Chapter 4.2 - Sampling Distributions

> "People think that if you collect enormous amounts of data you are bound to get the right answer. You are not bound to get the right answer unless you are enormously smart." - Bradley Efron

Scientists are more obsessed with data than statisticians— this sounds insane but I promise it's true.

No modern statistician worth their salt is sitting in front of a scientist saying that the problem with their study is that their sample size is too small. They might say the data is information deficient or general crap, but never that n is too small. The opposite direction is equally ridiculous since more data can't possibly harm your study.

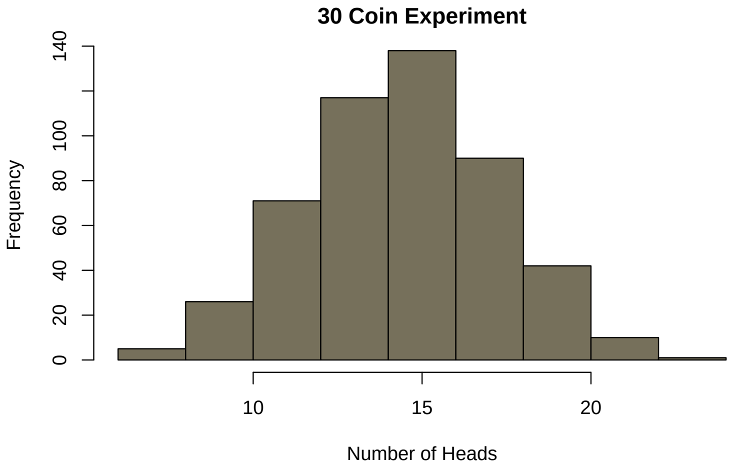
But scientists are frequently asking statisticians to tell them how much data they *need.* The answer to this question isn't helpful either:

Gather as much data as you can afford to.

This is an introductory textbook in a quantitative science which means we'll be covering methods derived by dead people and avoiding discussion around the tools living experts consider "useful". We're going to take a trip backwards in time to when statisticians cared about answering scientists queries about sample size (despite the fact that it is, in fact, a pointless question today). But we're not studying history for history's sake; we're studying history to see how almost the entirety of statistics is built on the foundation of one theorem. A theorem so powerful that it's responsible for an *incomprehensibly large* number of scientific discoveries.

We rolled some dice, we added up their value, we did that over and over again, we ended up at the normal distribution. We sampled deer body weights and averaged them up, we ended up at the normal distribution. . . why in the world did this work out that way? Is this just for dice and deer? You know that the answer is no.

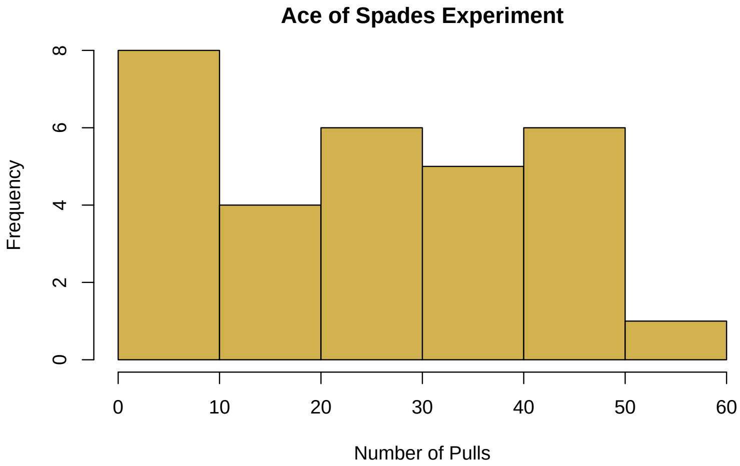
We flip a coin 30 times, we add up the number of heads, we repeat this 500 times, what happens?



