Section 2: Learning from Data: Estimation, Confidence Intervals, and Testing Hypotheses

Carlos Carvalho and Rob McCulloch

1. IID Normal: Models, Parameters, and Estimates

We will be using IID draws from a normal distribution as a model for data.

Surprisingly often, real data looks like IID normal draws!!

What does this mean??

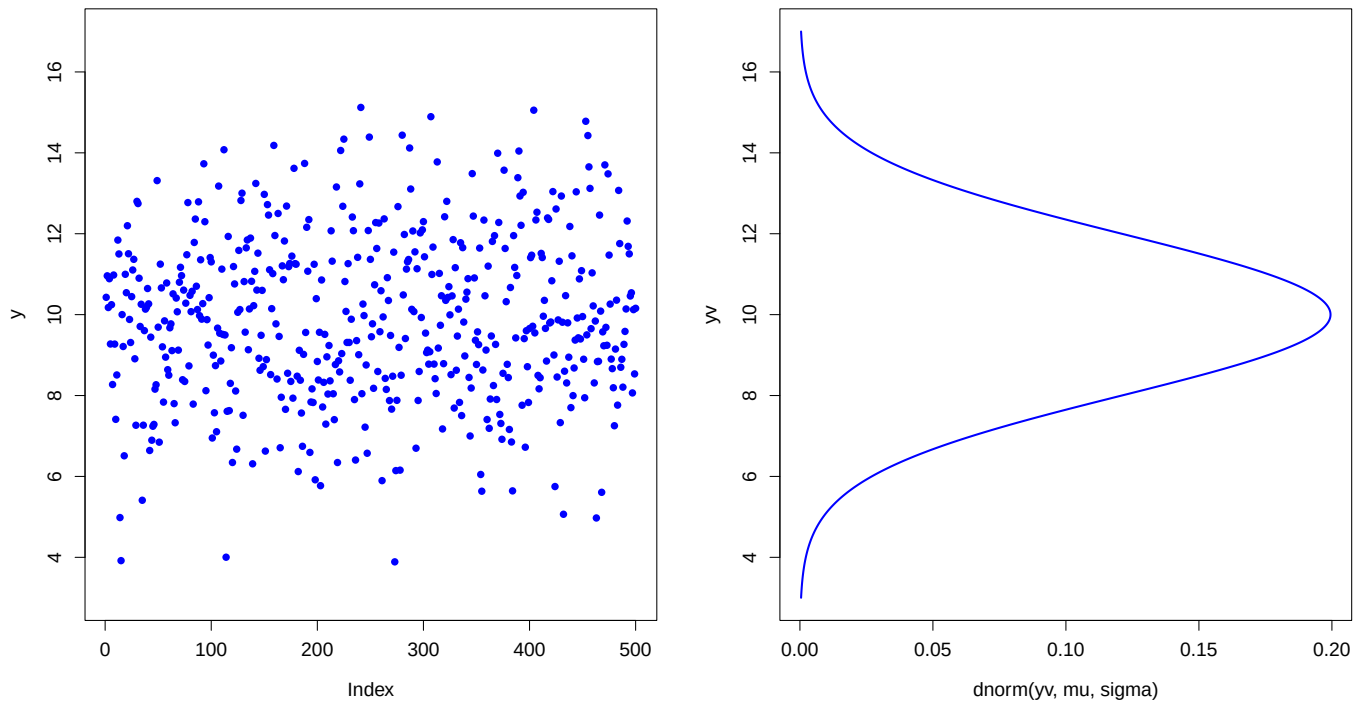
What do IID normal draws look like?

Let's simulate 500 draws from the $N(10, 4)$.

$Y_i \sim N(10, 4), IID.$

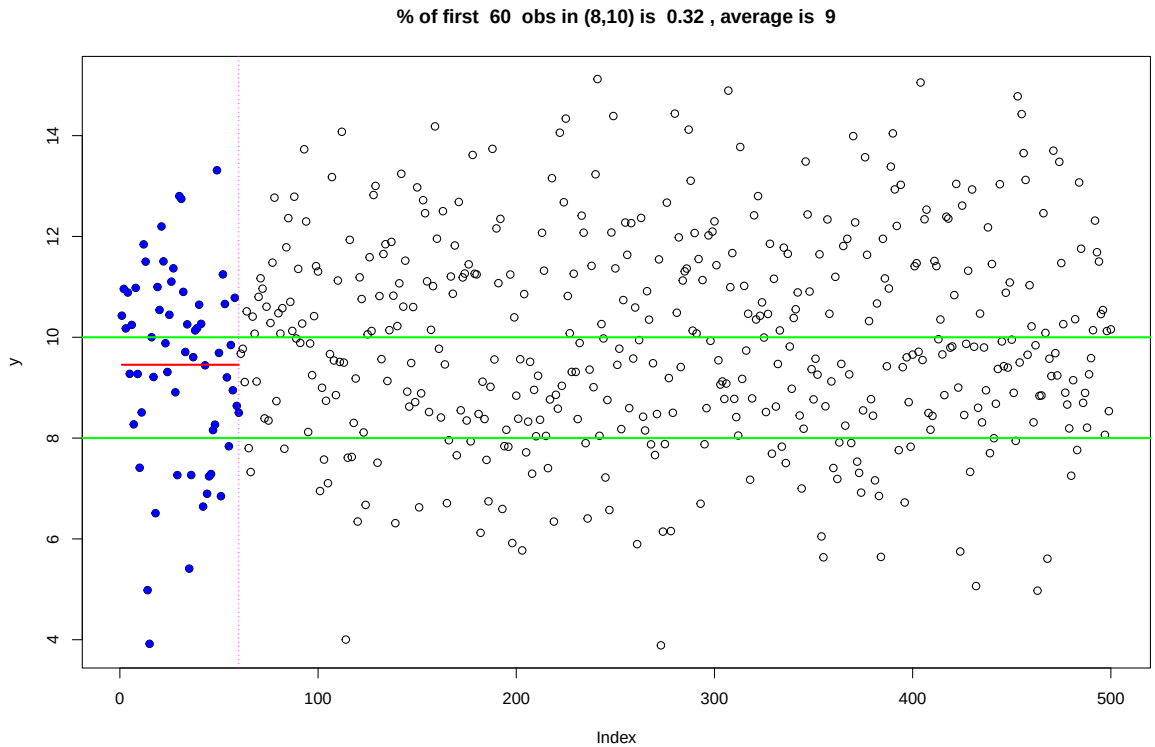
What do they look like?

