Confidence Intervals and Bootstrap

Stat 120

January 20 2023

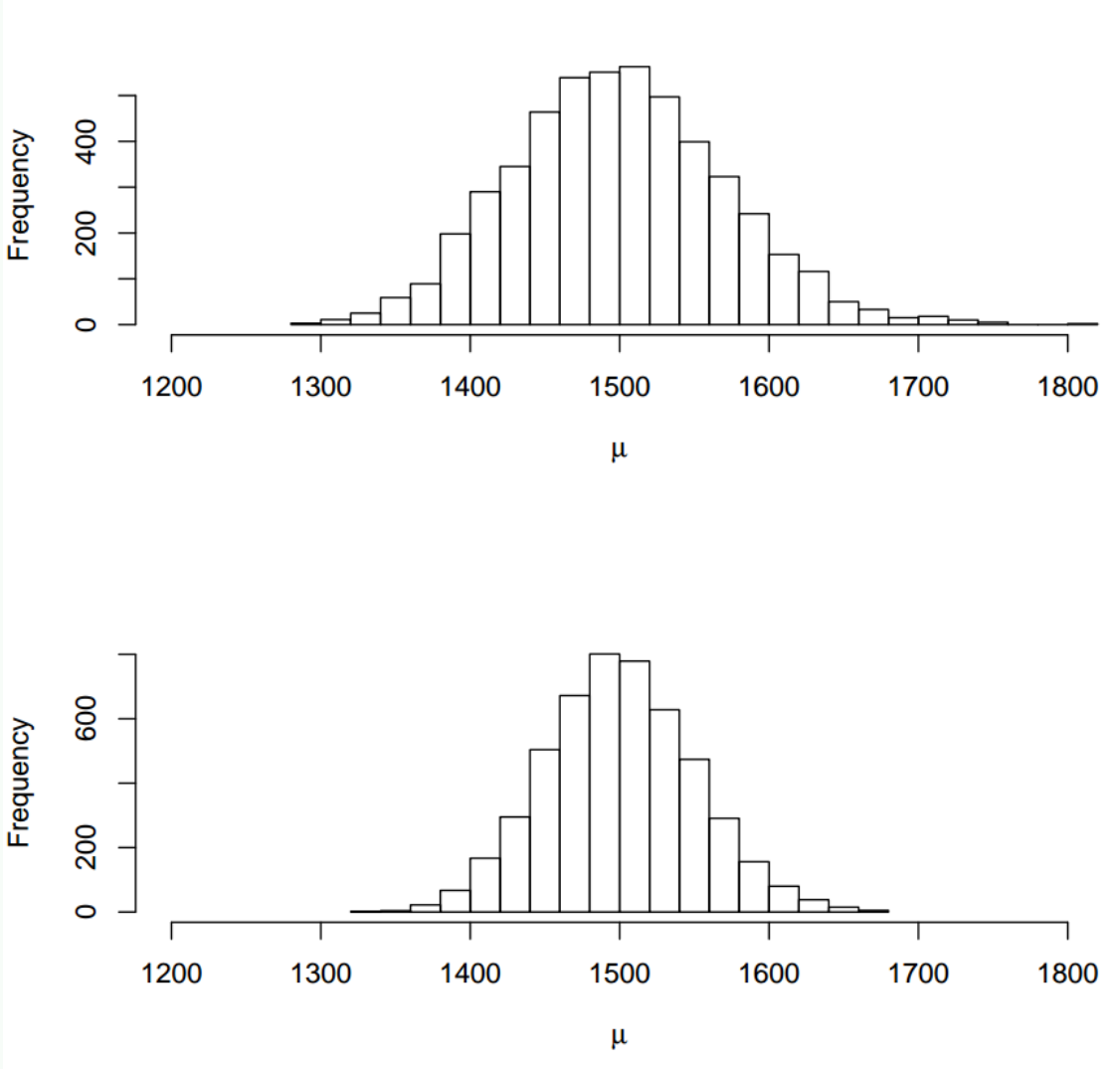
Question!

The higher the standard error of a statistic, the ..... the uncertainty surrounding the statistic.