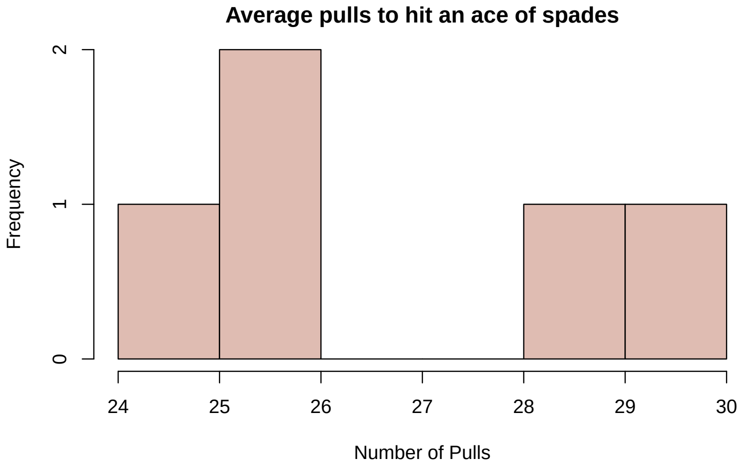
We pull cards from a deck until we hit an ace of spades, record the number, do that 30 times and average the number of pulls.

```
set.seed(73)
cat("Pull count to hit the ace of spades:", "\n", sample(1:52,30))
```

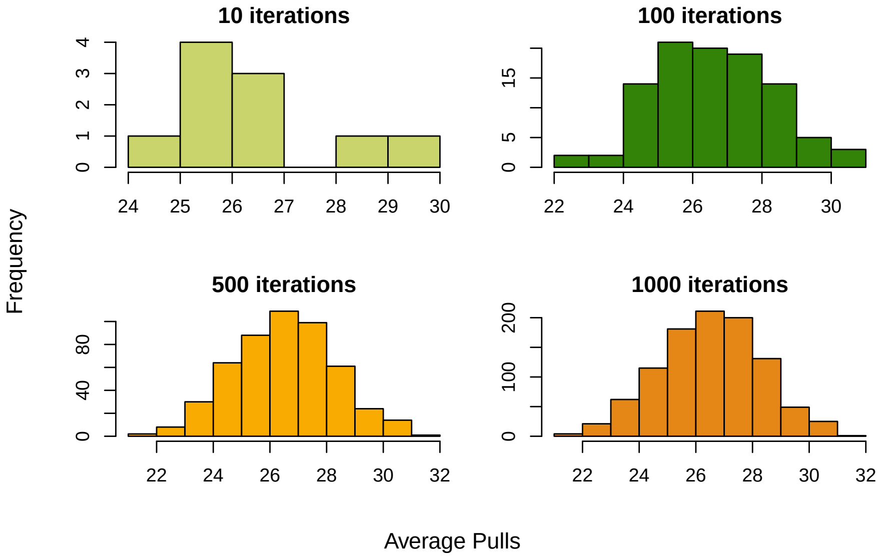
```
## Pull count to hit the ace of spades:
##  7 24 49 17 47 52 19 20 31 44 27 50 4 8 35 5 1 36 21 41 34 39 26 23 45 10 9 29 3 12
```



It's a discrete count from a 1/52 chance. The histogram of the individual pulls is very clearly **not** normally distributed. But statisticians can confidently say that the sample mean **is** normally distributed. Yet, it doesn't look normally distributed at 5 repeat experiments even though that's *a lot* of time spent pulling cards from a deck:



But as soon as we know, the shape changes the more data we generate:



What's happening here is a showcase of the **Central Limit Theorem**, the one theorem to rule them all. Before we can properly define this powerhouse of statistical science we need to gather some background on how this sort of phenomenon even arises.

Distributions of Statistics

The height of any given student in a lecture hall is a random variable, that should be an agreeable statement by now. If you close your eyes and point to a random person in the room you're performing an experiment and observing a realized value for a random variable. Fair enough.

If we select two people and add up their heights we're just observing the realization of a different random variable. While their heights are considered individual realizations of our "height" random variable, the summation of their two heights is a different "two heights added together" random variable.

