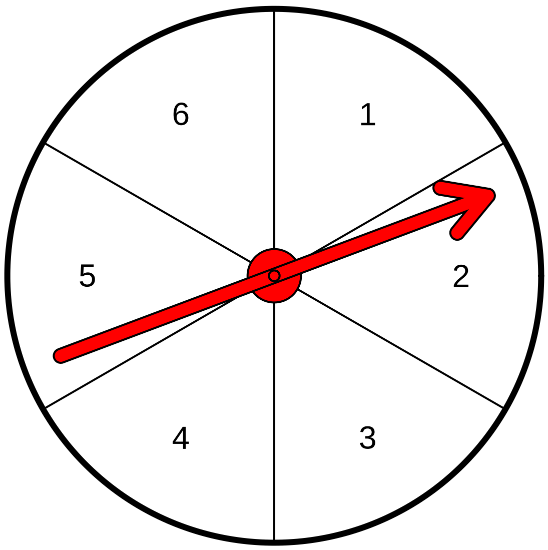
```
print(success_count)
```

```
## [1] 6
```

```
print(phat)
```

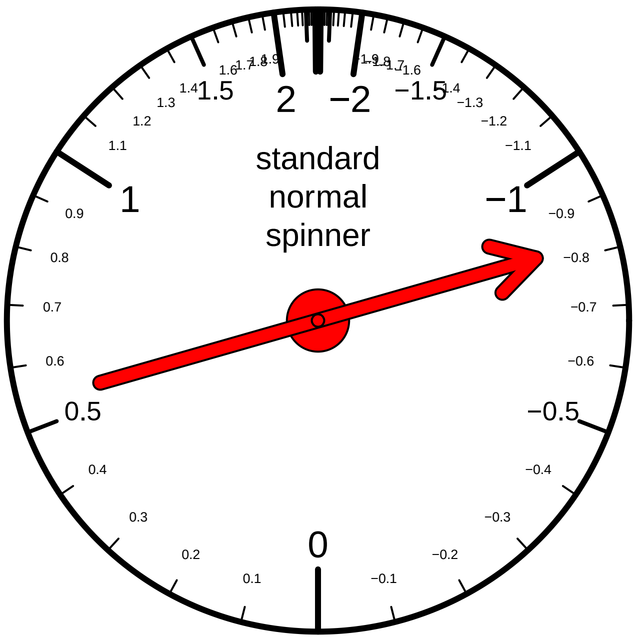
```
## [1] 0.6
```

Discrete spinner



- population mean (unit 2)
$$\mu = 3.5$$
- population standard deviation (unit 2)
$$\sigma = 1.7078251$$
- sample
$$3, 6, 3, 2, 2, 6, 3, 5, 4, 6$$
- sample size
$$n = 10$$
- sample total (sample sum)
$$\sum x = 40$$
- sample mean
$$\bar{x} = 4$$
- sample standard deviation
$$s = 1.63$$

Continuous spinner



- population mean (unit 3)

$$\mu = 0$$

- population standard deviation (unit 3)

$$\sigma = 1$$

- sample: -1.2070657, 0.2774292, 1.0844412, -2.3456977, 0.4291247, 0.5060559, -0.57474, -0.5466319, -0.564452, -0.8900378
- sample size

$$n = 10$$

- sample total (sample sum)