1. higher
2. lower

Sampling distribution

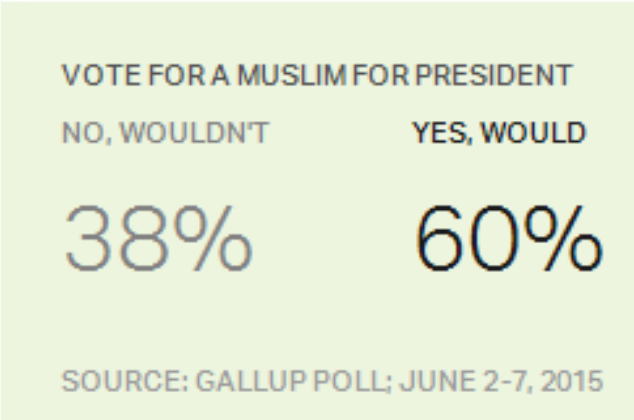
Sampling distributions. Top n=50, Bottom n=100



Interval Estimation

- ***Point estimates*** *are almost always not accurate*
- *Uncertainty in point estimates measured by the* ***Standard Error (SE)***
- *A plausible range of values for the population parameter is more reliable*
- ***Interval Estimate:*** *An interval estimate is an interval of numbers within which the parameter value is believed to fall*

A Gallup Poll



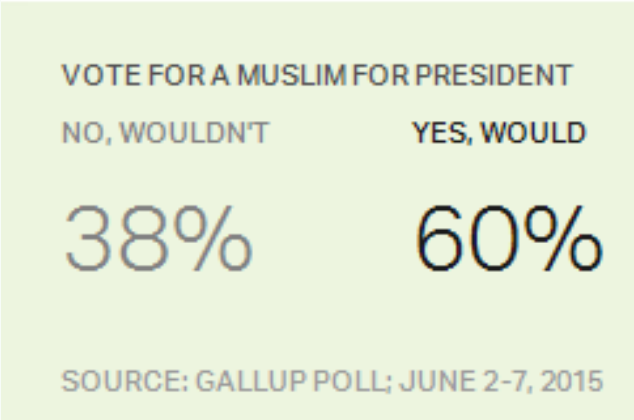
How accurate is an estimate of 60%?

Survey Methods

Results for this Gallup poll are based on telephone interviews conducted June 2-7, 2015, with a random sample of 1,527 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, the margin of sampling error is ±3 percentage points at the 95% confidence level. All reported margins of sampling error include computed design effects for weighting.

Link to the Gallup poll

A Gallup Poll



" ... the margin of sampling error is $\pm$ 3 percentage points at the 95% confidence level."

- Interval estimate: $60\% \pm 3\% = (57\%, 63\%)$
- The percentage of American adults who would vote for a Muslim for president is likely between 57% and 63%.
- Would a majority of US adults vote for a qualified Muslim presidential candidate?

Margin of Error

- *The margin of error measures how accurate a point estimate is likely to be in estimating a parameter.*
- *To determine the margin of error, we can use the statistic's sampling distribution and standard error*

Confidence Intervals

- A **confidence interval** is an interval containing the most believable values for a parameter
- A confidence interval is centered on the **point estimate** and extends a certain number of **standard errors** on either side of the estimate
- The **confidence level** tells us what percent of the intervals will contain the population parameter.

A 95% confidence interval will contain the true parameter for 95% of all samples.

Gallup Poll Result Interpretation

" ... the margin of sampling error is ± 3 percentage points at the 95% confidence level."

95% confidence means that 95% of all samples will yield a sample proportion that is within 3 percentage points of the population proportion

95% Confidence Interval

If the sampling distribution is relatively symmetric and bell-shaped, a 95% confidence interval can be estimated using

$$\text{statistic} \pm 2 \times \text{SE}$$

95% confidence means 95% of all samples will yield a statistic that is within 2 SE of the population parameter

Confidence Intervals are ...