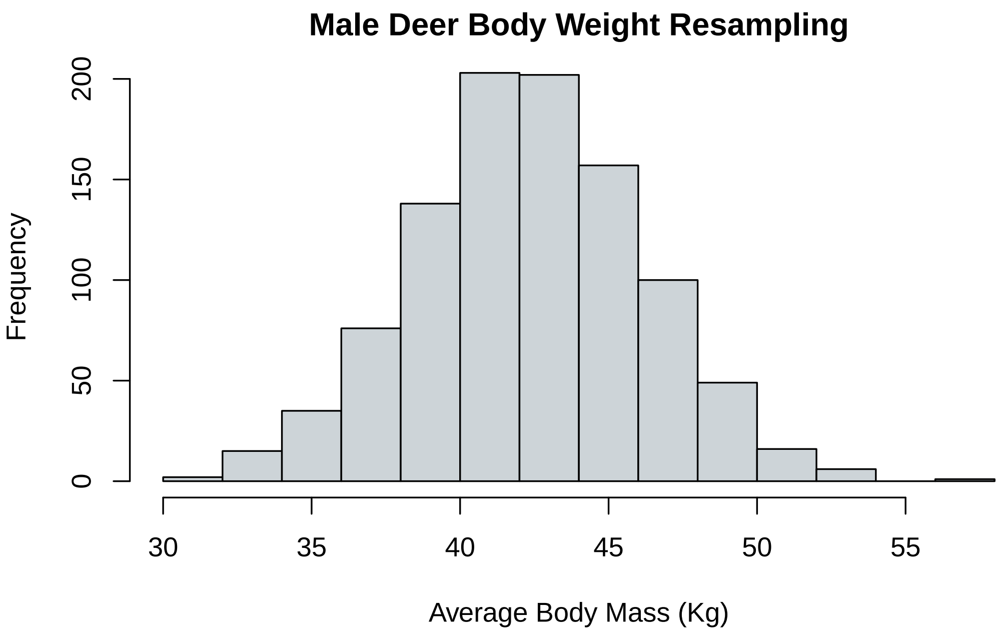
If we instead record the height for every student in the lecture hall and average them up we haven't done something dissimilar from adding up two heights to produce a new random variable. After all, an average is just a form of weighted sum. This means that the average height of the lecture hall is *also* a random variable. We even go so far as to say *every* sample mean is a random variable.

Given that a summary statistic like the mean is a random variable then all summary statistics should be random variables. If they're random variables then they should have a distribution. Don't let this thought experiment get away from you, it may seem chaotic but we always find order through statistics.

> "Though this be madness, yet there is method in't." — William Shakespeare, *Hamlet*, Act 2, Scene 2, lines 193–206

Let's start with a simple concept: Normally distributed data produces normally distributed statistics.

Body mass is a *truly* normally distributed variable (we'll have to trust that statement for now). If we look back to when we sampled deer body weights to generate a "fake" (simulated) experiment of measuring several Wisconsin's worth of deer:



We can play around with the original data on deer body weights, which we'll treat as a hypothetical population, and the experimental resampled data, which we'll treat as a hypothetical sample from the hypothetical population. When we calculate the "population" mean and the "sample" mean we should see that they're *nearly identical*:

```
# average of original "population" male  body weights
mean_pop=mean(subset(deer$Body.mass.in.kg,deer$Sex=="Male"))

# average from the resampling experiment
mean_sample=mean(males)
```

```
##
##  Population Mean: 42.2724
##  Sample Mean: 42.35554
```

What about the variance and standard deviation? That's a little problematic.

```
# variance and standard deviation from original "population" male body weights
var_pop=var(subset(deer$Body.mass.in.kg,deer$Sex=="Male"))
stdev_pop=sqrt(var_pop)

# variance and standard deviation from resampling experiment
var_sample=var(males)
stdev_sample=sqrt(var_sample)
```

```
##
## Population Variance: 431.3276
## Population Standard Deviation: 20.76843
## Sample Variance: 15.1552
## Sample Standard Deviation: 3.892968
```

A useful method for determining the difference between two methods is to calculate their ratio. The ratio isn't always a recognizable value but sometimes we can stumble into a good explanation for the difference.