Here is the sequence plot of the draws:



Because of the independence there is no obvious pattern in the sequence plot!!!

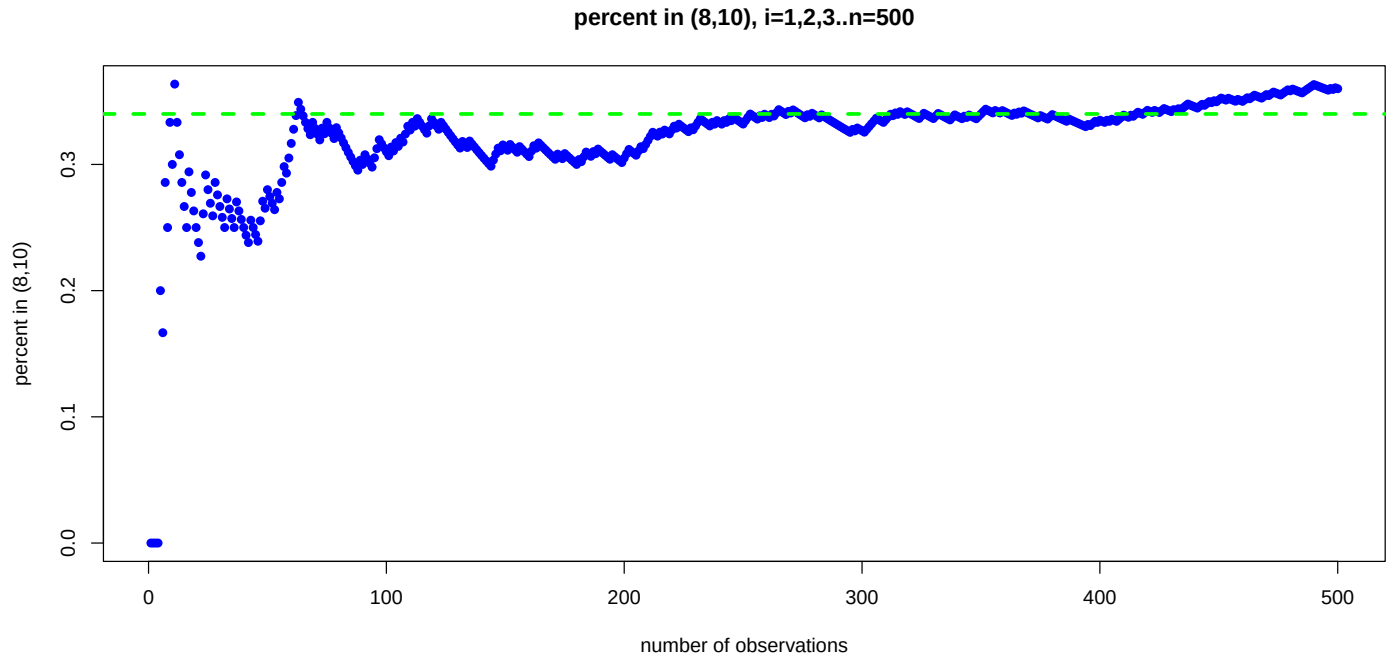
Here I take the first 60 observations and compute the percentage in (8,10) and the average.



What will happen to the percentage and average as I take more and more observations??

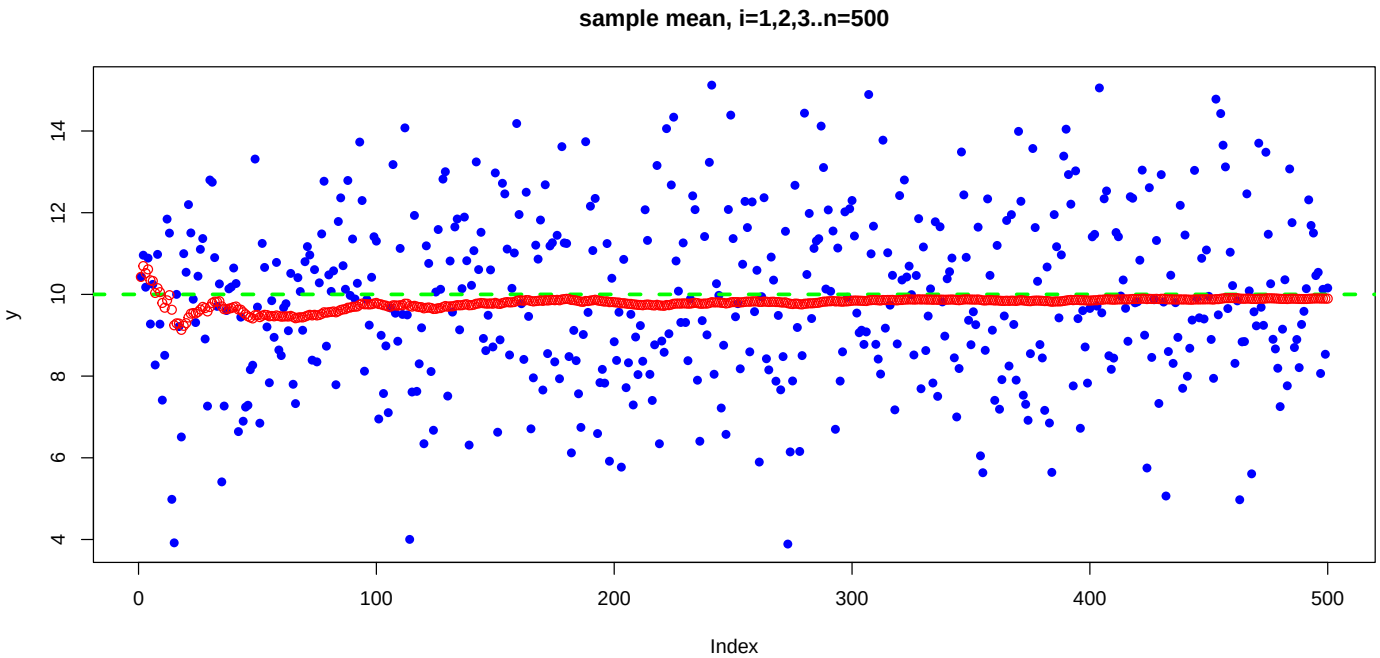
Half of .68 is .34!!

