Now What? Sampling Distributions

November 2025

Forward the Foundation

We have discussed probability in different forms

- Uniform distribution (eg a single die roll)
- Hypergeomteric distribution (eg 2 aces in a hand of cards)
- Binomial distributions (eg X heads in ten coin flips?)
- Poisson distribution (ie raw counts that look normal-ish)
- Normal distribution (ie bell shaped curve)

Forward the Foundation

Under certain assumptions, the last three distributions can be approximated with a normal distribution (= N(mean, variance))

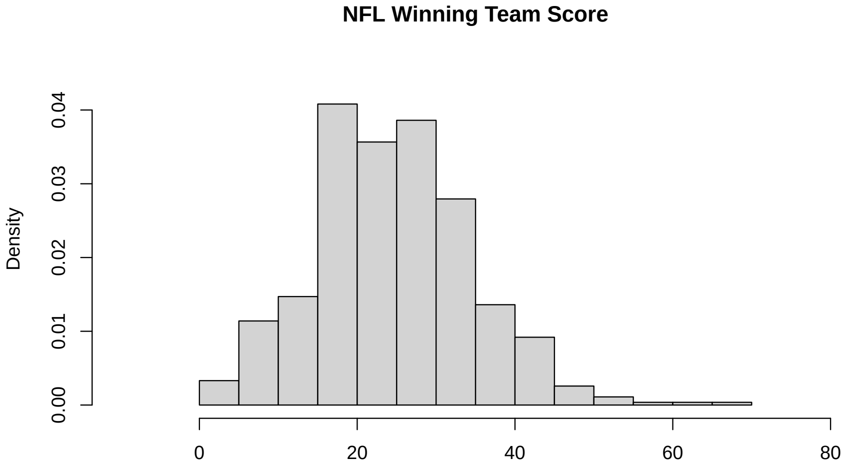
$$\begin{aligned}
\text{Normal Distribution} &\sim N(\mu,\ \sigma^2) \\
\text{Binomial Distribution} &\sim N(np,\ \frac{p(1-p)}{n}) \\
\text{Poisson Distribution} &\sim N(\mu,\ \mu) \\
\text{Poisson Distribution (with overdispersion)} &\sim N(\mu,\ \sigma^2)
\end{aligned}$$

Leveraging this, we can talk about...

