Eltecon Data Science Course by Emarsys

Measuring uncertainty

András Bérczi

October 14, 2020

Homeworks from last week

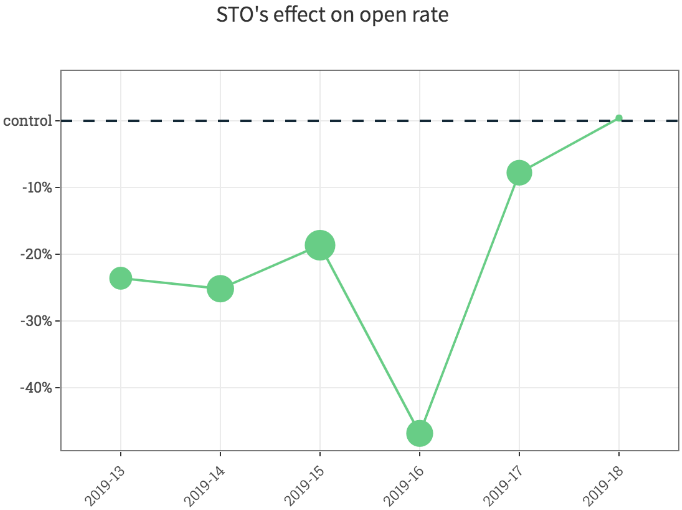
Any questions about final project?

Measuring uncertainty

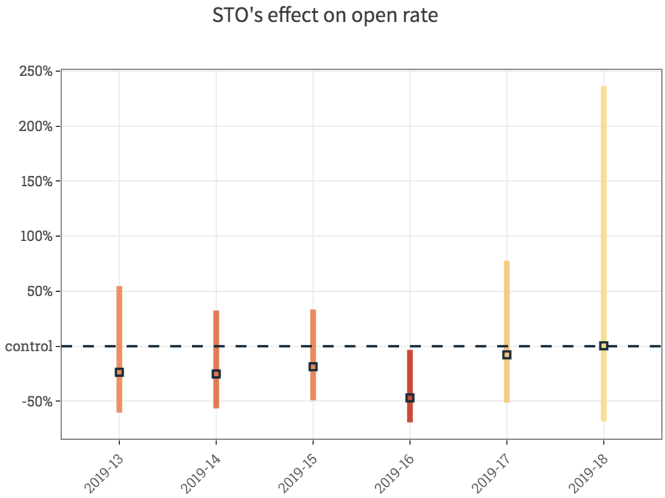
We can always measure something from our data. . .

. . . but how sure can we be about our measurement?

We can always measure something from our data. . .



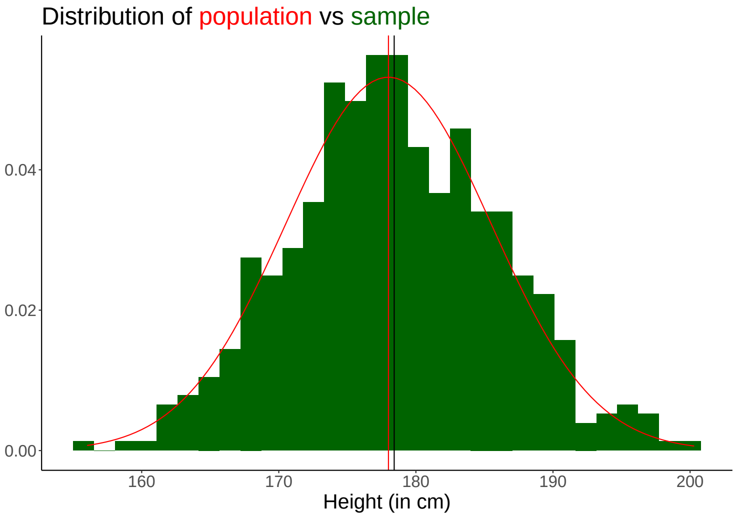
But not necessarily significant!