Here we plot the number of observations vs. percent in (8,10). Green line is drawn at .34.



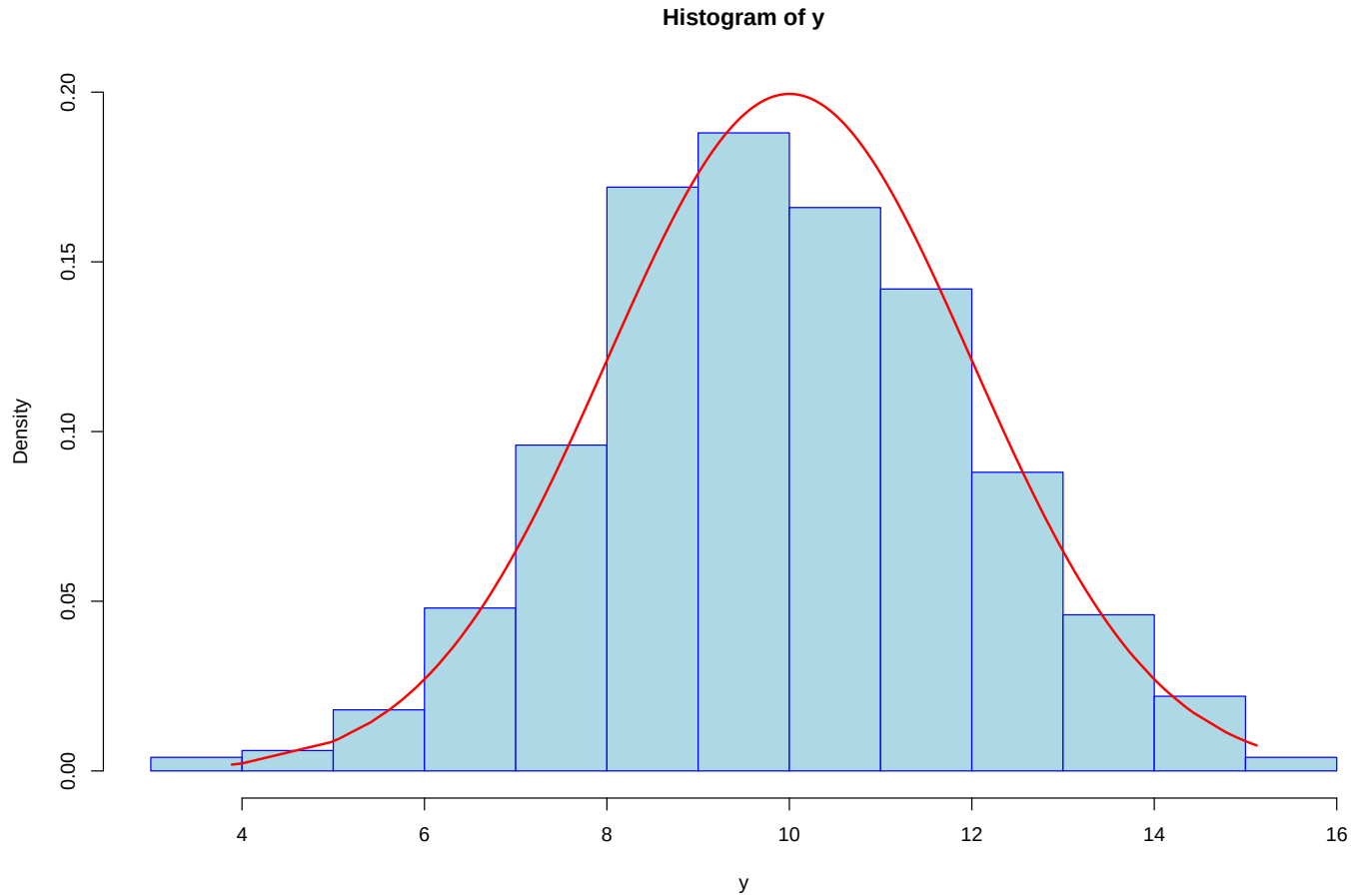
The observed frequency of times in the interval gets close to the probability of the interval as the sample size get large!!

Here we plot the number of observations vs. the average. Green line is at $\mu = 10$.



The observed sample mean gets close to the expected value of each Y_i as the sample size gets large !!

Here is the histogram with the normal $N(10, 4)$ density on top. The histogram has been scaled so that the *area* of each bar is the fraction of observations in the interval.



Histogram area $\approx$ percent obs in interval

$\approx$ probability of interval $\approx$ area under normal pdf.