$$\frac{\sigma^2}{s^2} = \frac{431.328}{15.155} = 28.46$$

This value might seem nonsensical, but let's think back: How many samples did we take for each average? We *ran* the experiment 1000 times but we only took 30 samples each time, which is close to the ratio between the two variances. We can also see that the ratio between the standard deviations is close to $\sqrt{30} \approx 5.477$.

$$\frac{\sigma}{s} = \frac{20.768}{3.893} = 5.335$$

Let's take the variance of the population and *divide* it by the sample size:

$$\frac{431.328}{30} = 14.377$$

While this might not look that close to the sample variance we calculated we need to remember that variance is a *squared* measurement. We're more concerned when there are very large errors rather than the single digit error we see here. This is showcased further when we convert to standard deviation:

$$\sqrt{14.377} \approx 3.792$$

The difference between this standard deviation and the "true value" we calculated is *less than* 3%, a level of error we shouldn't lose any sleep over.

We now have everything necessary to describe the distribution of the sample mean, $\bar{x}$.

$$\bar{x} \sim N(\mu_{\bar{x}}, \sigma^2_{\bar{x}})$$

$$\mu_{\bar{x}} = \mu$$

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Some quick algebra can also give us a general formula for the standard deviation of $\bar{x}$:

$$\sigma_{\bar{x}} = \sqrt{\sigma^2_{\bar{x}}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sqrt{\sigma^2}}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

But how are we supposed to handle data that isn't normally distributed? How do we know what the distribution of a sample mean is without gathering thousands of samples to justify our assumptions?

Student's t-Distribution

When we took draws from a deck of playing cards until we hit the ace of spades I made the statement that statisticians *knew* that the average of 30 draws was normally distributed, despite it being a discrete count with uniform probability.

```r
set.seed(73)
cat("Pull count to hit the ace of spades:", "\n", sample(1:52,30))
```

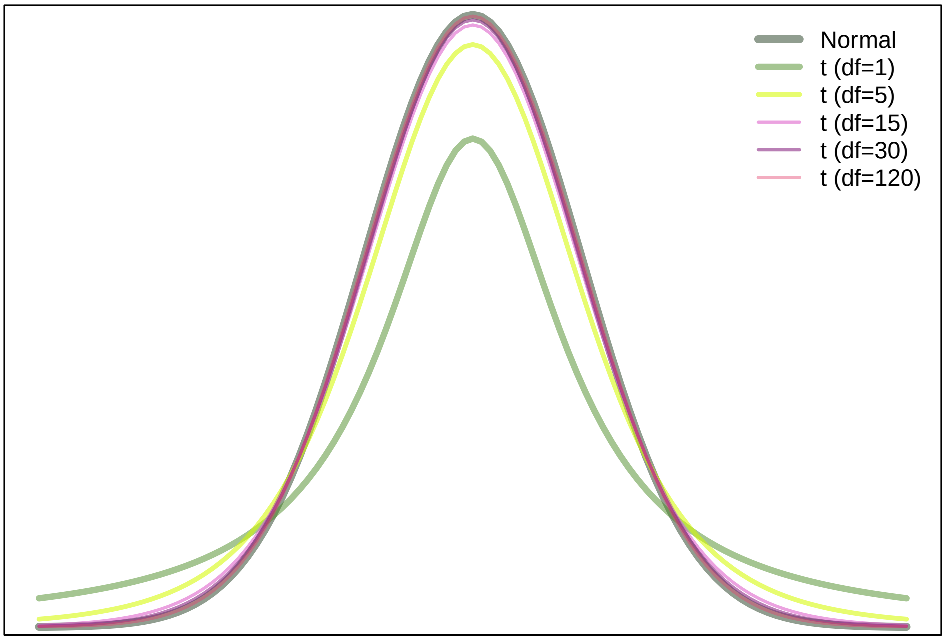
```
## Pull count to hit the ace of spades:
##  7 24 49 17 47 52 19 20 31 44 27 50 4 8 35 5 1 36 21 41 34 39 26 23 45 10 9 29 3 12
```

Yet every time I've shown this concept I've had to repeat the experiment and take averages hundreds of times before we begin to see the bell shaped curve. It seems like we could get that result without specifically averaging 30 draws if we're repeating the experiment so much, so why does this number 30 keep popping up? That's more of a history lesson than a proof or simulation study.