- where things are most probable
- how weird is the result we saw

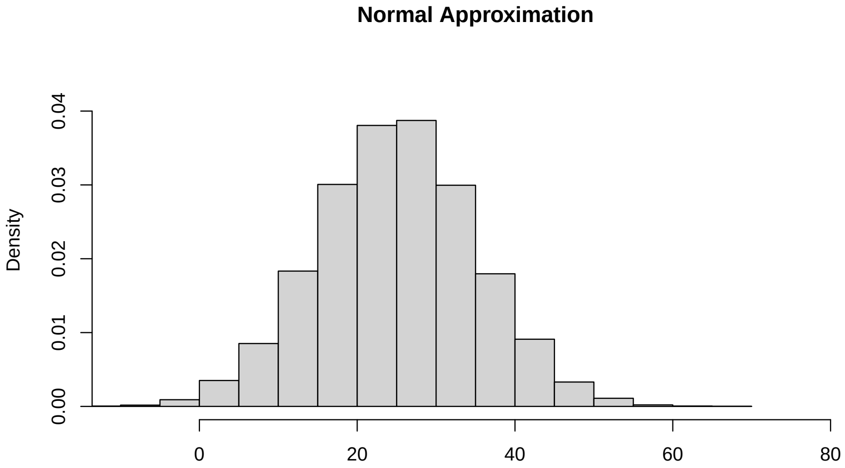
NFL Winning Scores Revisited

The real scores



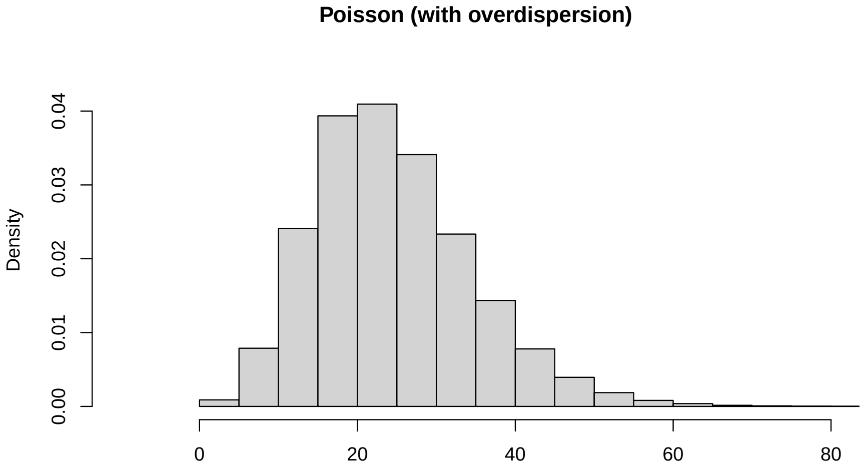
NFL Winning Scores Revisited

Normal Approximation for the real scores



NFL Winning Scores Revisited

Poisson Approximation for the real scores



Forward to Where?

It turns out there is another critical aspect to the normal distribution besides the fact people mislabel other distributions with it. But first we need to make a pit stop...

If something is random and can be repeated ad nauseam then we can build a histogram of the outcomes. That long term behavior builds a distribution (with probabilities involved).

Forward to Where?

It turns out there is another critical aspect to the normal distribution besides the fact people mislabel other distributions with it but first we need to make a pit stop...

If something is random and can be repeated ad nauseam then we can build a histogram of the outcomes. That long term behavior builds a distribution which can be useful (think probability)

So what does that imply for the mean of a random sample?

A sample's mean has...a distribution?

Yes! We can take a sample (of say 10 craft beers from around the state of Iowa) and record the mean of some info on them (eg find the average alcohol content).

And then do it again.

And then do it again.

And then do it again.

And then do it again.

And then do it again.

And then do it again.

And then do it again.

And then do it again.

And then do it again.

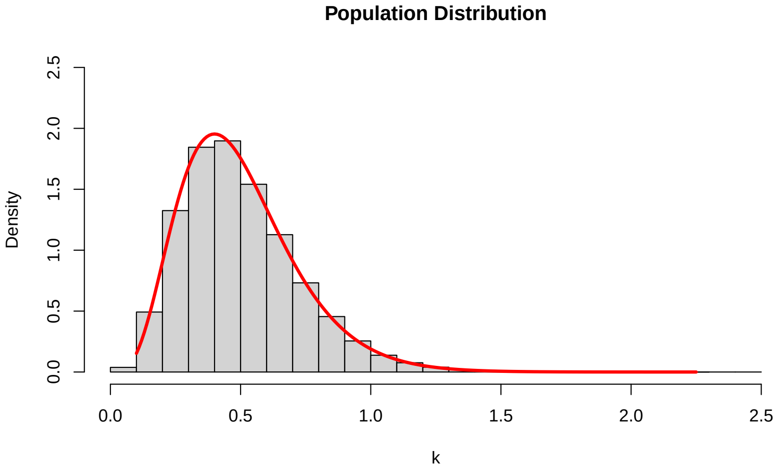
Sampling Distribution

The probability of an event is the long run frequency of that event occurring.

- So we can take an infinite number samples, each with 10 beers, (...please Grinnell?) and we would know the long term behavior for the sample mean of the alcohol content of 10 randomly selected beers.
- Called a **sampling distribution**
  - It's the long term behavior (distribution) of a statistic if we were to rerun the experiment/study forever and keep calculating it.

