

# Uncertainty

Jason Willwerscheid



2022-10-31

↪

↪

# Review

Last week we started discussing statistical modelling. Recall the difference from exploratory data analysis:

- ▶ Exploratory data analysis: Stephen Curry has made 41 of 44 free throws. 
- ▶ Statistical model: Stephen Curry's free throws follow a binomial distribution with  $n = 44$  and  $p = 0.93$ . (There are other possible models! For example, maybe we have reason to think  $p$  is higher or lower.) 

# Review

So what is the point of statistical modelling?

- understand how data is generated
- gives range of outcomes
- predict new values

# Review

So what is the point of statistical modelling?

1. To understand the data-generating process.
2. To predict future results.
3. To be able to describe uncertainty in the results.

# Review

To understand the data-generating process:

- ▶ For example, if the binomial model is an accurate model, then the free throws are independent. In other words, there is no such thing as a “hot hand.” Exploratory data analysis describes the data *as it is* and says nothing about how the data was generated.

# Review

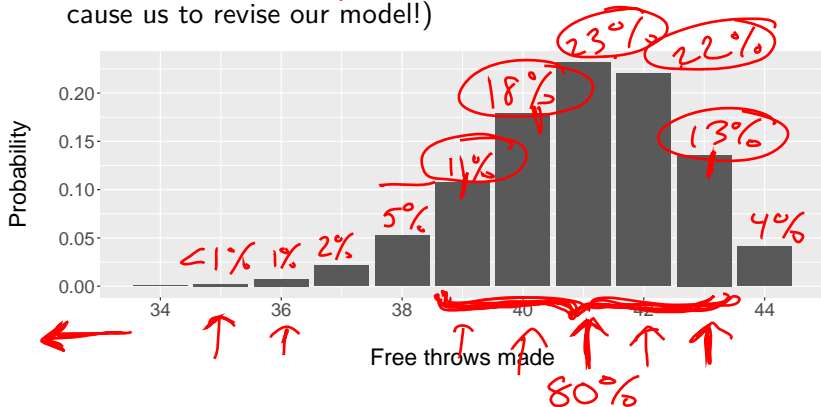
To predict future results:

- ▶ If  $p = 0.93$  is a good estimate, then we can expect Curry to make roughly 41 of his *next* 44 free throws. If, on the other hand,  $p = 0.86$  is a better estimate, then we should expect him to make closer to 38 of 44. Without at least an implicit model, it is impossible to know what new data will look like.

# Review

To be able to describe uncertainty in (new) results:

- ▶ If  $p = 0.93$  is a good estimate, then 39 of 44 free throws made would not be terribly surprising (it's only about half as likely as 41 of 44), but 35 of 44 would seem very unlikely (and might cause us to revise our model!)



# Uncertainty

Two kinds of uncertainty:

- ▶ Uncertainty in *new* results (given a statistical model).
- ▶ Uncertainty in the *model*: for 41 of 44 free throws,  $p = 0.93$  and  $p = 0.9$  are both pretty plausible.

$$X \sim \text{Binomial}(n=44, p=0.93)$$



# Prediction Intervals

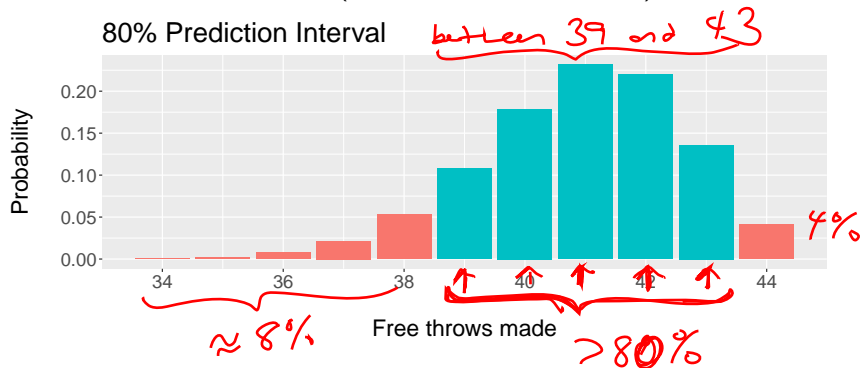
Uncertainty in *new* results (*given* a statistical model):

- ▶ One way to describe uncertainty is via **prediction intervals**.  
Look at the distribution and observe where 80% (or 90%, or 95%, or 99%) of results occur.



