Confidence intervals

Outline for today

Confidence intervals

Parametric confidence intervals: for π and μ

Computational methods: the bootstrap

Announcement: Final projects

All final project presentations are due at 11:59pm on Sunday (April 30th)

5 minute presentations: 3 slides

Final project written reports are due the Sunday after the last day of class (May 7th)

Big Data Baseball chapter 12

Good ending?

"Math had played such a big role in the pirates' turnaround."

And with this one sentiment so concludes BIG DATA BASEBALL. we can all thank statistics for rescuing the Pirates career. i am interested in what their team would be like if they didn't adopt these methods. i cant say i'm invested in analytics so this book wasn't a top seller for me, but interesting.

- Julia K

Big Data Baseball chapter 12

Quote #12: "So it's an art in the sense of what we get from the raw information doesn't always tell the whole story… At the end of the day you can make the argument that [the art] is just refining the data, but in a way there are still situations that come up where there is gray area and you have to massage through it" (p 200).

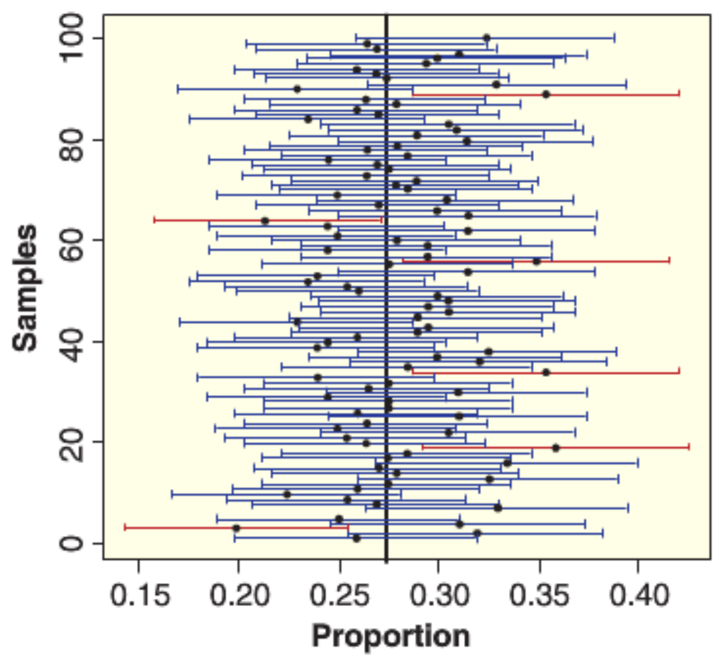
Reaction: I found this quote to be a really interesting take away from the story this book has presented. The book has outlined the inception, development and take-over of sabermetrics in the field of baseball, and its dramatic effects on team performance. Throughout the chapters, it appeared there was such reluctance to accept and listen to this data because it rejected the thought of subjective decision making by team managers and individual players. But this quote really just shows that it was a transition of how to use that subjective ability within the given data sets, not a total replacement of them.

- Helen (James had a similar Q&R)

Confidence Intervals

A **confidence interval** is an interval computed (from sample data) by a method that will capture a *population parameter* a specified proportion of times

I.e., if we apply this procedure 100 times, we will capture the population parameter say 95% of those times