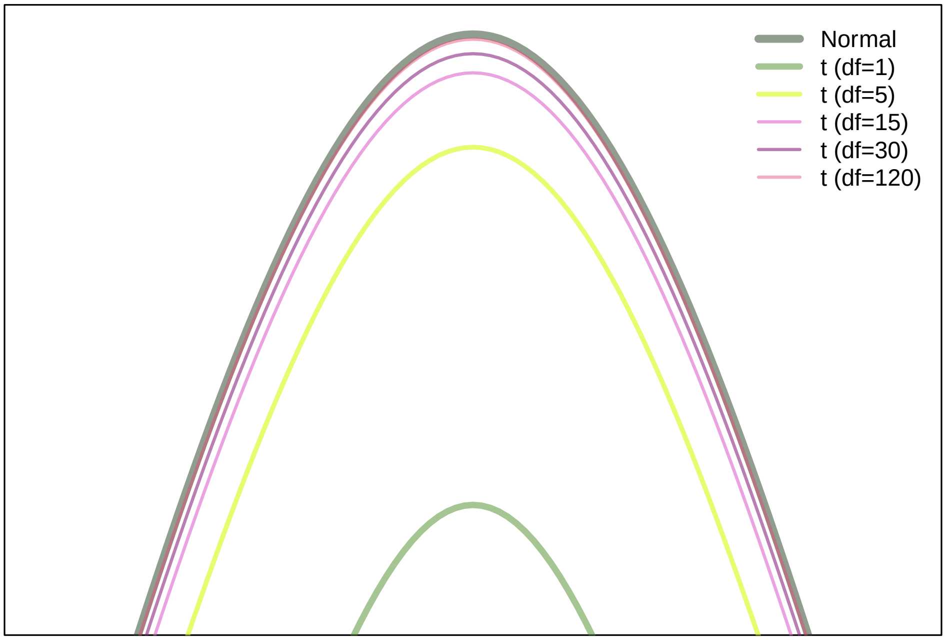
> It's the start of the 20^{th} century and Guinness (yes, the beer) is trying to cement it's dominance in the market. They begin hiring experts in chemistry, engineering, and statistics. Their thinking was that a proper industrialization of their brewing process required better formulas, better systems, and the elimination of variance in their product.
>
> William Sealy Gosset, a young chemist turned statistician, is hired onto the staff of the Dublin brewery and tasked with assessing and improving product quality while reducing production costs. At the time the field of statistics was dominated by Pearson's sampling philosophy, samples of 120 or greater from a normally distributed population produced normally distributed means. This is a very inconvenient barrier when you're trying to reduce the cost of quality control experiments. Gosset, while working with very low volume experiments, discovered that means from small samples *weren't quite normally distributed*.
>
> Gosset developed an internal report for Guinness documenting the rough distribution of these small sample means and it was met with excitement. He was encouraged to consult with Karl Pearson, one of the fathers of mathematical statistics, and with his guidance Gosset published his work. However due to a Guinness company policy, Gosset had to publish anonymously. He chose the pseudonym "Student", hence the name "Student's Distribution".
>
> His work was cemented by Sir Ronald A. Fisher who chose to express the random variables described by the Student's Distribution with the letter t. This leads us to today where we commonly refer to it as the "t-distribution" or "Student's t-distribution".

While it's a fun history lesson it's also a valuable piece of context behind the purpose of the t-distribution. The t-distribution is symmetric, bell-shaped, and centered around 0. It *sounds like* the standard normal distribution but the t-distribution is differentiated by its one parameter, degrees of freedom (df), which is generally calculated as $n - 1$.

As *df* shrinks, the *tails* of the *t*-distribution become **heavy**; they decrease at a lower rate which means that extreme values are *more likely.* As *df* increases towards ∞ the *t*-distribution *converges* to (becomes) the standard normal distribution.