The Weights Data:

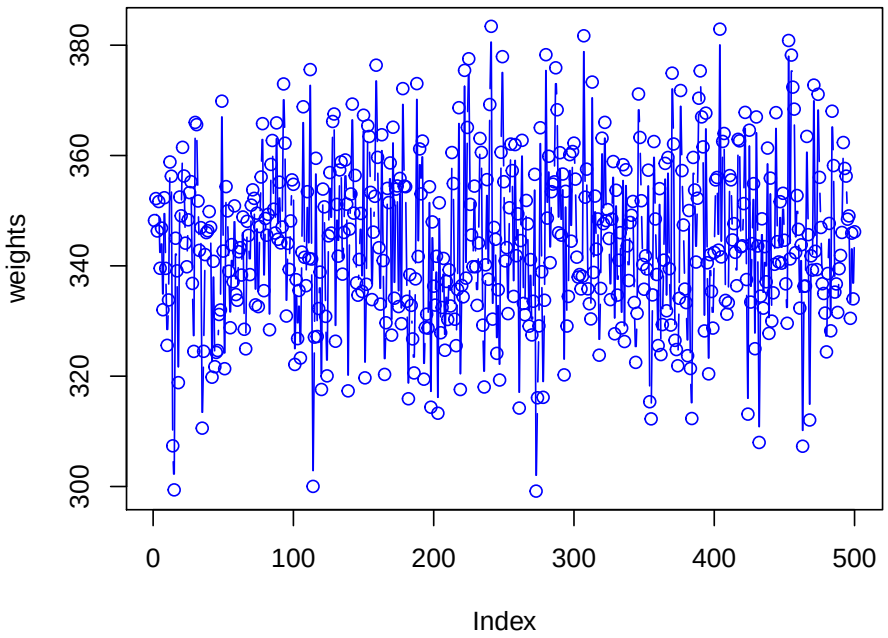
We are managing the process the put cereal into boxes.

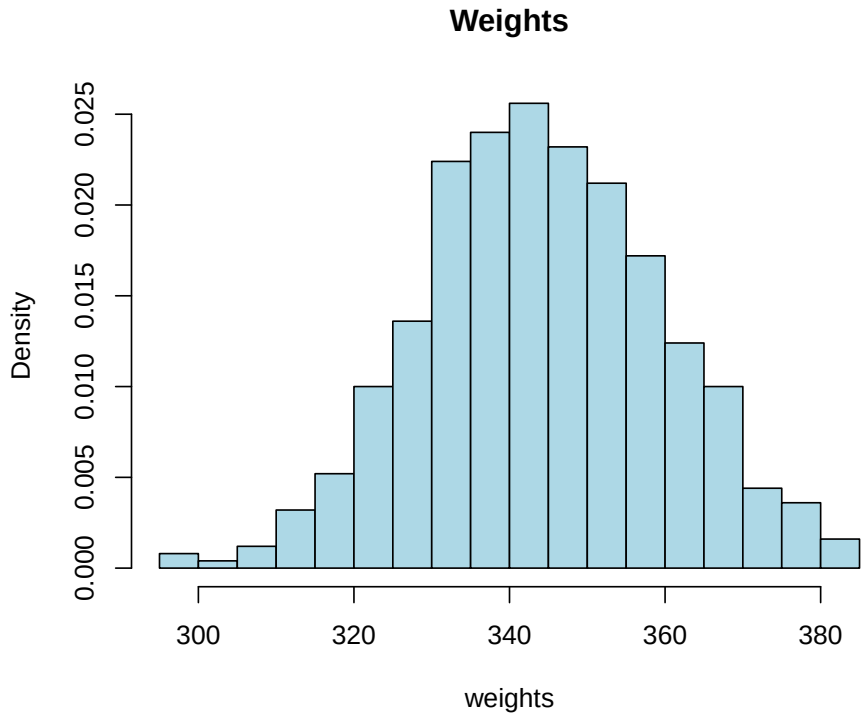
The amount going into a box should be 350 grams.

An acceptable range is (330,370).

We have data on the amount of cereal going into 500 boxes.

Here is the sequence plot and histogram for the weights data:





They look like IID normal draws !!!!

Given the weights "look normal", we would like to use the data to come up with good values for μ and σ.

We will use the sample mean to *estimate* the normal mean μ.

We will use the sample variance s_y^2 to *estimate* the normal standard deviation σ.

$$\bar{y} = \frac{1}{n} \sum y_i.$$

$$s_y^2 = \frac{1}{(n-1)} \sum (y_i - \bar{y})^2.$$

We have seen that the sample mean gets close to μ as the sample size gets bigger.

Similarly, the average squared distance (use n instead of $n-1$) gets close to σ^2.

Note:

s_y^2 is the *sample variance* of the numbers in y.

$s_y = \sqrt{s_y^2}$ is the *sample standard deviation* of the numbers in y.