Wits and wagers

E.g., how tall was the tallest man

Parameter exists in the ideal world

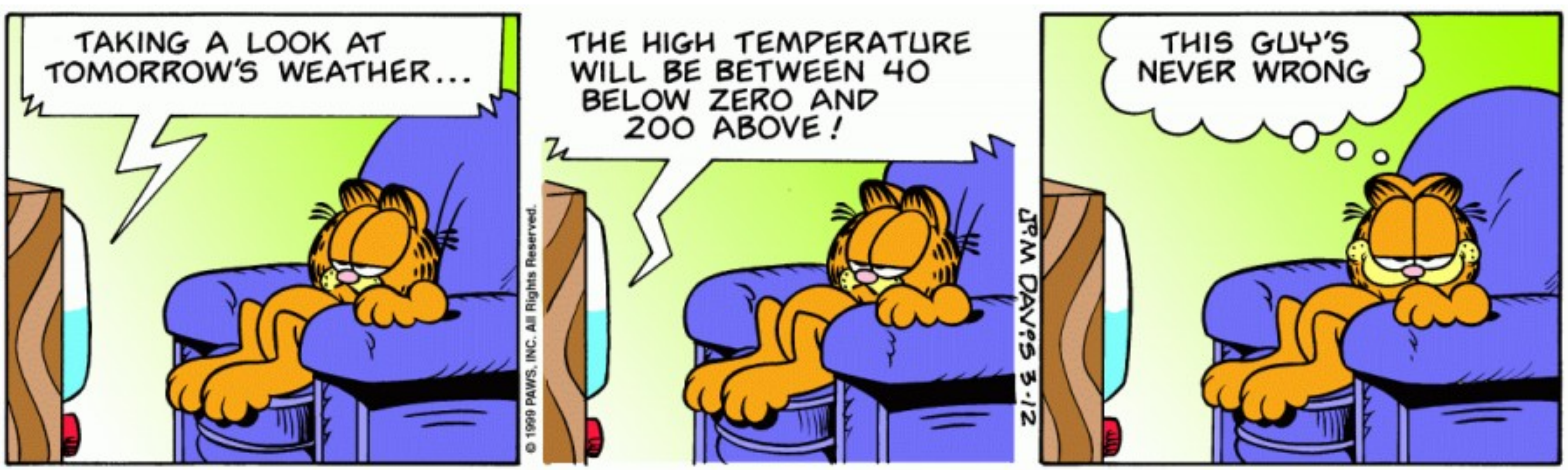
We toss intervals at it

95% of those intervals capture the parameter

Wits and wagers

There is a tradeoff between

- Interval width
- **Confidence level**: proportional intervals that will contain the true value



Example: The Gallup poll

42% of American voters identify as independents, plus or minus 2%

How do we interpret this?

It says that the <u>population parameter</u> lies somewhere between 40-44%

i.e., if they sampled all voters they would get a value within this range

- on 95% of polls conducted

Example: A-Rod's ability

In 2012, A-Rod had 529 PA and an OBP of .353

- i.e., he had 187 hits out of 529 plate appearances

Remember that we did a hypothesis test that showed it was very unlikely that A-Rod's real OBP was

.300 or less

Question: What is a range of OBP values that are likely to contain the real value of A-Rod's ability π

We want to have a procedure, that say 95% of the time, will construct an interval that contains π

Confidence Intervals

Does anyone have any ideas how we could construct such a confidence interval?

One way: the set of all values for π that we can't reject the null hypothesis in a hypothesis test

- i.e., all π such that:

$$Pr(X = 187;\, n = 529,\, \pi) \geq .05$$

Based on this method we get a confidence interval of [.319 .387] for plausible values of π

Computing confidence intervals

There are several other methods to compute confidence intervals as well, but first…

Sampling distributions

A **sampling distribution** is the distribution of our statistic of interest

Recall that in hypothesis testing, a null distribution is created by assuming the null hypothesis H_0 is true

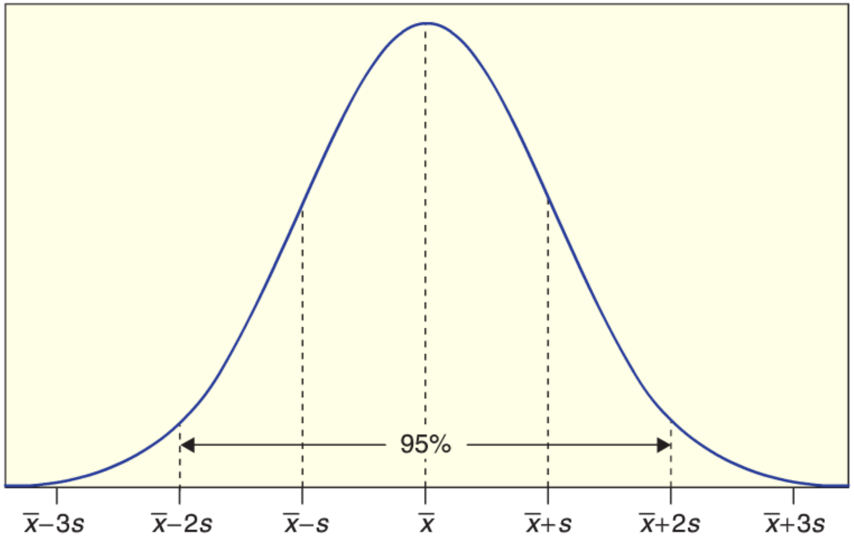
A sampling distributions is the <u>actual</u> distribution of our statistic