# Prediction Intervals

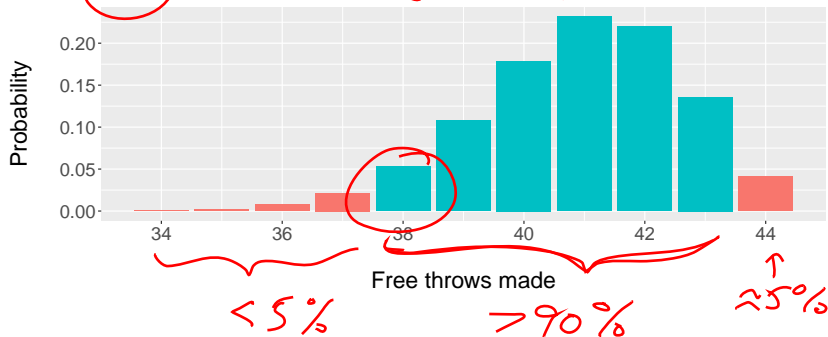
Uncertainty in *new* results (given a statistical model):



# Prediction Intervals

Uncertainty in *new* results (given a statistical model):

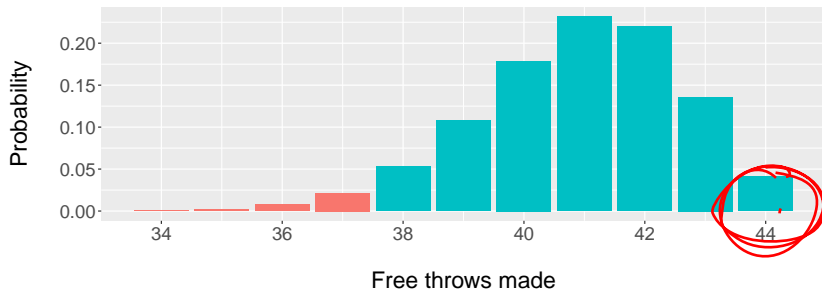
90% Prediction Interval between 38 and 43



# Prediction Intervals

Uncertainty in *new* results (*given* a statistical model):

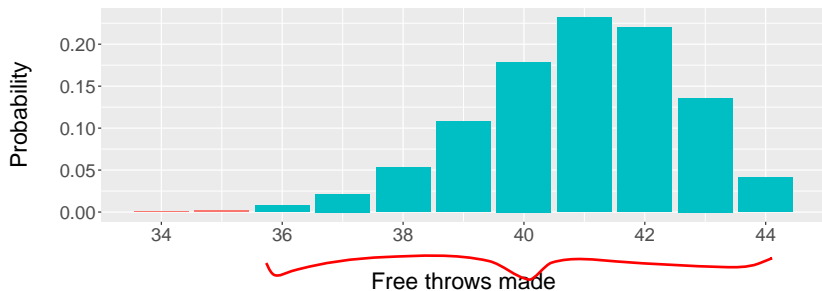
95% Prediction Interval



# Prediction Intervals

Uncertainty in *new* results (*given* a statistical model):

99% Prediction Interval



# Uncertainty

Uncertainty in new results can be described exactly only when we have an exact statistical model. But we almost never have the “true” distribution of the data (e.g., the “true” value of  $p$ ). So how can we come up with a model in the first place? More generally, how can we deal with uncertainty in the model?