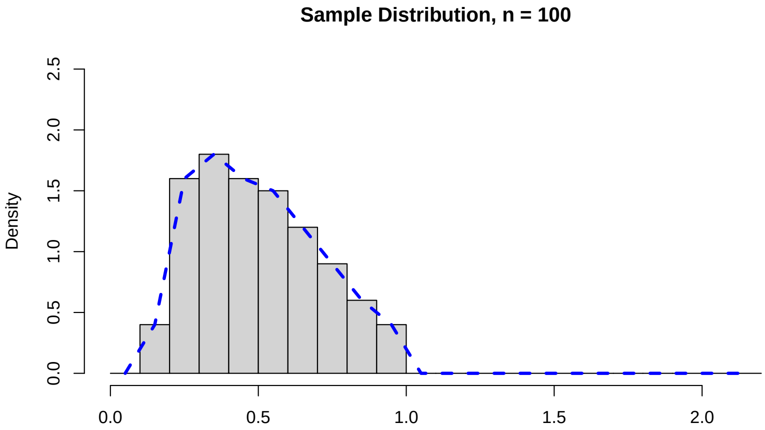
Ex: Population Distribution

Population distributions are usually impossible to know so here is a a simulated distribution with a Gamma distribution



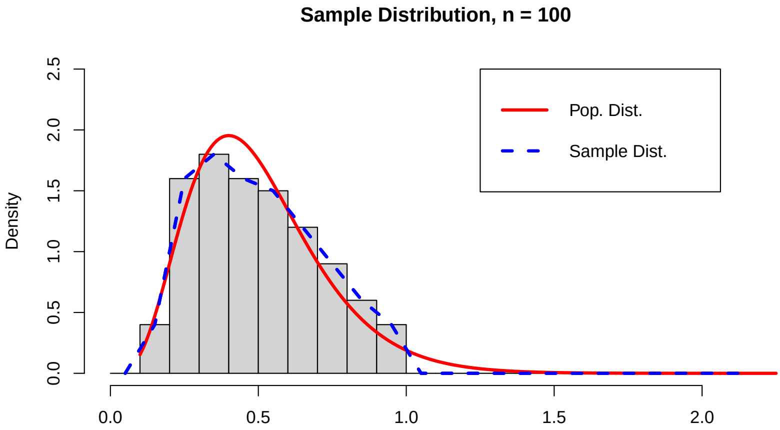
Ex: Sample Distribution

And we can take a sample from our population and plot the results



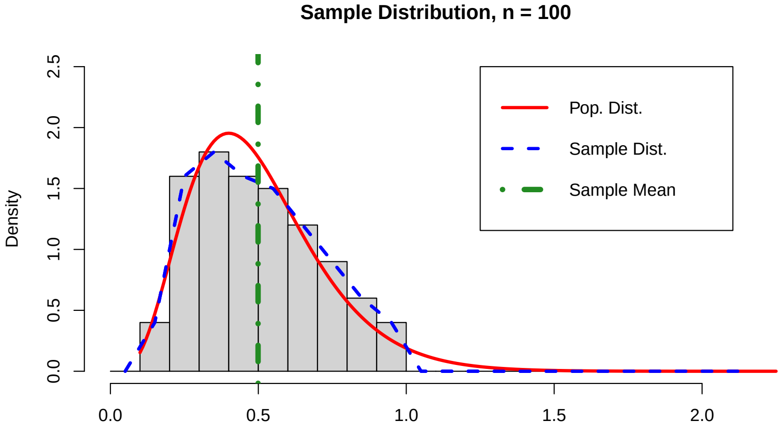
Ex: Sample Distribution

Compared to the population distribution we aren't that bad