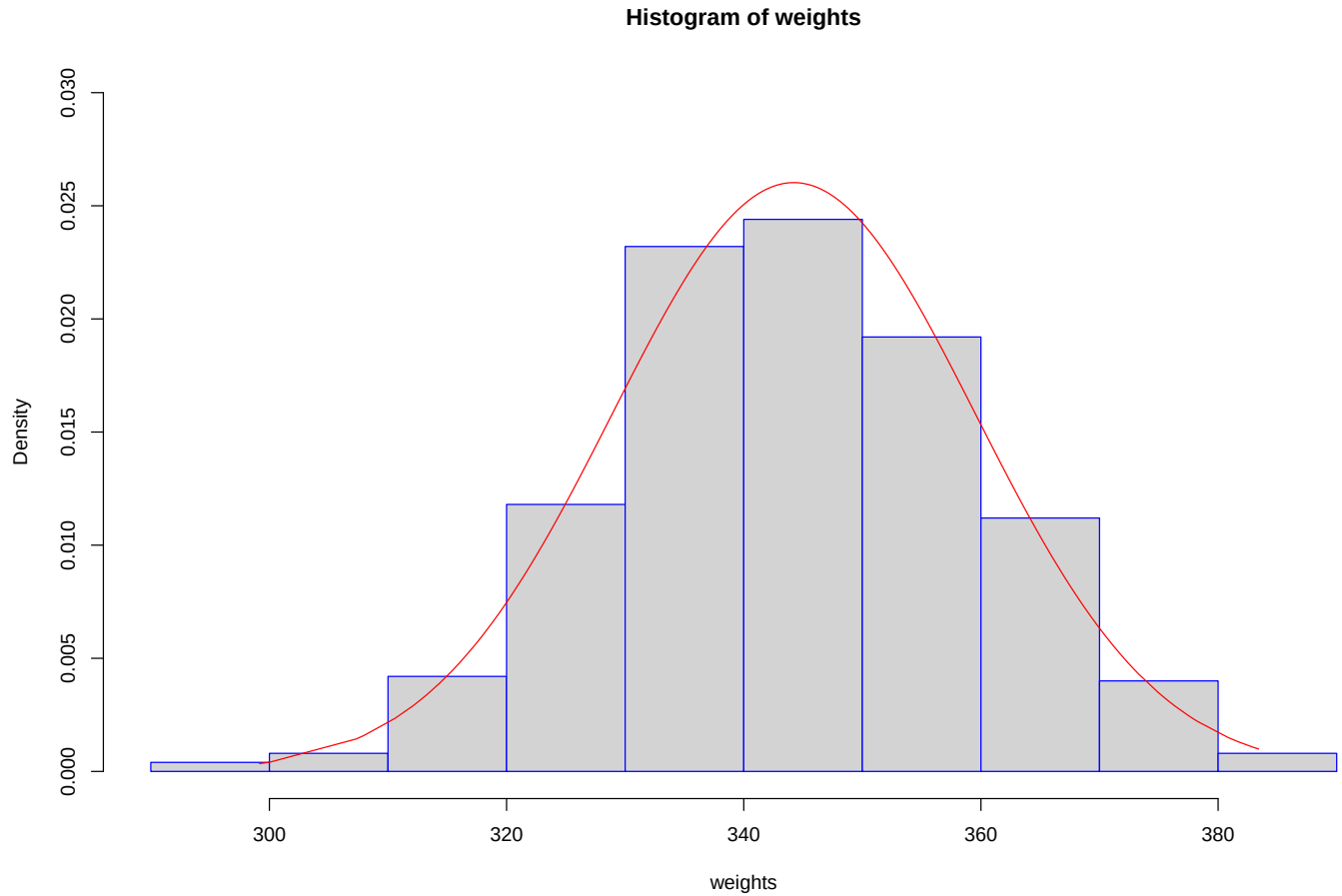
For our weights data, the mean is 344.2 and the standard deviation is 15.3 so our estimate of μ is 344.2 and our estimate of σ is 15.3.

Did it work??

Here is the histogram of our data with the

$$W \sim N(344.2, 15.3^2)$$

density plotted over it.



works great !!!!

Our Model:

Let W_i be the random variable represented the uncertain amount of cereal going into to i^{th} box.

$$W_i \sim N(344.2, 15.3^2), \text{ IID}$$

We have already observed the values $W_i = w_i,\ i = 1, 2, \ldots, 500$.

It *looks like* our model could plausible have generated these values!!

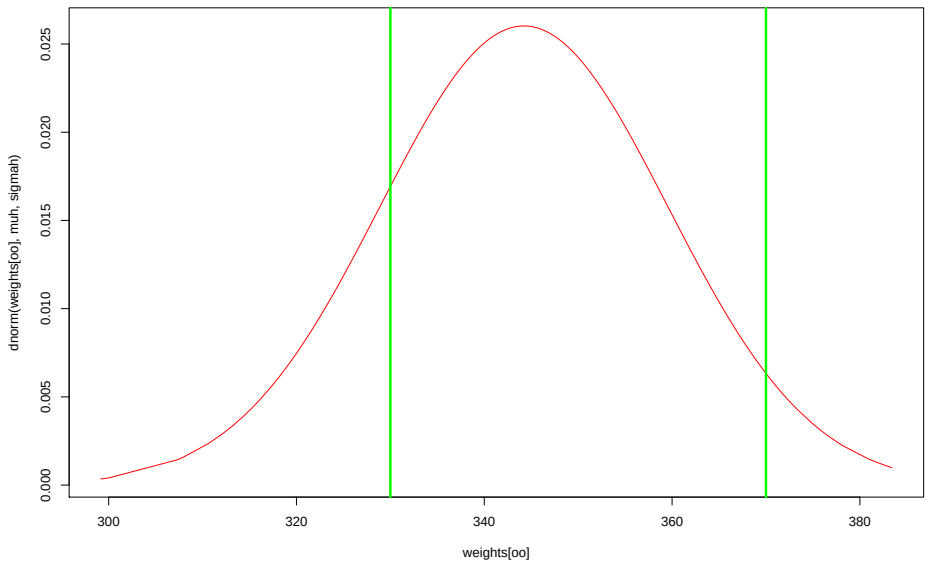
We can use the model to think about our process:
looks like μ is too small and σ is too big!!!

We can use the model to predict future performance.

Using the Model:

According to our model, what is the probability that the weight in the *next box* we fill will be in the acceptable range of (330,370) ?

```
> pnorm(370,344.2,15.3) - pnorm(330,344.2,15.3)
[1] 0.7774519
```



Not too good !!!

We need to improve the process !!

Big Picture:

The sample mean and standard deviation are often used as summaries of the data:

the average of the values, how spread out the values are.

In the context of the IID normal model we are using the sample mean and standard deviation as estimates of the *parameters* μ and σ in the model $W \sim N(\mu, \sigma^2)$.

A very general process in statistics is:

- Build a *probabilistic model* which could have generated the kind of data you see.
- Estimate the parameters of the model from the data.

2. Confidence Interval for a Normal Mean

For our weights data, our estimate of μ is 344.2.

What would we like μ to be??

We want to put 350 in!!

It looks like μ is too small, but our estimate does not have to be right. All we have said so for is that if n is "big enough" we should be alright.