It's hard to catch but if we zoom in on the peaks of these curves we should see a graphical explanation for the "*golden sample size*":



The explanation is that it doesn't exist.

The t-distribution will become a useful tool for the concepts in chapters 6 and 7, but as of right now it's primary contribution is helping us realize that sample size isn't as relevant as we might think.

The general rule of thumb in most statistics classrooms is that at $n = 30$ we can "*assume normality*" for sample means. This is due to the t-distribution, more specifically it's due to a textbook that Fisher published where he included a t-table (much like the z-table) that only went *up to df* $= 30$ before jumping to ∞ (the standard normal distribution). He did this because it made the table fit neatly onto one page.

When we're dealing with real statistical analysis we *should* collect as much data as possible. But to say there's a specific sample size where the sample means are *guaranteed* to be normal is a bit of a stretch.

We can assume almost any sample mean we encounter will *eventually* become normally distributed as $n \to \infty$, (we call this the asymptotic distribution), and we can apply the rule of thumb that any sample size at $n = 30$ should produce a normally distributed mean. In reality though a sample mean can be normally distributed as early at $n = 2$ or it could still fail at being normally distributed at *arbitrarily large* sample sizes.

The Central Limit Theorem

> It's difficult to explain the Central Limit Theorem *completely* without basic calculus or analysis. Any example is going to feel a little incomplete. If you're craving a better understanding I highly recommend you look into the formal proofs for the theorem; if you have the pre-requisite math skills they can be quite beautiful.

The basis of the Central Limit Theorem is that any standardized summation of samples from a population will inevitably become normally distributed, *regardless* of the original distribution of the data, as we continue to sample infinitely. This is the premise behind the deer body weight, coin flipping, and card pulling examples. Despite only one of those examples using normally distributed data they all produced symmetric, bell-shaped histograms.

We've seen this showcased a tiresome amount of times, so it's best to instead cement our statements on the Central Limit Theorem.

$$\bar{x} \sim N(\mu_{\bar{x}}, \sigma^2_{\bar{x}}) \ , \quad \text{as } n \to \infty$$

In practice this rule becomes reliable at $n = 15$ and next to impossible to break at $n = 30$. As such, in most cases, if we don't know anything about the original population our sample mean arose from then we should try to avoid assumptions of normality until we have $n = 30$ or we gain better context on the population.

Applying this to a real problem:

Data from a heart health survey suggests that the average age for smokers is $\mu = 40$ with a variance of $\sigma^2 = 200$. A random sample of 40 residents from Midwest towns is collected and their average age is calculated. What is the distribution of that sample mean, $\bar{x}$?

$$\bar{x} \sim N(\mu_{\bar{x}}, \sigma^2_{\bar{x}}) \ , \quad \text{since } n > 30$$

$$\mu_{\bar{x}} = \mu = 40$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma_{\bar{x}}^2}{n} = \frac{200}{40} = 5$$

$$\bar{x} \sim N(40, 5)$$

This concept can be applied further to calculate probabilities. What's the probability that the sample mean is 33?

$$\sigma_{\bar{x}} = \sqrt{\sigma_{\bar{x}}^2} = \sqrt{5} = 2.24$$

$$P(\bar{X} < 33) = P\left(Z < \frac{33 - 40}{2.24}\right) = P(Z < -3.125)$$

Since the z-score is halfway between two possible values on the z-table we would usually calculate the probability for -3.12 and -3.13 and average them. In this case we'll find that either these are the same probability or negligibly close:

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	**0.0009**	**0.0009**	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007

$$P(Z < -3.125) = 0.0009$$

Sometimes it's useful to multiply the probability for a z-score by 1000 or more to help with interpretation.

$$0.0009 \times 10000 = 9$$

We could then interpret this as saying:

> ”9 out of every 10000 sample means calculated from this population would be less than 33”

It's important to recognize that we're referring to the probability of a *statistic* arising, not an individual value. There are obviously younger smokers in the population and we'd have a good chance of selecting a much younger individual at random. But the probability of taking a random sample of the same size and calculating a smaller sample mean is very low.