- If the null hypothesis is true, the sampling distribution is the same as the null distribution
- If the null hypothesis is false, then the sampling distribution is different (depends on the actual parameter value)

Computing confidence Intervals

For more many statistics, such as $\hat{p}$ and $\bar{x}$ the sampling distribution is normal (if your sample size is large enough)

- E.g., a binomial distribution becomes normal when n becomes large

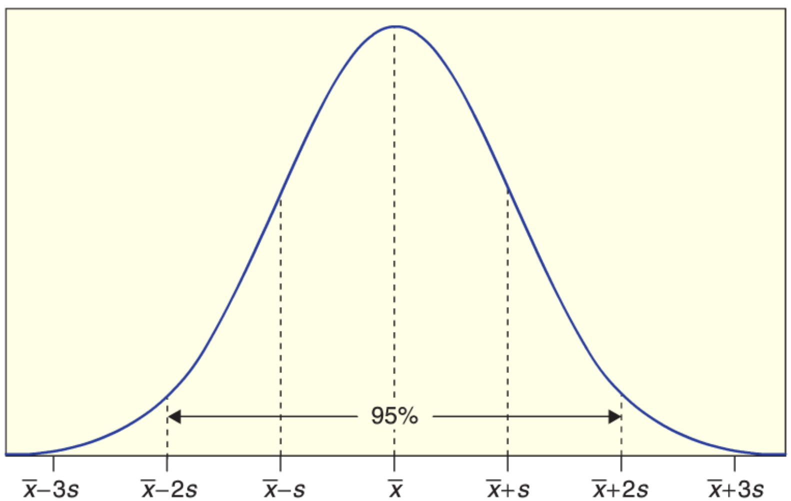


Recall that:

- 95% of the mass in a Normal distribution is within 2 standard deviations from the mean
- Thus our sample statistics ($\hat{p}$ or $\bar{x}$) will lie within 2 standard deviations of the actual parameter values (π or μ)

Computing confidence Intervals

The standard deviation of the sampling distribution is called the standard error (denoted SE)



Claim: If we knew the standard error, we could compute a confidence interval

- Why is this true?

If we add and subtract $2 \cdot SE$ to our statistic of interest ($\hat{p}$ or $\bar{x}$) , 95% of the time it will contain the population parameter

Computing confidence Intervals

There are formulas for computing the standard error

- For a sampling distribution of proportions $\hat{p}$:

$$SE = \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

- For a sampling distribution of means $\bar{x}$:

$$SE = \frac{s}{\sqrt{n}}$$

remember $\bar{x}$ comes from a t-distribution

Several methods focus on estimating the standard error and then create confidence intervals that are equal to:
statistic $\pm$ $2 \cdot SE$

Example: confidence Intervals on a proportion (π)

A 95% confidence interval is: $\hat{p} \pm 2 \cdot SE$

where

$$SE = \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

For A-Rod's 2012 data we have:

- OBP = .353, PA = 529
- What is $\hat{p}$ and n?
- What is SE?
- What is a 95% CI?