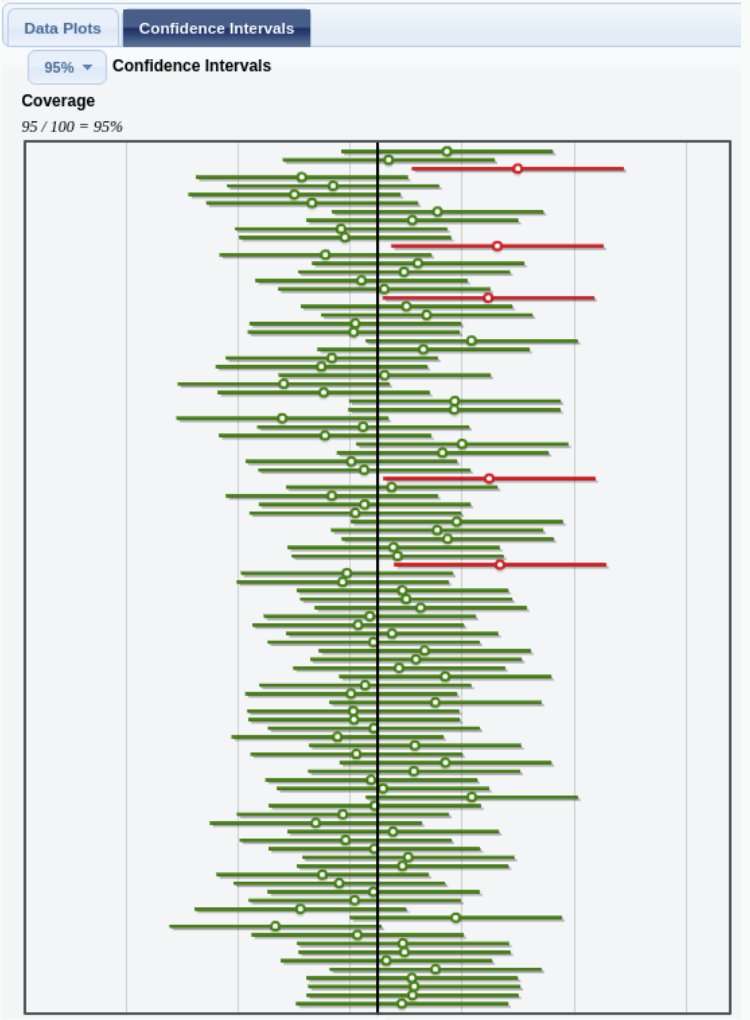
- *always about the population*
- *not probability statements*
- *only about population parameters, not individual observations*
- *only reliable if the sample statistic they're based on is an unbiased estimator of the population parameter*

A short demo

Let's all go to Statkey web app.

Take Home Points

- *The parameter is **fixed***
- *The statistic is **random** (depends on the sample)*
- *The interval is also **random** (depends on the statistic)*
- *Confidence level is the **proportion** of intervals that capture the true parameter*



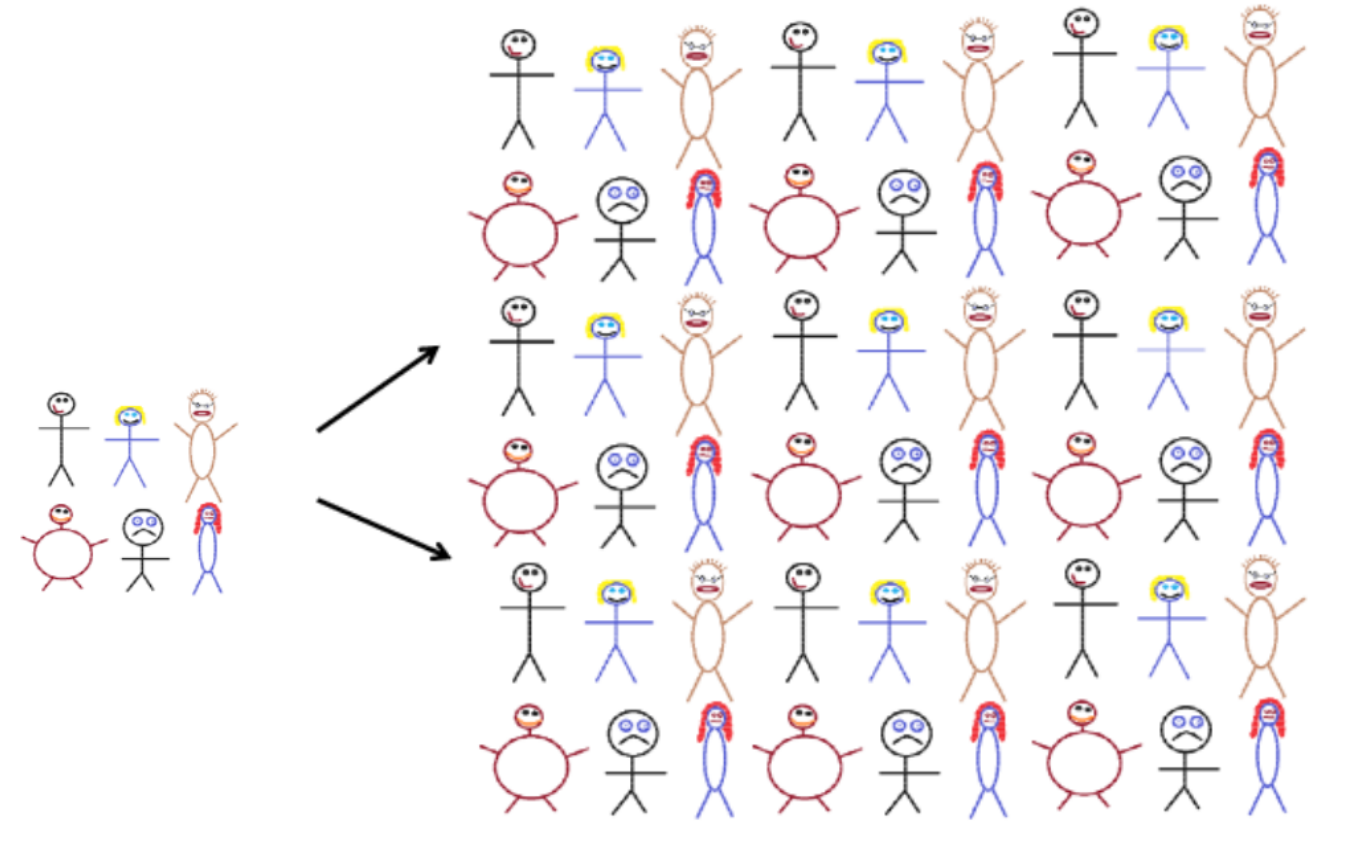
Repeated Sampling of 100 95% Confidence Intervals, Truth = Vertical Line