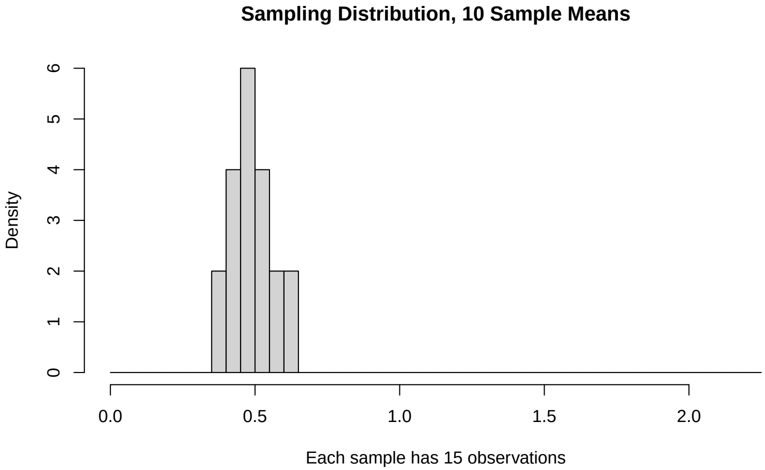
Ex: Sample Distribution

We can also mark down where the mean of our sample is



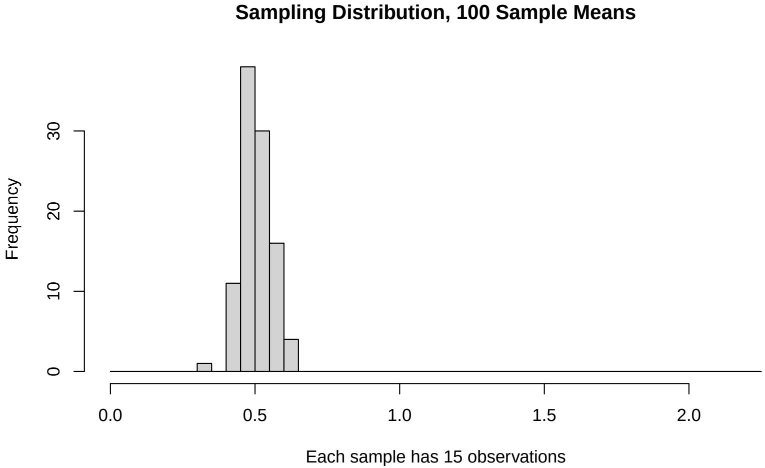
Ex: Sampling Distribution

We rerun an experiment, redo the survey, collect a new set of volunteers, etc... and then plot the recorded *statistics*



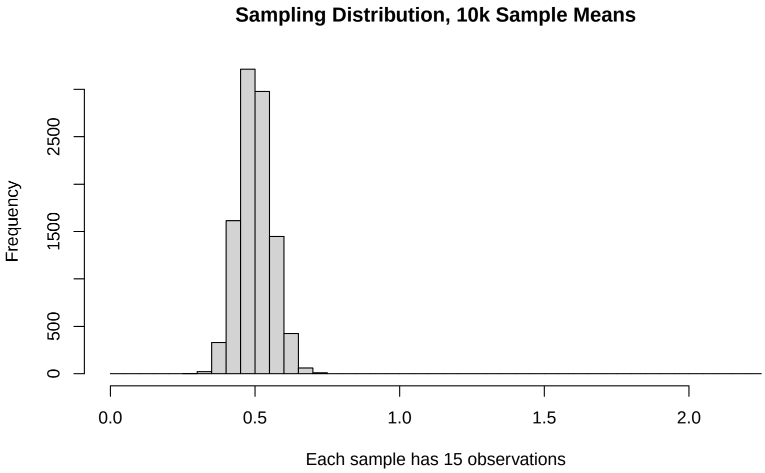
Ex: Sampling Distribution

And we can bump up the number of samples we are taking (eg run even more experiments)