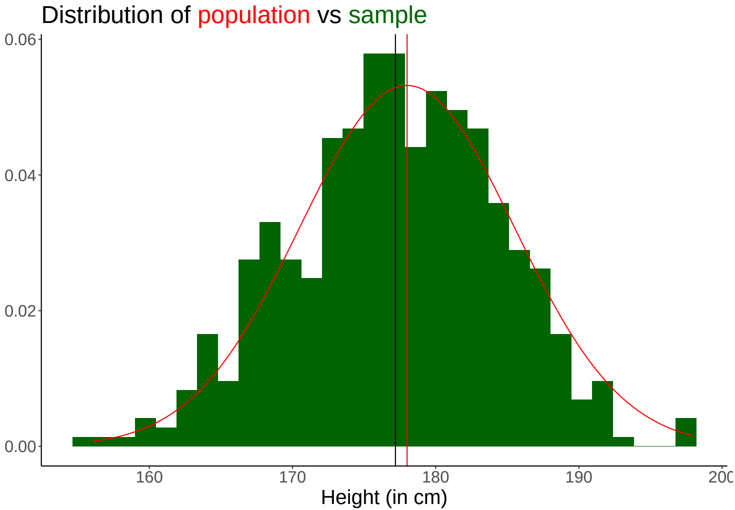
Why do have uncertainty in the measurement?

- If you knew the whole population, there wouldn't be uncertainty in your measurement
- But we only see 1 'segment' of the data = we have a sample of the population

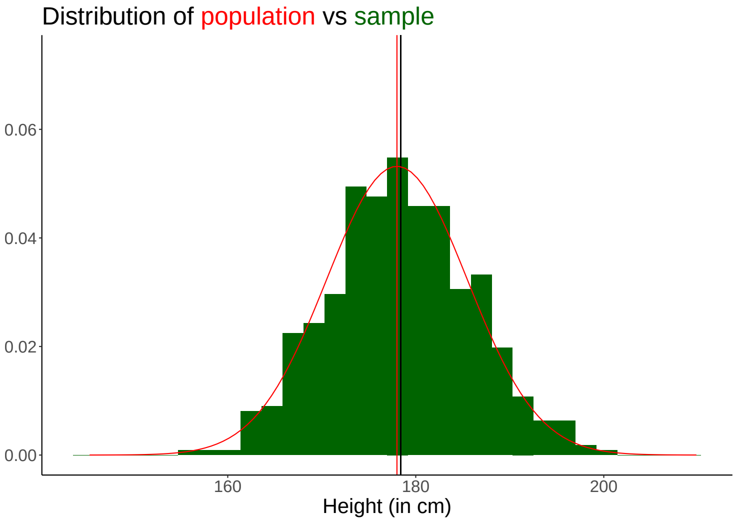
Sampling from a population



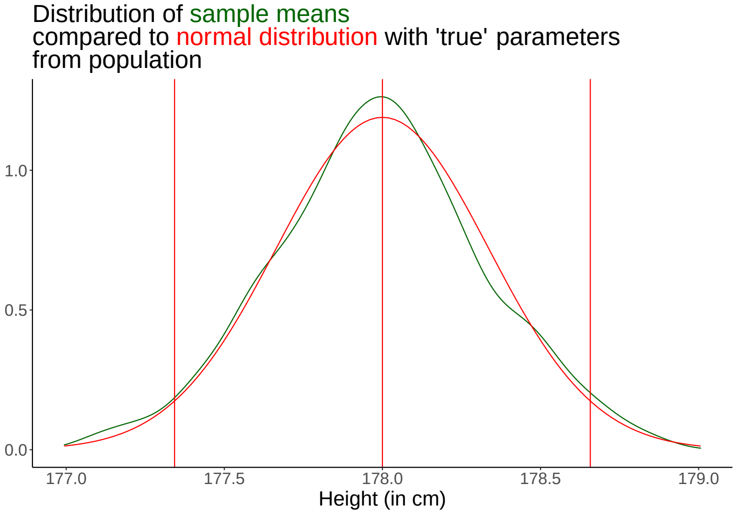
Sampling from a population



Sampling from a population



Distribution of sample means - LLN + CLT



Law of Large Numbers

The average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed. - Wikipedia

Central Limit Theorem

When independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a bell curve) even if the original variables themselves are not normally distributed. - Wikipedia

What are Confidence Intervals?