We need to have to have some idea of the possible error for a given sample size n.

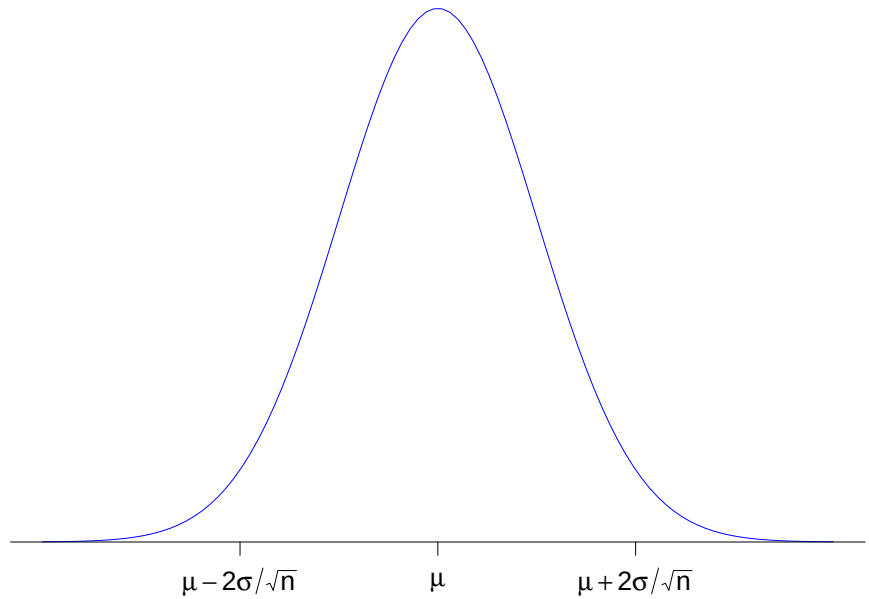
The *standard error* and associated *confidence interval* will quantify our possible error in estimation.

Imagine we are *about* to get a sample of size n and then use the sample mean to estimate μ, how is the sample mean related to μ?

For $Y_i \sim N(\mu, \sigma^2)$, IID,

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

This tells us how different our estimate is likely to be from the true μ!

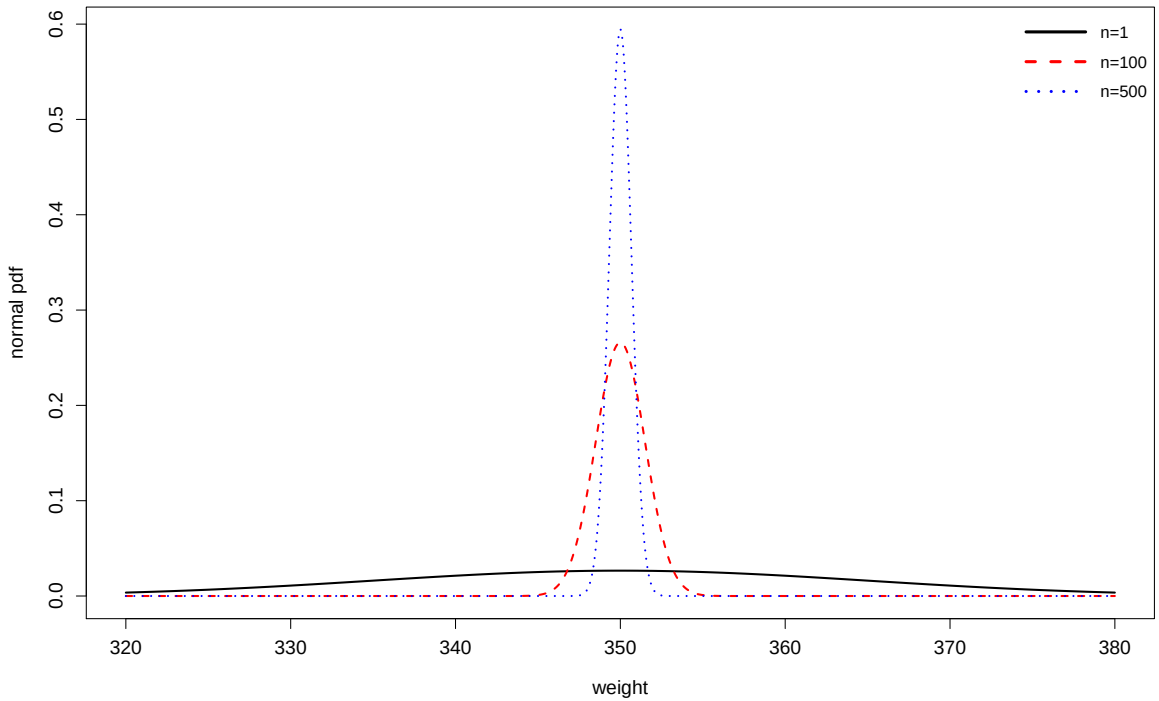


There is a 95% chance the error (the difference between μ and what $\bar{Y}$ turns out to be) will be less than $\pm 2\frac{\sigma}{\sqrt{n}}$!!!

In our weight example (Y is W), let's suppose the true values are $\mu = 350$ and $\sigma = 15$.

Here are the normal pdf's of $\bar{W}$ for different sample sizes.

Before you get your sample, $\bar{W}$ is a random variable !!!!!



We can probabilistically quantify how close $\bar{Y}$ is to μ given the sample size n !!

So, there is a 95% chance that our error will be less than $\pm 2\frac{\sigma}{\sqrt{n}}$.

and

$$\frac{\sigma}{\sqrt{n}} \approx \frac{s_y}{\sqrt{n}}$$

where the error in in this estimate ($\sigma \approx s_y$) is small enough that we don't have to worry about it for $n \geq 20$.

So, with probability 95%, our estimation error (using $\bar{Y}$ to estimate μ) is:

$$\pm 2\frac{\sigma}{\sqrt{n}} \approx \pm 2\frac{s_y}{\sqrt{n}}$$

For our weights data, $\bar{Y}$ turned out to be $\bar{y} = 344.22$.