# Hypothesis Testing

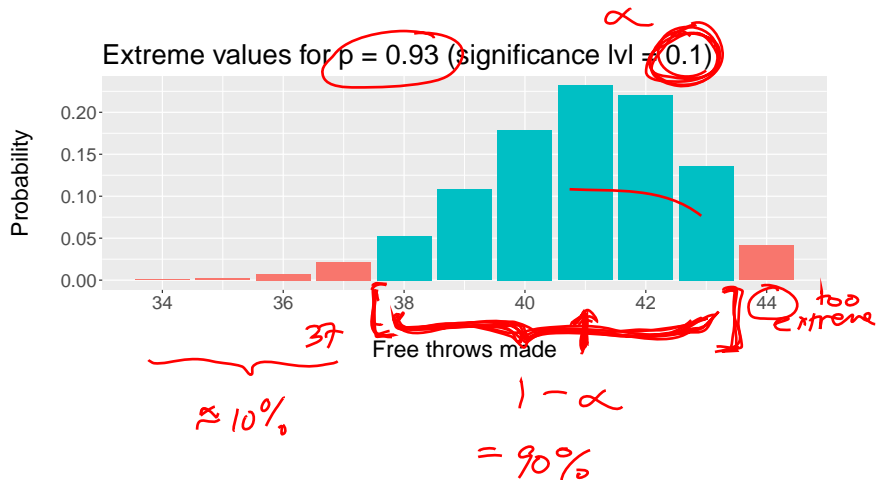
One idea: *choose* a value of  $p$  and see whether the data is “plausible” for that value of  $p$ .

# Hypothesis Testing

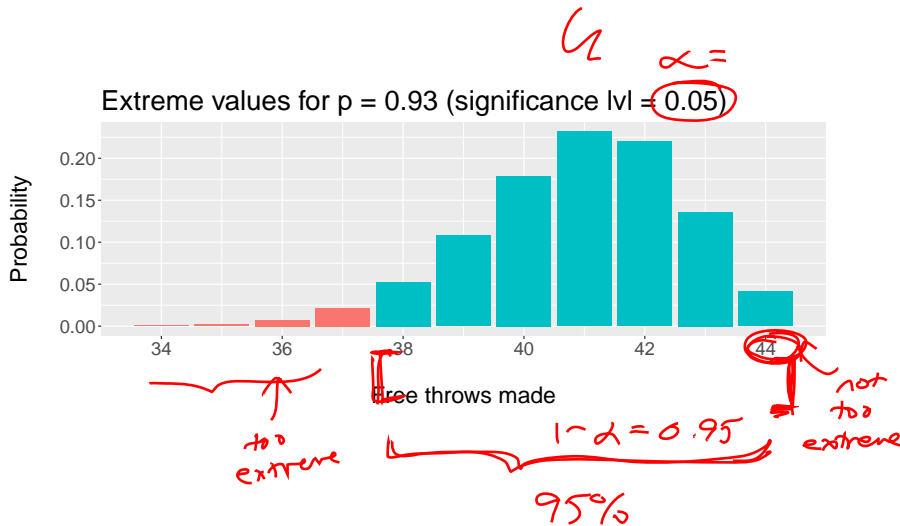
“Plausible” means: what is the probability that a value as “extreme” as the observed value occurs? If the probability is less than some **significance level**  $\alpha$  (usually, 0.05, or 0.1, or 0.01), then we deem that value of  $p$  implausible.