- The normal table gives us the fact that P[-1.96 $<$ Z $<$ 1.96] $=$ 0.95.
- With a sample of n values from a population with mean μ and standard deviation σ, the Central Limit theorem gives us the result that Z $= \sqrt{n}\frac{\bar{x}-\mu}{\sigma}$ is approximately normally distributed with mean 0 and with standard deviation 1.

What are Confidence Intervals?

Start from P[-1.96 < Z < 1.96] = 0.95 and then substitute for Z the expression $\sqrt{n}\,\dfrac{\overline{X}-\mu}{\sigma}$.

This will give us

$$P\left[-1.96 \;<\; \sqrt{n}\,\frac{\overline{X}-\mu}{\sigma} \;<\; 1.96\right] \;=\; 0.95$$

We can rewrite this as

$$P\left[-1.96\frac{\sigma}{\sqrt{n}} \;<\; \overline{X}-\mu \;<\; 1.96\frac{\sigma}{\sqrt{n}}\right] \;=\; 0.95$$

Now subtract $\overline{X}$ from all items to get

$$P\left[-\overline{X}-1.96\frac{\sigma}{\sqrt{n}} \;<\; -\mu \;<\; -\overline{X}+1.96\frac{\sigma}{\sqrt{n}}\right] \;=\; 0.95$$

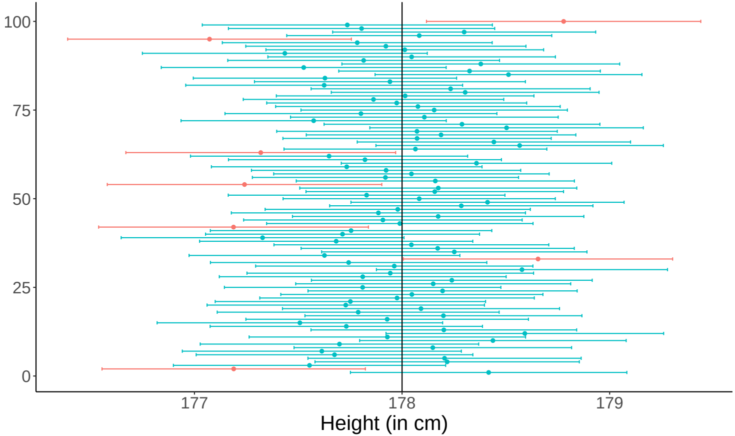
Multiply by -1 (which requires reversing inequality direction) to obtain

$$P\left[\overline{X}+1.96\frac{\sigma}{\sqrt{n}} \;>\; \mu \;>\; \overline{X}-1.96\frac{\sigma}{\sqrt{n}}\right] \;=\; 0.95$$

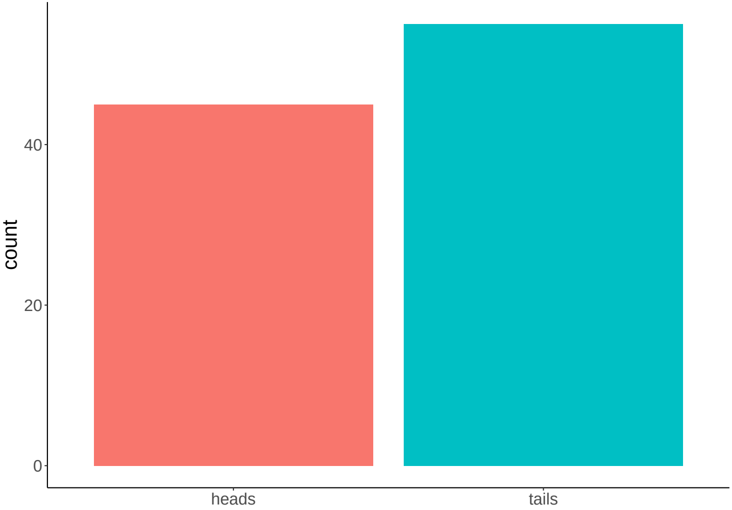
http://people.stern.nyu.edu/gsimon/Pamphlets/

What are Confidence Intervals?