What to do when we only have one sample - **BOOTSTRAP!**

- **Repeated sampling** is needed to compute the standard error of a sample statistic
- Can **estimate the SE** from a bootstrap distribution
- Use this SE to compute a **confidence interval** for an unknown parameter

Bootstrap

Original sample (left) to population (right)



Creating a bootstrap sample is the same as using the data simulate a "population" that contains an infinite number of copies of the data.

Bootstrap Sampling in R

- *resample a set of observations with replacement*
- *same data points can appear multiple times*

	Data	Statistic
Original sample	$x_1, x_2, \ldots, x_n$	$\bar{x}_n$
Resample	$x_1^*, x_2^*, \ldots, x_n^*$	$\bar{x}_n^*$

```r
# R-code
# In base R
boot <- sample(x, size, replace = TRUE)
```

```r
# R-code
library(CarletonStats)
boot(x)
```

Bootstrap Steps

1. ***Generate a bootstrap sample.***
2. ***Compute the statistic of interest for your bootstrap sample.***
3. ***Repeat steps (1) –(2) many times. Plot the distribution of all your bootstrap statistics***

This is the bootstrap distribution!!

Bootstrap Distribution

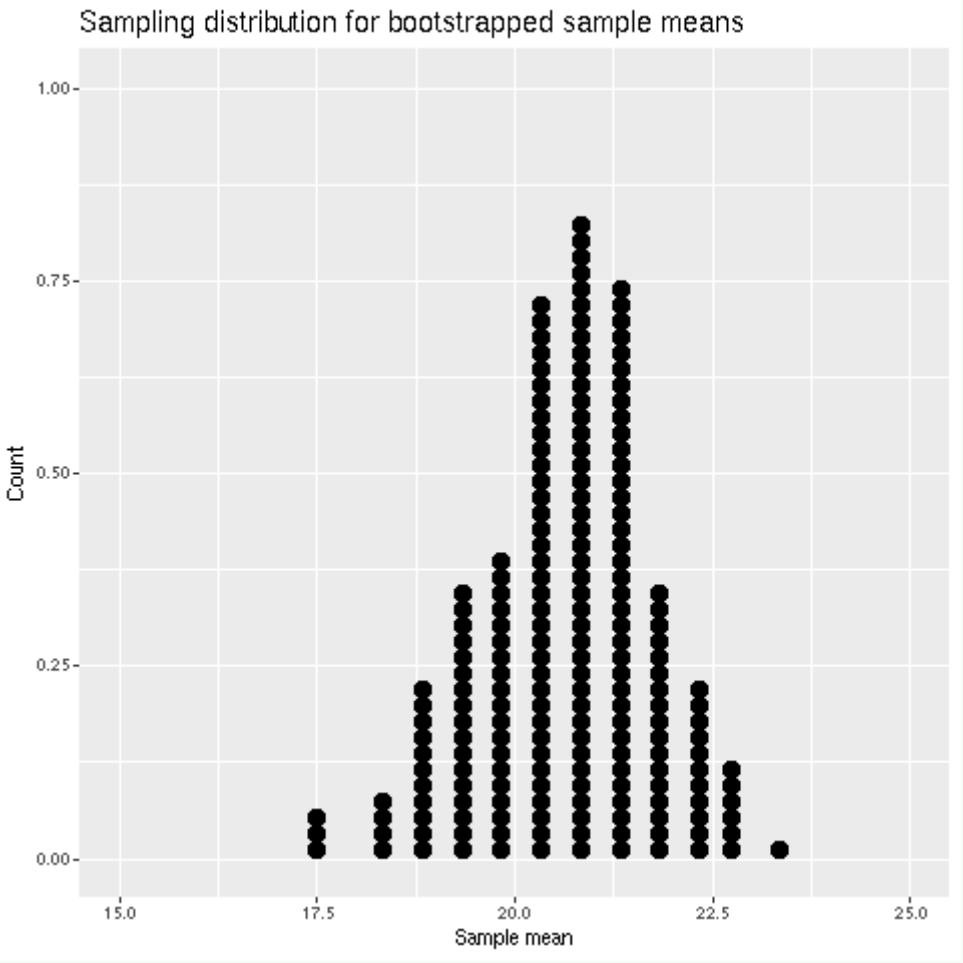
- *A **bootstrap distribution** is the distribution of many bootstrap statistics.*
- *The standard deviation of this distribution is called the **bootstrap standard error** of the statistic.*
- *The bootstrap distribution is **centered** near the original sample mean.*