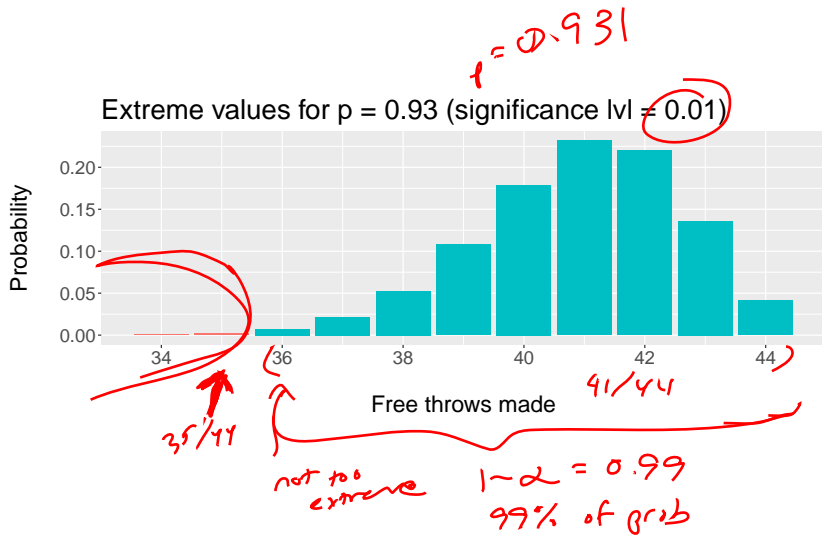
# Hypothesis Testing



# Hypothesis Testing



# Hypothesis Testing

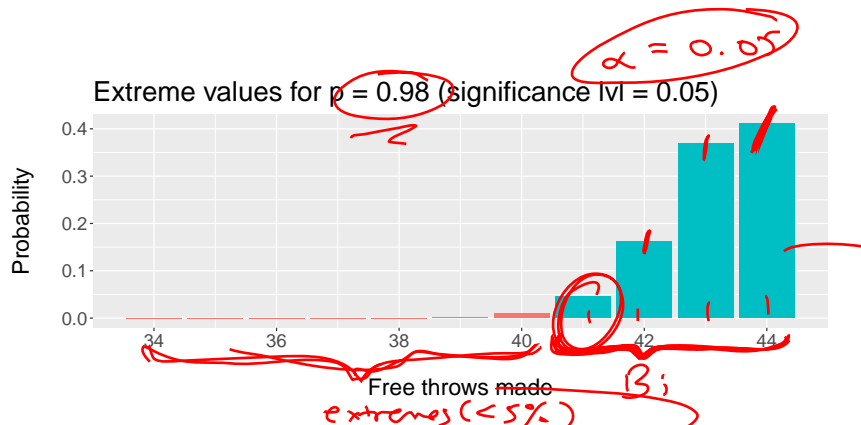


# Hypothesis Testing

Procedure and terminology for hypothesis testing:

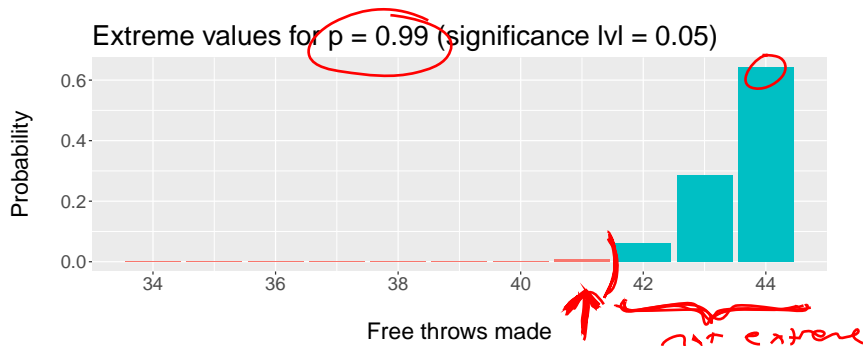
- ▶ We have a statistical model for the data (e.g., a value of  $p$ ). We *assume* that this model is the exact, “true” model. This is the **null hypothesis**.
- ▶ We choose a **significance level**  $\alpha$ . That is, how implausible will the data have to appear in order to reject the null hypothesis?
- ▶ The significance level determines the extreme values. If the data appears as one of these extremes, then we will “reject the null.” Otherwise, we will “fail to reject the null.” (Note: in hypothesis testing, we never “accept” the null! We simply say that there is not enough evidence to reject it.)

# Hypothesis Testing



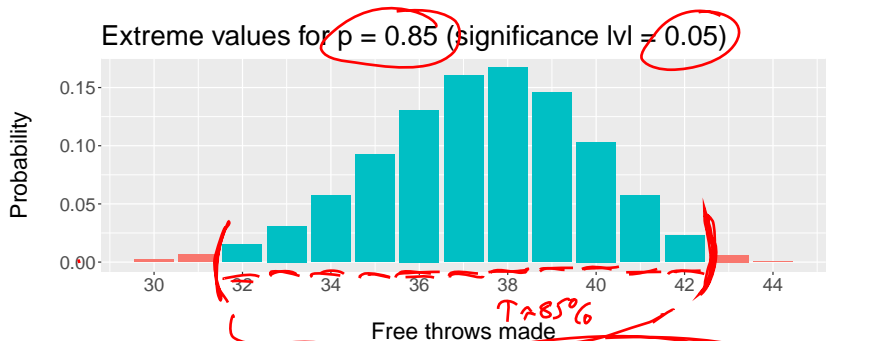
For  $p = 0.98$ , making 41 of 44 free throws is plausible (at  $\alpha = 0.05$ ). We do not reject this model.