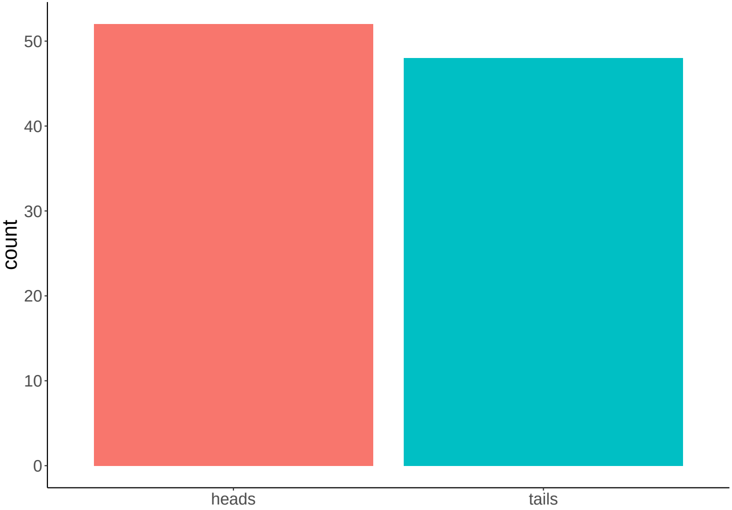
Mean and CI from different samples:
About 95% of the CIs contains the true mean,
but 5% does not contain (just by chance)



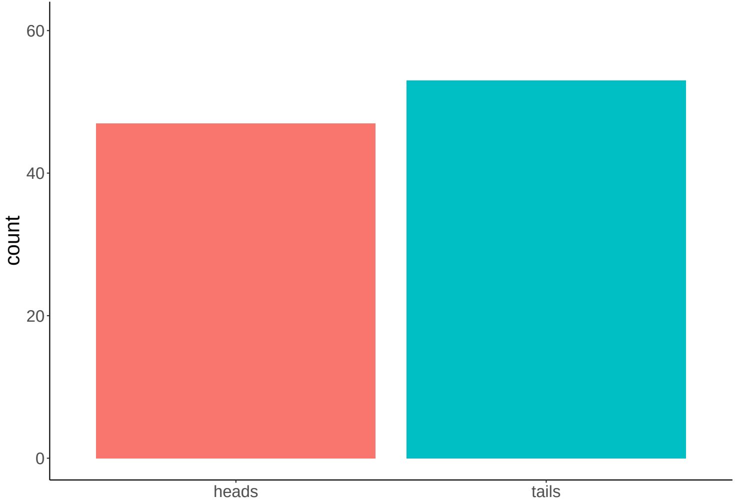
Distribution of sample means with different distributions



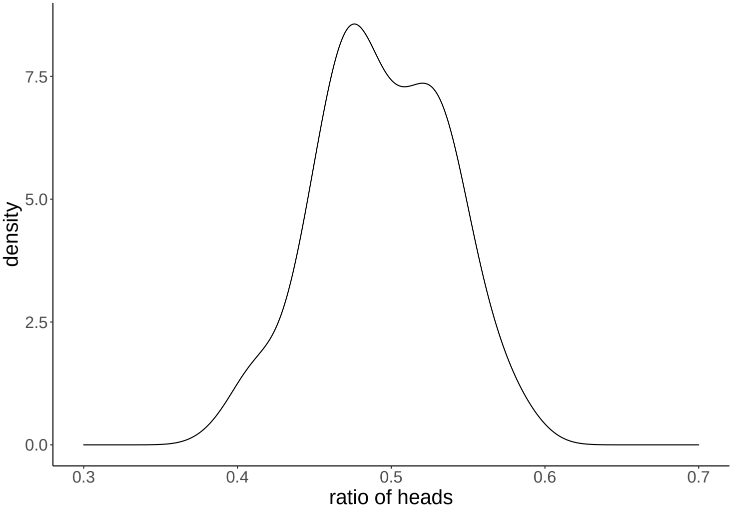
Distribution of sample means with different distributions