$$\sum x = -3.8315741$$

- sample mean

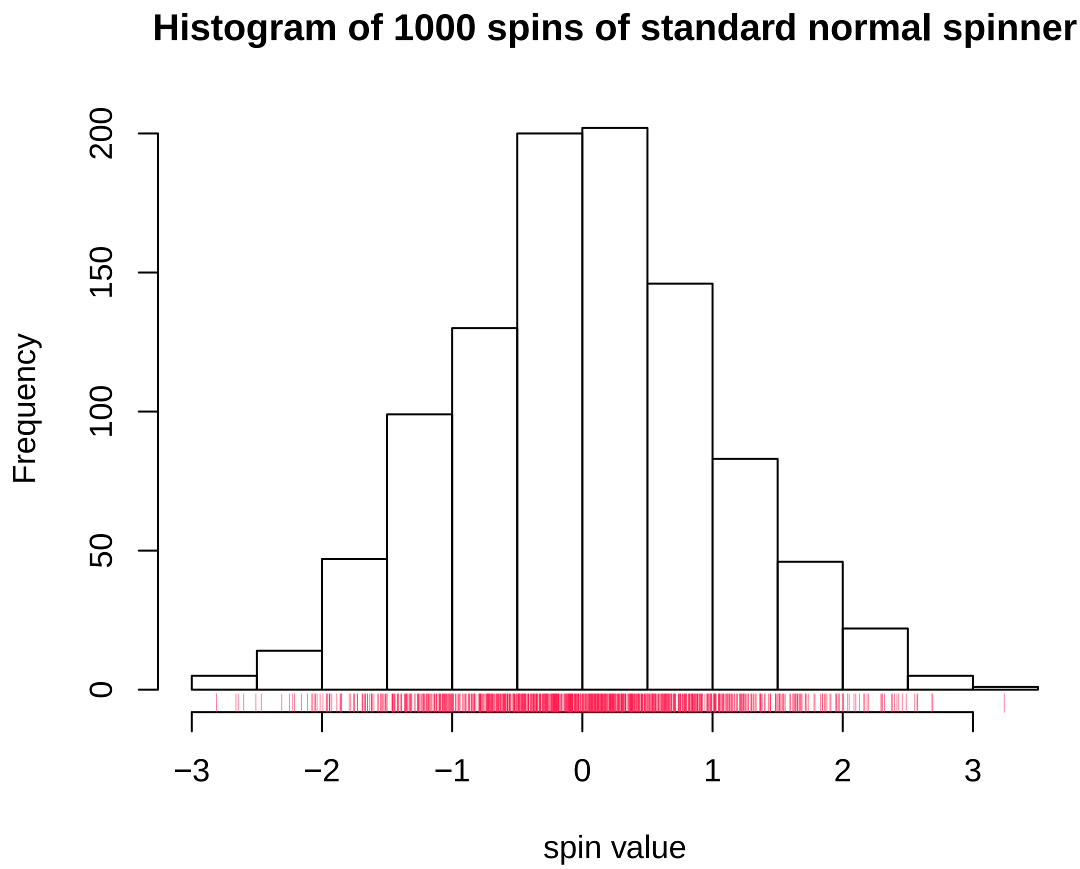
$$\bar{x} = -0.3831574$$

- sample standard deviation

$$s = 0.9958$$

Histogram and frequency table

```
x = rnorm(1000)
hist(x,main="Histogram of 1000 spins of standard normal spinner",xlab="spin value")
rug(x,col=rgb(1,0.1,0.3,0.5))
```



interval

frequency

-3 to -2.5

5

-2.5 to -2

14

-2 to -1.5

47

-1.5 to -1

99

-1 to -0.5

130

-0.5 to 0

200

0 to 0.5

202

0.5 to 1

146

1 to 1.5

83

1.5 to 2

46

2 to 2.5

22

2.5 to 3

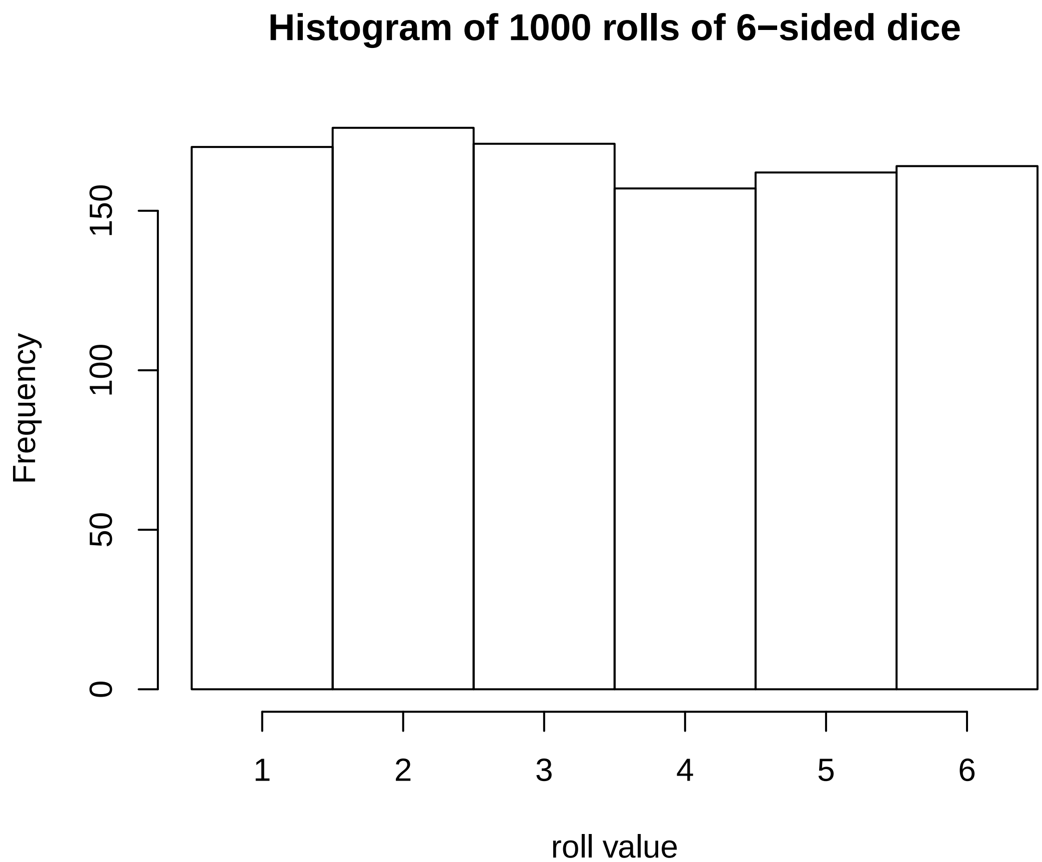
5

3 to 3.5

1

Dice hist and freq tab

```
x = sample(1:6,1000,T)
b = seq(0.5,6.5,1)
hist(x,main="Histogram of 1000 rolls of 6-sided dice",xlab="roll value",breaks=b)
```



interval

frequency

0.5 to 1.5

170

1.5 to 2.5

176

2.5 to 3.5

171

3.5 to 4.5

157

4.5 to 5.5

162

5.5 to 6.5

164

roll_value

frequency

1

170

2

176

3

171

4

157

5

162

6

164

Notation reference

n = sample size, how many measurements

$\#(\ldots)$ = how many measurements satisfy … criterion. (nonstandard notation)

- Binary
  - p = population proportion
  - $\hat{p}$ = "p hat" = sample proportion
- Discrete
  - μ = "mu" = population mean
  - σ = "sigma" = population standard deviation
  - $\sum x$ = "sum of x" = sample total
  - $\bar{x}$ = "x bar" = sample mean
  - s = sample standard deviation (with Bessel correction by default)
- Probability
  - $P(\ldots)$ = probability that … happens