Bootstrap Distribution

Suppose $X = \{20, 24, 19, 23, 22, 16\}$

$$
\begin{aligned}
X_1^* &= \{16, 19, 16, 23, 22, 24\} \\
X_2^* &= \{22, 19, 22, 19, 23, 19\} \\
X_3^* &= \{20, 22, 24, 16, 24, 16\} \\
\vdots \quad & \quad \dots \\
\vdots \quad & \quad \dots \\
X_N^* &= \{19, 24, 19, 19, 19, 22\}
\end{aligned}
$$

N = total number of simulations/samples

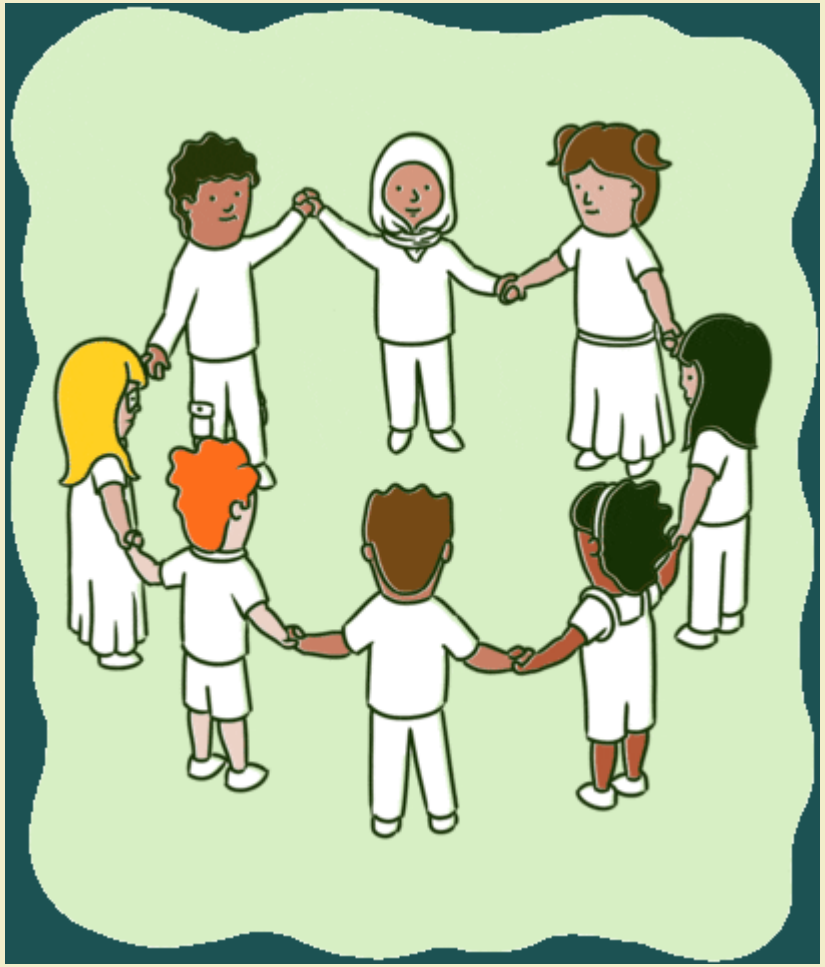


Summary

- *Interval estimates let us infer a set of plausible values for a parameter*
- *A 95% confidence interval will contain the true parameter for 95% of all samples*
- *Usually, we do not have access to the population and cannot do repeated sampling*
- *To get an estimate of standard error, can generate bootstrap samples by sampling with replacement form the original sample, using the same sample size*
- *Can use the bootstrap standard error to construct bootstrap confidence intervals*

Your Turn 1

25:00



Go over the remaining portion of in class activity and let me know if you have any questions!