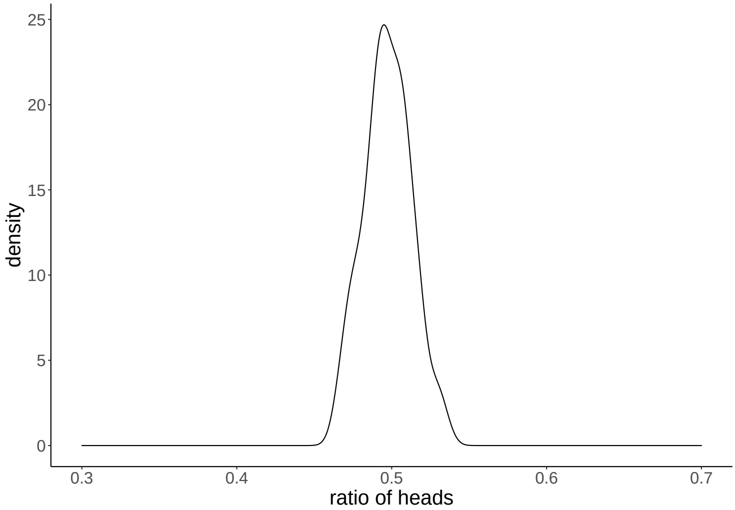
Distribution of sample means with different distributions



Distribution of sample means with different distributions



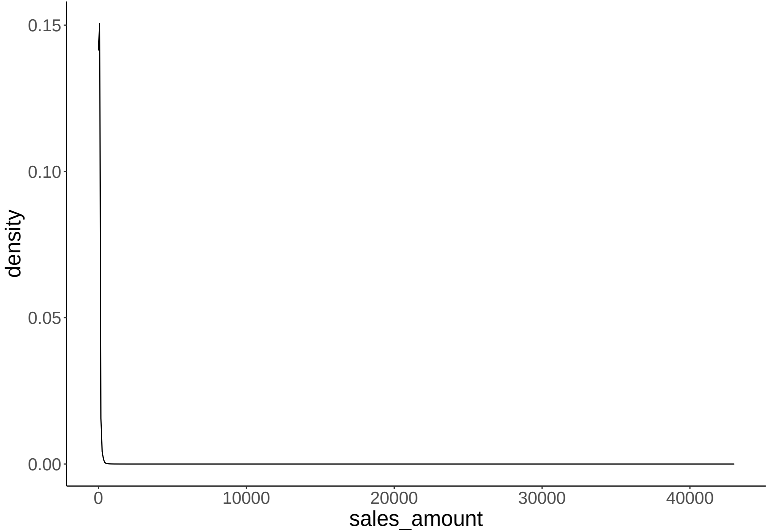
Why does sample size matter?



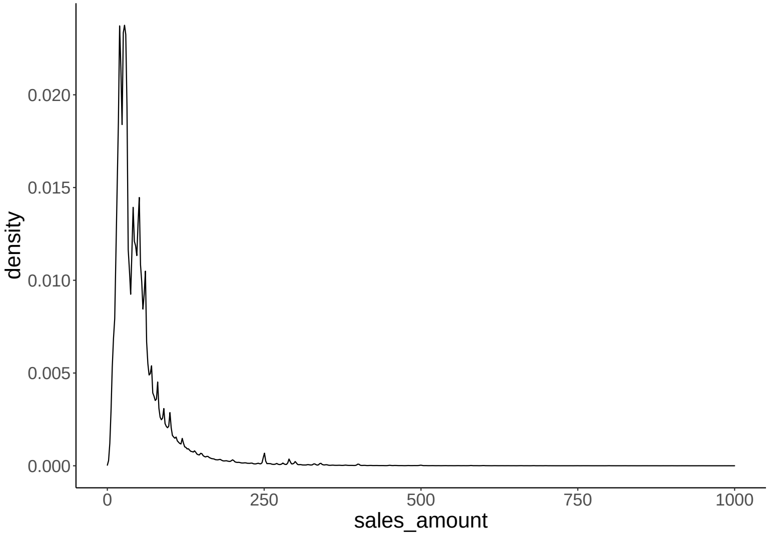
What are the key assumptions?