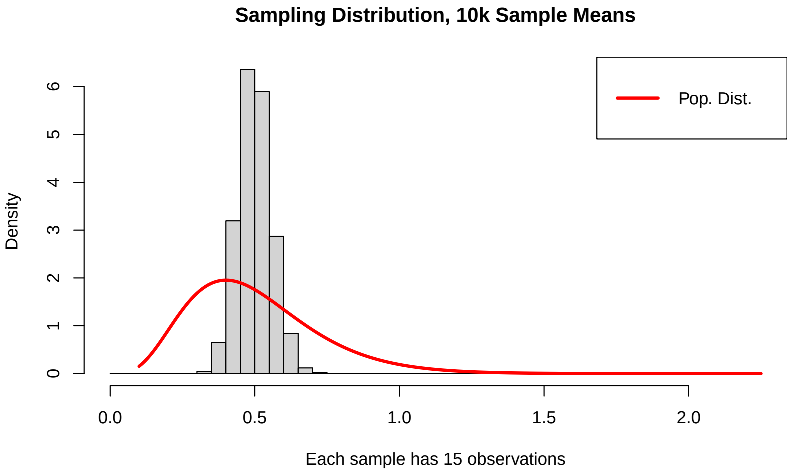
Ex: Sampling Distribution

And again



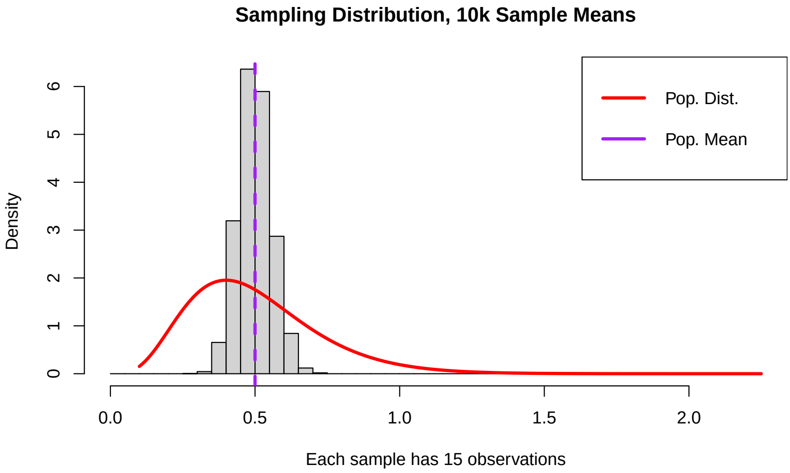
Ex: Sampling Distribution

And this looks nothing like our population distribution



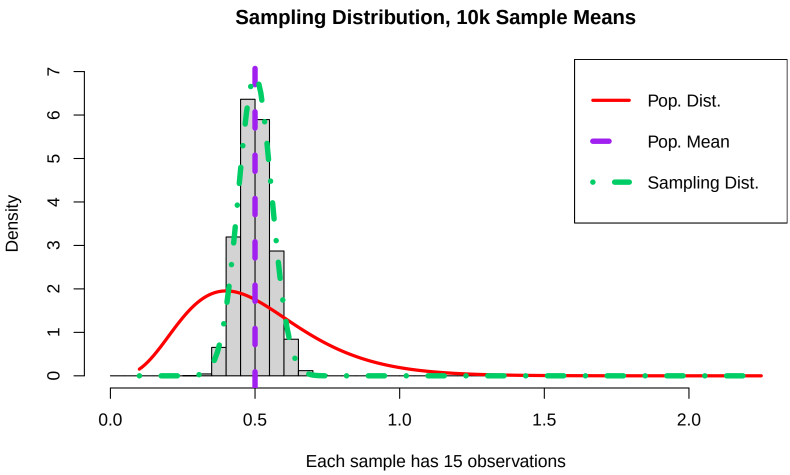
Ex: Sampling Dist Center

Buuuuuut it does line up really well with our population's mean



Ex: Sampling Dist Shape

And looks solidly normal



Sampling Distribution: Interpretation Do's and Don'ts

This distribution references ONLY the mean alcohol content of 10 randomly selected craft beers (the average of observations)

- Where do we think the middle 95% of our samples' means will be? (ie confidence interval)
- What is the probability that I'd see a mean this large given my true mean is 0? (ie hypothesis test)

This distribution tells us almost nothing about what you get if you randomly ordered a craft beer at the bar (ie an observation)

- Sampling distribution is the long term behavior of a statistic, not of the observations (eg middle 95% of the raw data can't be discussed)
- All things are stated in terms of that statistic

Three Distributions

Population distribution:

Sample Distribution:

Sampling Distribution:

Three Distributions

Population distribution: The “true” distribution from where the data comes from. It’s the distribution with all possible observations.

- The alcohol content of all craft beers in the state

Sample Distribution: The distribution of the observations we have

- Hopefully looks a lot like the population distribution
- The alcohol content of 10 craft beers from around the state

Sampling Distribution: The long term behavior of a sample statistic and can only be used to discuss that statistic (eg mean, proportion, variance, etc...)