# Hypothesis Testing



For  $p = 0.99$ , making 41 of 44 is *not* plausible (at  $\alpha = 0.05$ ). We reject this model.

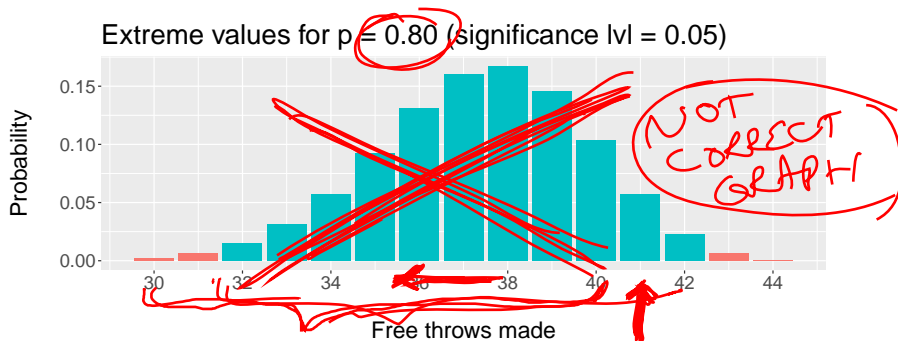
# Hypothesis Testing



For  $p = 0.85$ , making 41 of 44 free throws is plausible (at  $\alpha = 0.05$ ). We do not reject this model.

# Hypothesis Testing

95% of answers  
(30, 41)



For  $p = 0.8$ , making 41 of 44 is *not* plausible (at  $\alpha = 0.05$ ). We reject this model.



# Confidence Intervals

4

To describe uncertainty in  $p$ , we can collect all the values of  $p$  for which the data is plausible!

- ▶ 41 of 44 free throws gives a 95% confidence interval for  $p$  equal to (0.846, 0.981). (Note the terminology: use  $\alpha = 0.05$  to get a 95% confidence interval; use  $\alpha = 0.1$  to get a 90% confidence interval; etc.)
- ▶ Would a 90% confidence interval be wider or narrower?

$1 - \alpha = 10\%$  more extreme values  
→ more likely to reject model  
→ rejecting more values of  $p$   
→ fewer plausible values of  $p$   
→ narrower interval

# Confidence Intervals

Confidence intervals are weird:

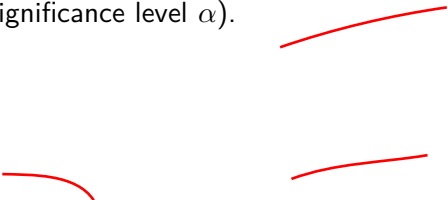
- ▶ Colloquially, you can say that we are 95% sure that  $p$  is between 0.846 and 0.981. But that statement has no mathematical meaning.
- ▶ A 95% **confidence interval** for  $p$  of (0.846, 0.981) does *not* mean that the true value of  $p$  has a 95% chance of being between 0.846 and 0.981. The true value of  $p$  is fixed. It is either in the interval or not. There is no randomness.

# Confidence Intervals

The usual definition of a confidence interval is as follows: Given that  $p$  is what it is, if you repeated the experiment of throwing 44 free throws a large number of times, and obtained a confidence interval for every experiment, then 95% of confidence intervals would contain the true value of  $p$ .

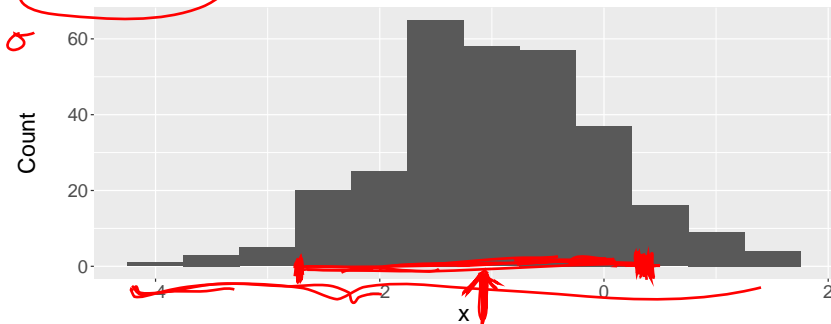
# Confidence Intervals

Confused? Me too. I prefer to think of it as the range of values of  $p$  that would not be too surprising, given the observed data (that is, the range of values for which the observed data does not qualify as “extreme,” given the significance level  $\alpha$ ).