$s_y = 15.33$.
$s_y/\sqrt{n} = 15.33/\sqrt{(500)} = .686$.

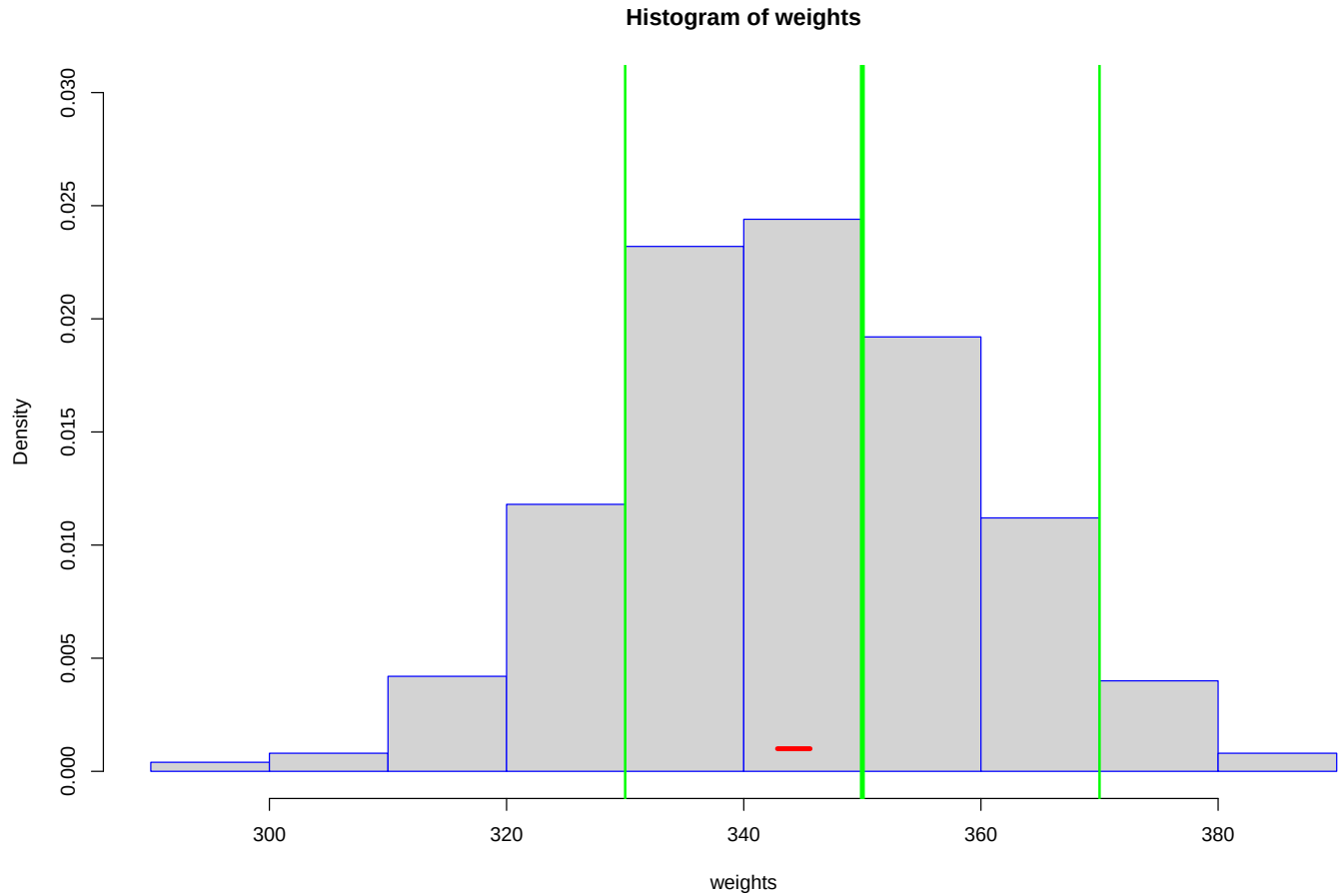
So, estimate $\pm$ error is:

$$344.22 \pm 2(.686) = 344.22 \pm 1.37 = (342.85, 345.59).$$

Pretty small!!!

Acting as if $\mu = 344$ is reasonable.

Here is the histogram of the weights data with the confidence interval for the mean plotted using the red bar.



The confidence interval gives us a sense of how big our error might be in estimating the true mean with the sample mean.

Summary:

Let's summarize the ideas, the procedure, *and* the jargon.

For, $Y_1, Y_2, \ldots, Y_n \sim N(\mu, \sigma^2)$, iid,

$$\bar{Y} = \frac{1}{n} \sum Y_i$$

is the *estimator* for μ.

Given our normal model, we *plan* to get a sample of size n and use the average as as estimate of μ.

Before we take the sample, $\bar{Y}$ is a random variable,
it is our *estimator*.

After we get the a sample, $\bar{Y}$ will turn out to be $\bar{y}$.
$\bar{y}$ is our *estimate*.

The sampling distribution of the estimator is

$$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n}).$$

The sampling distribution tells us what kind of estimate we are likely to get from our estimator given the true values of the *parameters* μ and σ.

Our estimate is likely to be close to the true value μ if:

- σ is small so that each Y_i tends to be close to μ.
- n is big.