- Under certain circumstances, it looks normally distributed
- The distribution for the mean alcohol content of an infinite number of a sample of 10 beers

Infinite possible combinations of 10 beers?

That sounds a whole lot like the permutation test we saw.

We exhaustively listed all possible combination of jurors and used that to say the probability we'd have 1 or 0 black jurors

So what is the sampling distribution doing?

The sampling distribution is using assumptions to approximate all possible outcomes while last week we wrote out all possible outcomes (call it the permutation distribution). That is, the sampling distribution is approximating what we did with permutations

Which distributions do we *know* we know?

Population distribution:

Sample Distribution:

Sampling Distribution:

Which distributions do we *know* we know?

Population distribution: Unknown

- is the “truth”
- usually must estimate

Sample Distribution: We know this one!

- This is the distribution of our actual data
- Ideally looks like the population distribution

Sampling Distribution: Unknown

- Estimated using statistics from our sample
- and with some assumptions
  - which must be checked!
- we believe it’ll be normally distributed

What do we calculate from these distributions?

Population distribution: Parameter

- A **parameter** is some numeric summary of our population
- Usually impossible to calculate

Sample Distribution: Statistic

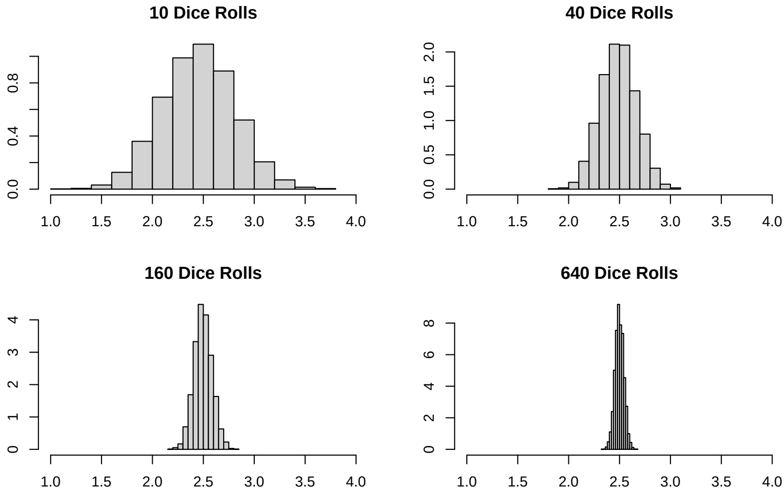
- A **statistic** is a summary of our sample
- We can calculate this

Sampling Distribution: Probabilities for a Sample's Statistic(s)

- Can measure the probability we see a result this weird
  - Based on assumptions
  - Hypothesis test
- Can state where results are most likely
  - Confidence intervals for the sampled statistic

What do we notice from the dice lab?