Confidence Intervals

$$SE = \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

Confidence interval is : $\hat{p} \pm 2 \cdot SE$

For A-Rod's 2012 data we have:

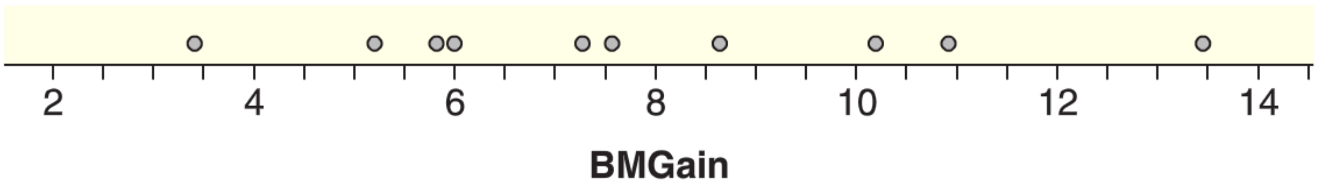
- $\hat{p}$ = .353, n = 529
- SE = .021
- 95% confidence interval is [.311 .394]

How would the SE and the confidence interval change if A-Rod had been up 1,000 times?

Example, confidence Intervals on a mean (μ): Fat mice

A study kept a light on at night which allowed mice to eat at night when they typically are resting. These mice gained a significant amount of weight compared to control mice kept in darkness which ate the same amount of calories.

The 10 mice with light gained an average of 7.9g with a standard deviation of 3.0g.



Find an approximate 95% CI for the amount of weight gained

$$CI = \bar{x} \pm 2.26 \cdot \frac{s}{\sqrt{n}}$$

A bit bigger than 2 because $\bar{x}$ Comes from a t-distribution with n - 1 = 9 degrees of freedom

Light at night makes mice fatter

What is the parameter we are estimating?

$$CI = \bar{x} \pm 2.26 \cdot \frac{s}{\sqrt{n}}$$

$\bar{x}$ = 7.9,

s = 3,

n = 10

t* = qt(.975, df = 9) = 2.26

7.9 ± 2.26 · 3/3.16 = [5.75 10.05] grams

Bootstrap standard errors

Plug-in principle

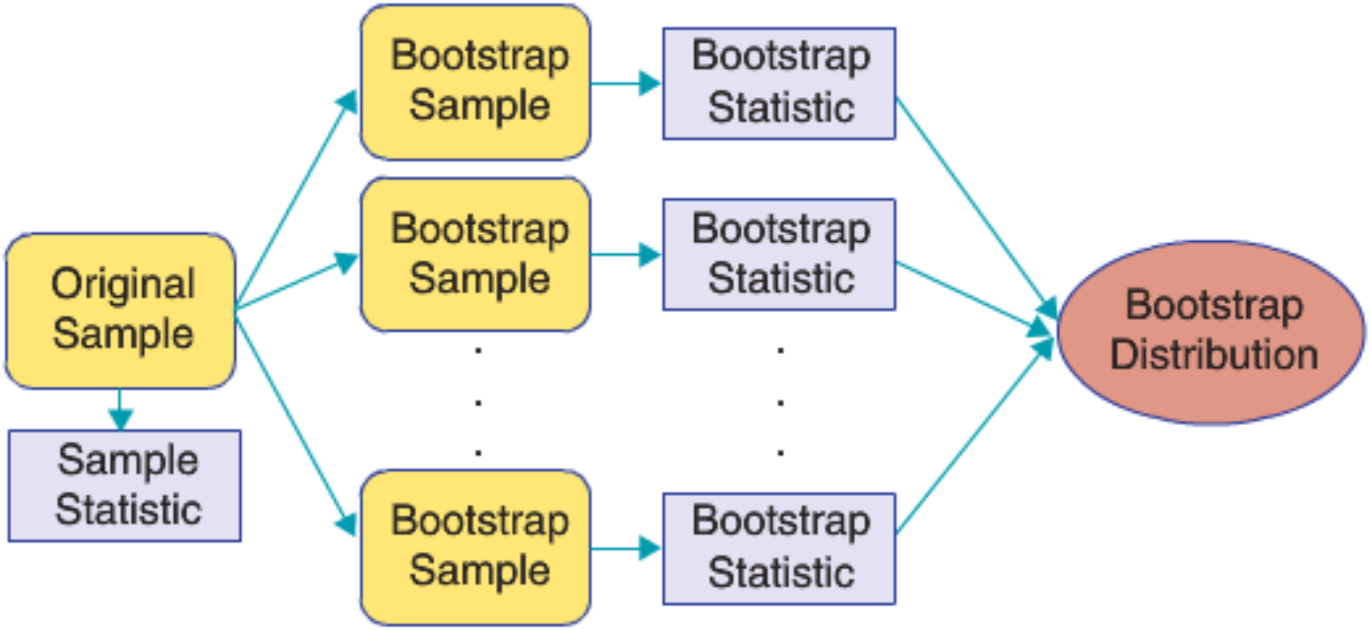
Suppose we get a sample from a population of size *n*.