- i.i.d. sampling
- finite variance / distribution is not ‘long tailed’

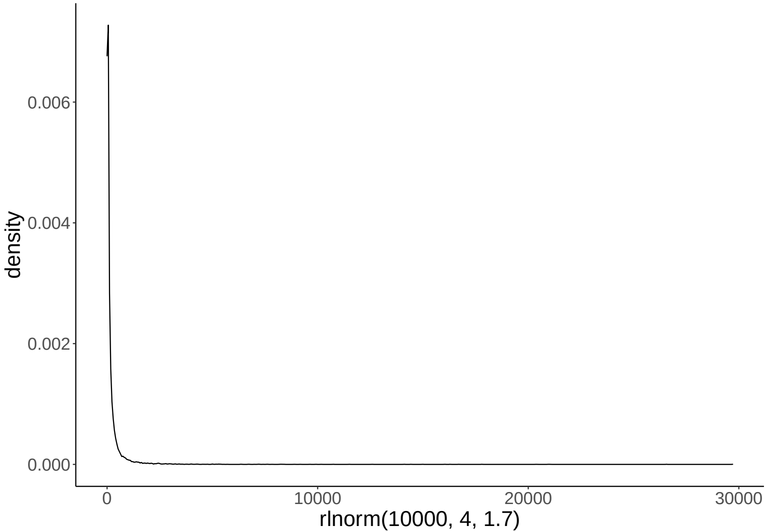
What if variance is infinite?



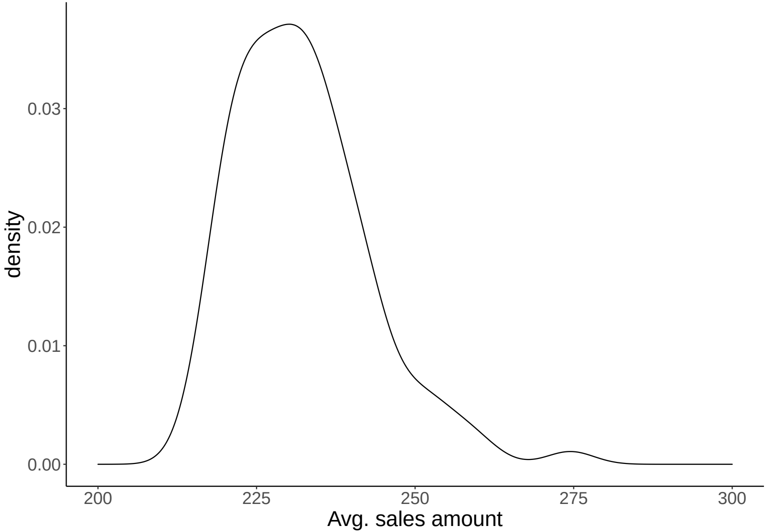
It stayes very skewed even if we zoom in



What if variance is infinite?