Based on the 10k dice simulations....d4

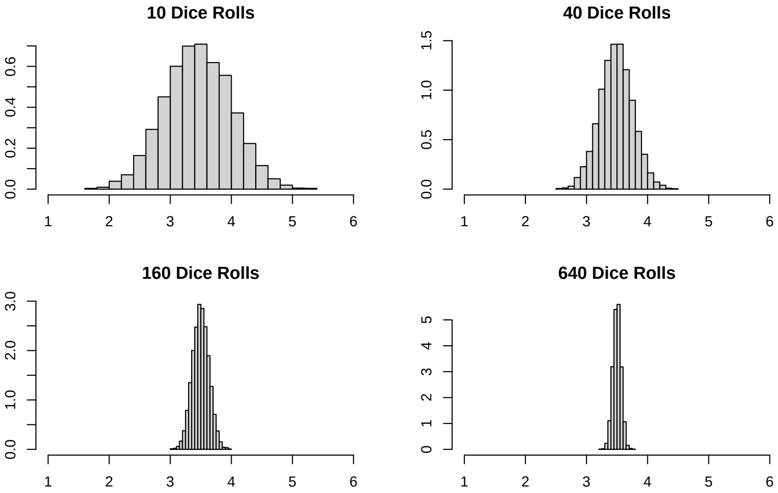

10 Dice Rolls

40 Dice Rolls

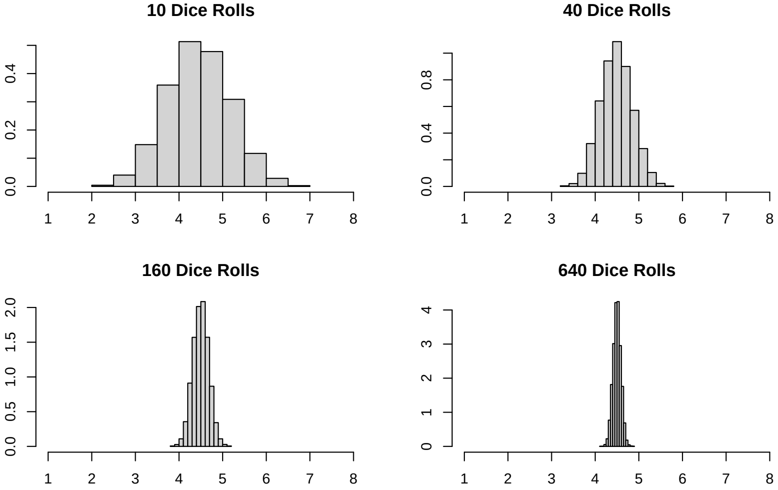
160 Dice Rolls

640 Dice Rolls

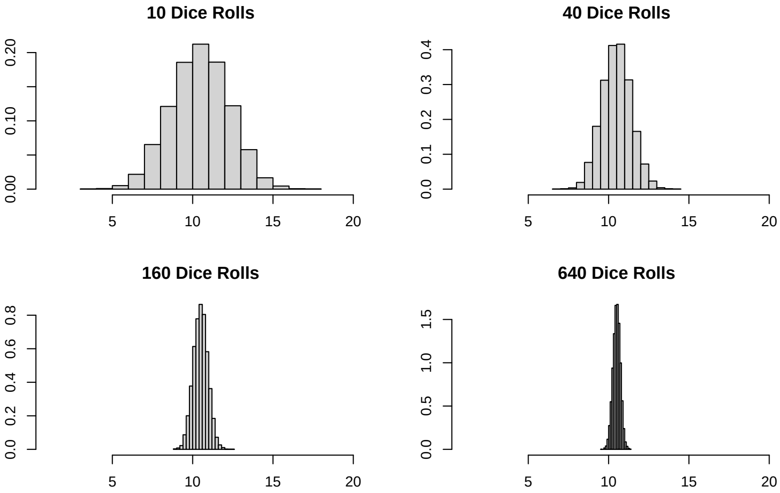
Based on the 10k dice simulations....d6



Based on the 10k dice simulations....d8



Based on the 10k dice simulations....d20



What did we notice?

What did we notice?

- More dice in a sample will tighten the sampling distribution
  - Ie the variance of a sampling distribution is a function of the number of observations in a sample (called sample size)
- It's bell-shaped, with no apparent outliers
  - Normal, under certain assumptions
- Centered at our expected value
  - Called **unbaised**. That is the long term average (of many samples) of our estimate is the parameter
  - Ie our estimate is not always, systematically wrong
- More sides lead to more spread out distributions
  - More variability in the pop. -> more variability in the sample -> more variability in the sampling distribution

What statistics have a sampling distribution?

What statistics have a sampling distribution?

All, to the best of my knowledge

Summary	Pop. Parameter (Unknown)	Sample Statistic (Known)
Mean	μ	$\bar{x}$
St. Dev.	σ	$s\ (= \hat{\sigma})$
Proportion	p	$\hat{p}$
Correlation	ρ	r
Regression Coeff.	β	$\hat{\beta}$

- Note $\bar{}$ (read as "bar") is used for a mean
- And $\hat{}$ (read as "hat") is for estimated values (generally)
- Mean and Reg. Coeff. are this unit; Proportion is next

Next Time

We will use the sampling distribution of a mean to...

- Z-tests (for known variance)
- t-tests (for unknown variance)
- confidence intervals for both

We will also dive into what assumptions are necessary to know the final shape of the distribution.