To calculate standard errors, we substitute our sample for the full population (plug-in principle)

We then sample *n* points <u>with replacement</u> from our sample, 1000's of times and get a bootstrap sample distribution.

The standard deviation of this distribution (standard error of the bootstrap distribution) is a good approximate for the standard error of the real sampling distribution.

Bootstrap process



95% Confidence Intervals

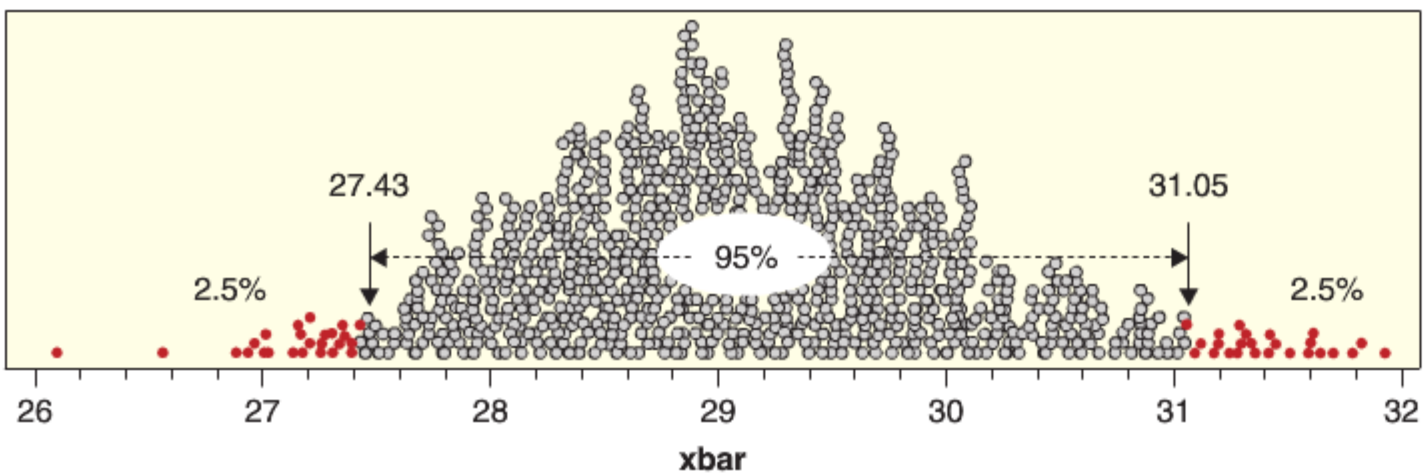
When a bootstrap distribution for a sample statistic is symmetric and bell-shaped, we can estimate a 95% confidence interval using:

$$\text{Statistic} \pm 2 \cdot SE$$

Where SE is the standard error estimated using the bootstrap

What if the bootstrap distribution is not bell-shaped?

If the bootstrap distribution is approximately symmetric, we can use percentiles in the bootstrap distribution to an interval that matches the desired confidence level.