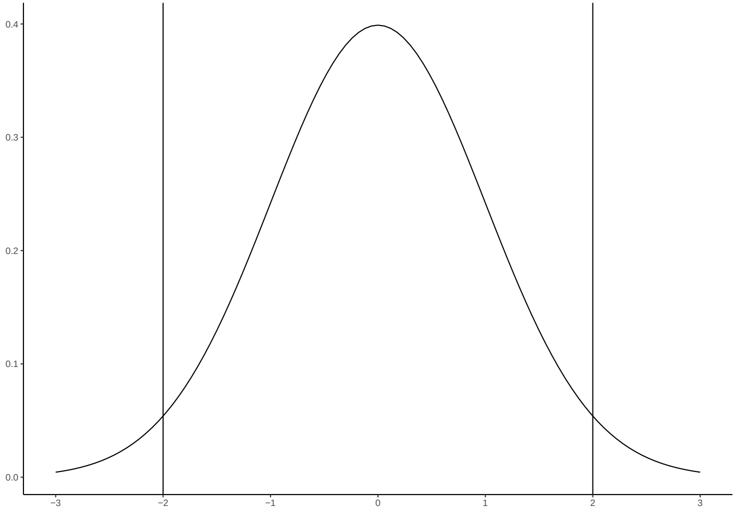
Distribution of avg. sales amount from samples



How can we calculate the uncertainty of our measurement?

- Based on variance of known distribution
- Monte-Carlo method
- Bootstrapping
- (and other methods as well of course)

Calculate uncertainty based on variance



How to calculate uncertainty from sampling distribution

By CLT + LLN, you can add uncertainty to your point estimate, such as:

$$\bar{x} \pm 1.96 * \frac{s}{\sqrt{n}}$$

$\bar{x}$ is the sample mean,

s is the standard deviation of the sample distribution,

n is the sample size

Bayesian uncertainty

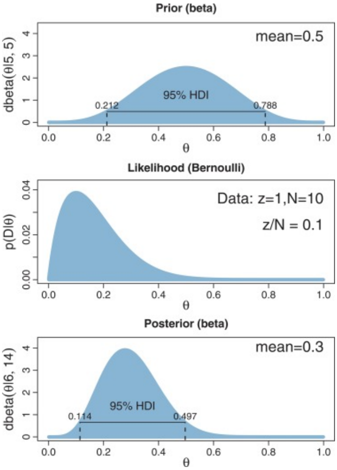
Confidence Interval: If we would resample from our population, 95% of times the Confidence Interval will contain the true, unknown parameter.

Credible Interval: There is a 95% chance that this interval contains the true parameter.

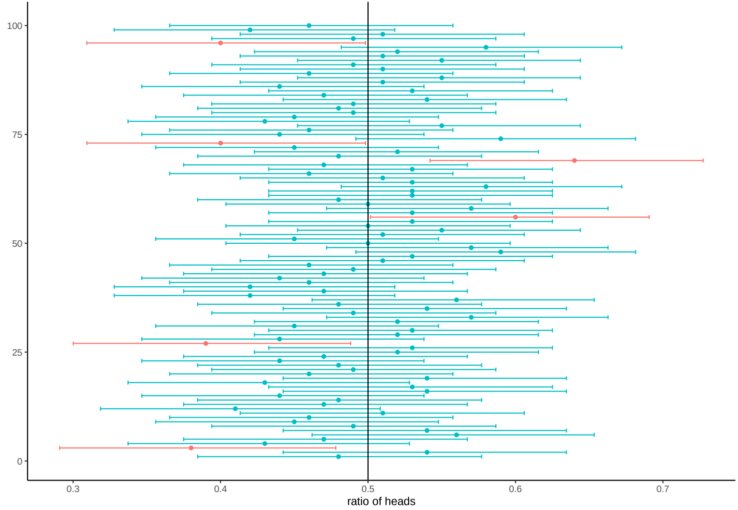
Difference in calculating uncertainty

Confidence Interval: Based on sampling distribution of means

Credible Interval: Based on data and prior belief



95% Credible interval



Confidence Interval vs Credible Interval

- Credible Interval is easier to understand
- Credible Interval gives smaller interval if we have some prior knowledge
- With wrong prior provided, our posterior distribution is going to be wrong as well!
- With a lot of data or with non-informative priors, the two intervals are about the same

head_ratio	cred_int_lower	cred_int_higher
0.48	0.4320385	0.5480755

head_ratio	conf_int_lower	conf_int_higher
0.48	0.3820784	0.5779216

How to plot uncertainty

Calculate uncertainty over time

Takeaways

- Always show uncertainty
- Think about your audience