Three red curved lines are drawn on the slide. One is positioned to the right of the text, and two are positioned below it, one to the left and one to the right.

## Example: Continuous Data

For continuous data, we might assume that the data is normally distributed, in which case we need to estimate the mean and standard deviation. Review: how can we estimate the mean here?



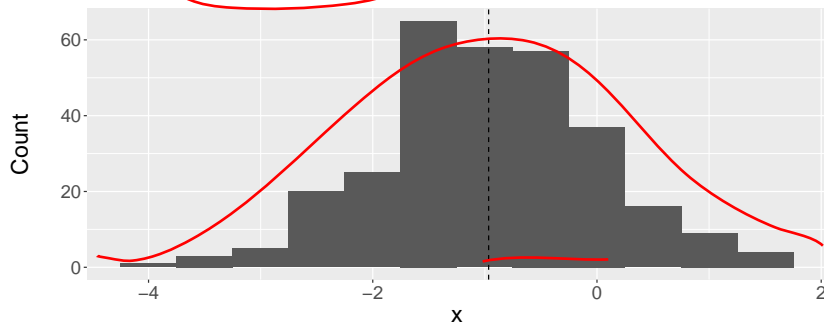
$\rightarrow x = \{-4.01, -3.89, -3.87, -3.85, \dots\}$

$\rightarrow x_B = \{-4.01, -4.01, -3.87, -3.87, -3.87, -3.85\}$

# Uncertainty

*"true" mean  
population mean*

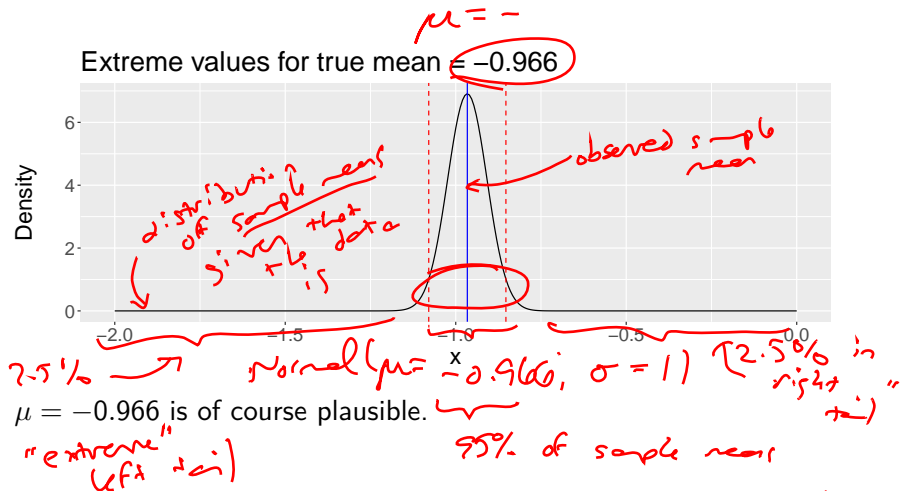
We can use the **sample mean** (here, -0.966).



## Example: Continuous Data

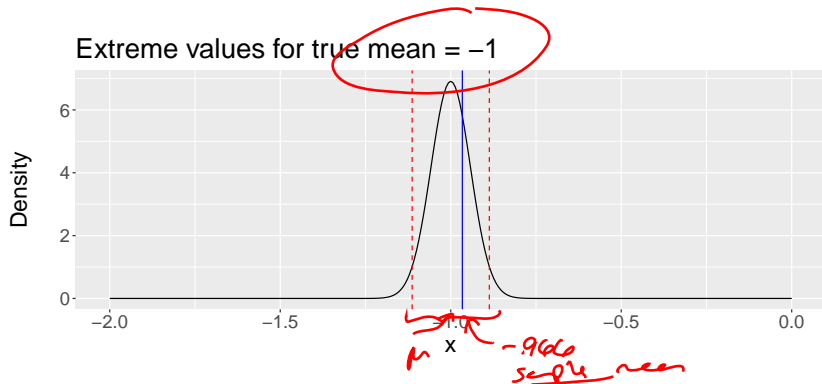
We can do hypothesis testing as before. For normally distributed data, we know the exact distribution of the **sample mean** (as it turns out, it is a  $t$  distribution). Thus we know how to ask, for various values of the “true mean”  $\mu$ , whether the observed sample mean is plausible.