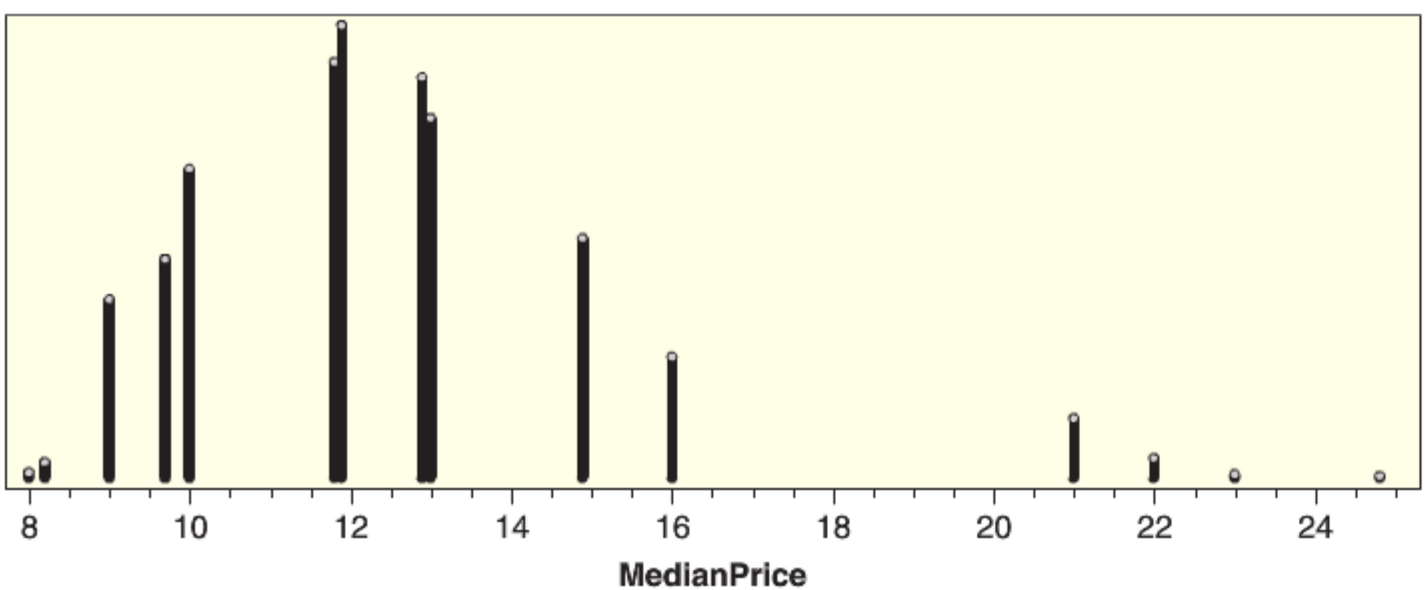
Findings CIs for many different parameters

This bootstrap method works for constructing confidence intervals for many different types of parameters

Caution: the bootstrap does not always work

Always look at the bootstrap distribution, if it is poorly behaved (e.g., heavily skewed, has isolated clumps of values, etc.), you should not trust the intervals it produces.

Median prices of Mustangs



Let’s explore bootstrap distributions

http://asterius.hampshire.edu:3838/intro_stats/boostrap_sampling_distribution/

What are the steps needed to create a bootstrap SE?

1) Start with a sample

2) Resample the points in the sample to get a bootstrap sample with replacement

3) Compute the statistic of interest on the bootstrap sample

4) Repeat steps 2-3 10,000 times

5) Take the standard deviation of this resampling distribution

Example: A-Rod's ability

In 2012, A-Rod had 529 PA and an OBP of .353

- i.e., he had 187 hits out of 529 plate appearances

How can we create a vector that has this data?

```
arod <- c(rep("Hit", 187), rep("Out", 529-187))
```

How can we verify this vector is correct?

- Check that we get the right OBP (i.e., the right $\hat{p}$)

```
obs.stat   <-   sum(arod == "Hit")/529
```

What are the steps needed to create a bootstrap SE?

```
boot_dist <- NULLfor (i in 1:10000) {

    bootstrap_sample <- sample(arod, replace = TRUE)
    boot_dist[i] <- sum(bootstrap_sample == "Hit")/529

}


hist(boot_dist)

SE.arod <- sd(boot_dist)

CI.lower  <- obs.stat - 2 * SE.arod
CI.upper  <- obs.stat + 2 * SE.arod
```

Comparing CIs

Bootstrap

- [.312   .395]

Formula:

- [.311 .394]

Final projects

All final project presentations are due at 11:59pm on Sunday (April 30th)

5 minute presentations: 3 slides

Final project written reports are due the Sunday after the last day of class (May 7th)