## Example: Continuous Data



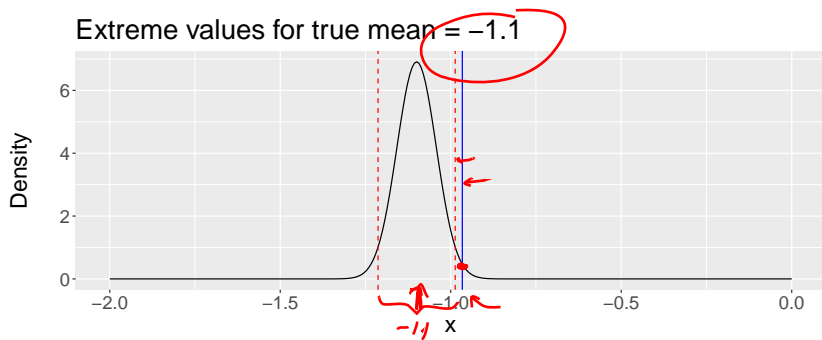


## Example: Continuous Data



With the significance level  $\alpha$  taken to be 0.05,  $\mu = -1$  is also plausible.

## Example: Continuous Data



With  $\alpha = 0.05$ ,  $\mu = -1.1$  is *not* plausible.

## Example: Continuous Data

A 95% confidence interval for  $\mu$  is  $(-1.08, -0.85)$

## Example: Continuous Data

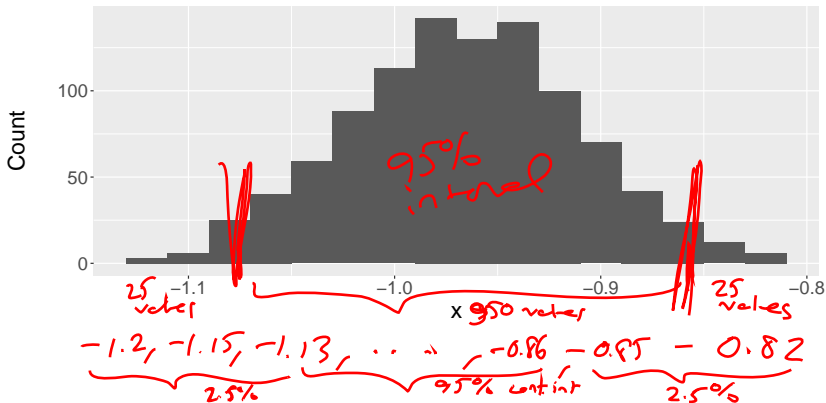
Binary data (0s and 1s; successes and failures; made and missed free throws) are easy to model; as long as events are independent, the binomial distribution is the only possible distribution and we only have to estimate the value of  $p$ . Continuous data is much more difficult to model, since we don't only need to make assumptions about the *parameters* of the distribution: assuming that the data is normal (or  $t$ , or whatever) is a pretty big assumption in itself!

# Bootstrap

One clever way to estimate uncertainty *without making any explicit assumptions about the true distribution* is called the **bootstrap**. The idea is to use the distribution of the data (also called the **empirical distribution**) as the “true” distribution and sample from that.

# Uncertainty

In the example above (which is in fact drawn from a true normal distribution), I have 300 observations. I can “bootstrap” sample means by taking 300 draws from the observed data (with replacement!) and calculating the sample mean for that new sample. I then repeat as many times as I like:



# Uncertainty

To get an interval estimate, I then find out where 95% (or 90%, or 99%) of the bootstrapped sample means lie:

```
##           2.5%           97.5%  
## -1.0739360 -0.8588168
```

↙ "exact" CI

$(-1.08, -0.85)$

The bootstrapped 95% confidence interval  $(-1.07, -0.86)$  is very similar to the confidence interval we got from hypothesis testing!

# Uncertainty

Of course, the bootstrap really shines when the data cannot be assumed to be a standard, known type of distribution like the normal distribution. We'll explore an example of this kind of bootstrap in the next demo. . . .