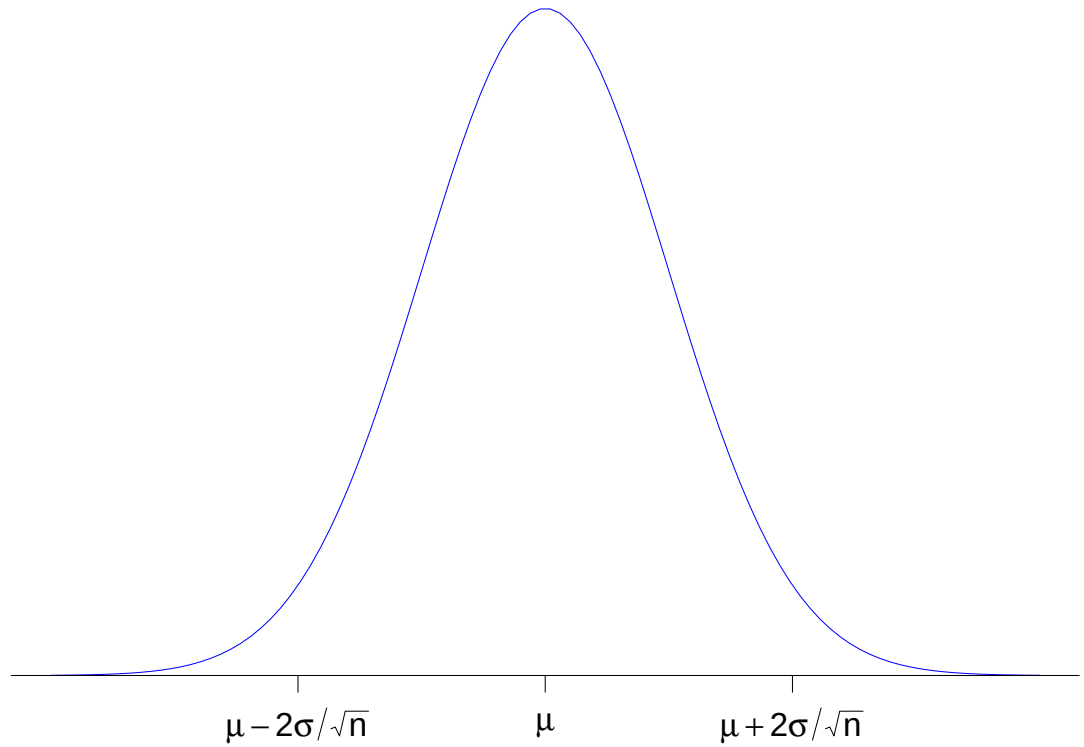
Note:

$$E(\bar{Y}) = \mu.$$

We say that $\bar{Y}$ is an *unbiased estimator*.

Any particular estimate will end up being too big or too small.

But, on average, they are right.



Note:

$$E(s_y^2) = \sigma^2$$

The sample variance is an unbiased estimator of σ^2.

This is why we divide by $(n-1)$, if we just divide by n, the estimate would tend to be too small.

The standard error of the mean is

$$se(\bar{y}) = \frac{s_y}{\sqrt{n}}$$

The standard error is an estimate of the standard deviation of $\bar{Y}$.

Given $Y_1, Y_2, \ldots, Y_n$ iid, $N(\mu, \sigma^2)$, for $n \geq 20$, the (approximate) 95% confidence interval for μ is

$$\bar{y} \pm 2\, se(\bar{y})$$

Before you take your sample, you have a 95% chance μ will be in the confidence interval!!

Small n:

For n less than about 20, just plugging in our estimate s_y can introduce too much error.

We are going to skip the details, but there is an adjustment you can make using the "tvalue".

> Given $Y_1, Y_2, \ldots, Y_n$ iid, $N(\mu, \sigma^2)$ the (exact) 95% confidence interval for μ is
>
> $$\bar{y} \pm tval\, se(\bar{y})$$

Tvals:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]
n	5.00	10.00	15.00	20.00	25.00	30.00	35.00	40.00	45.00	50.00	1000.00
tval	2.78	2.26	2.14	2.09	2.06	2.05	2.03	2.02	2.02	2.01	1.96

To get the tval: Excel: =tinv(.05,n-1), R: abs(qt(.025,n-1)).

Bottom line:

Confidence interval small: *GOOD, you know a lot*.

Confidence interval big: *BAD, you don't know a lot*.

Example:

Below is the histogram of monthly returns on "the Canadian market".
Canadian returns look normal!!

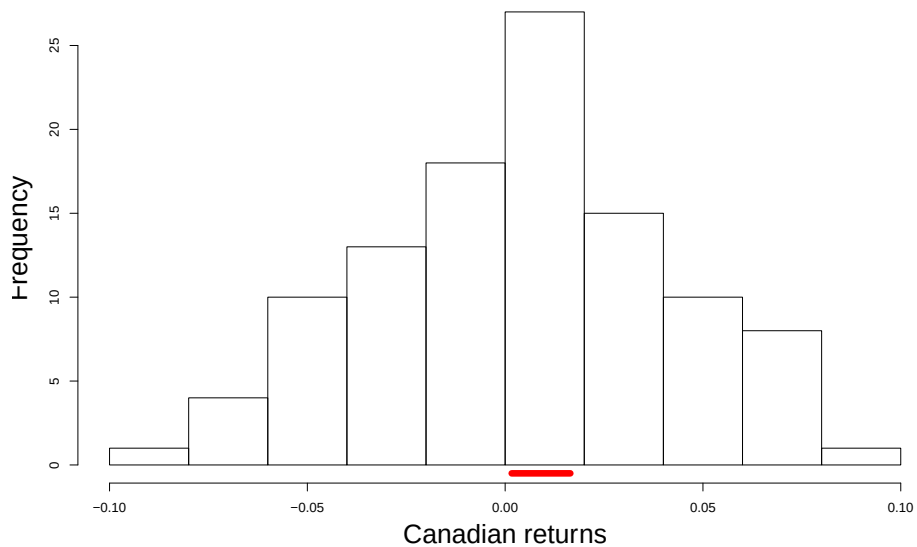
Let's get the 95% confidence interval for μ.

$n = 107.$
$\bar{y} = .009.$
$s_y = .038.$
$se(\bar{y}) = .0037.$
$2 * se = .0074.$

ci:

$.009 \pm .0074. \quad =$

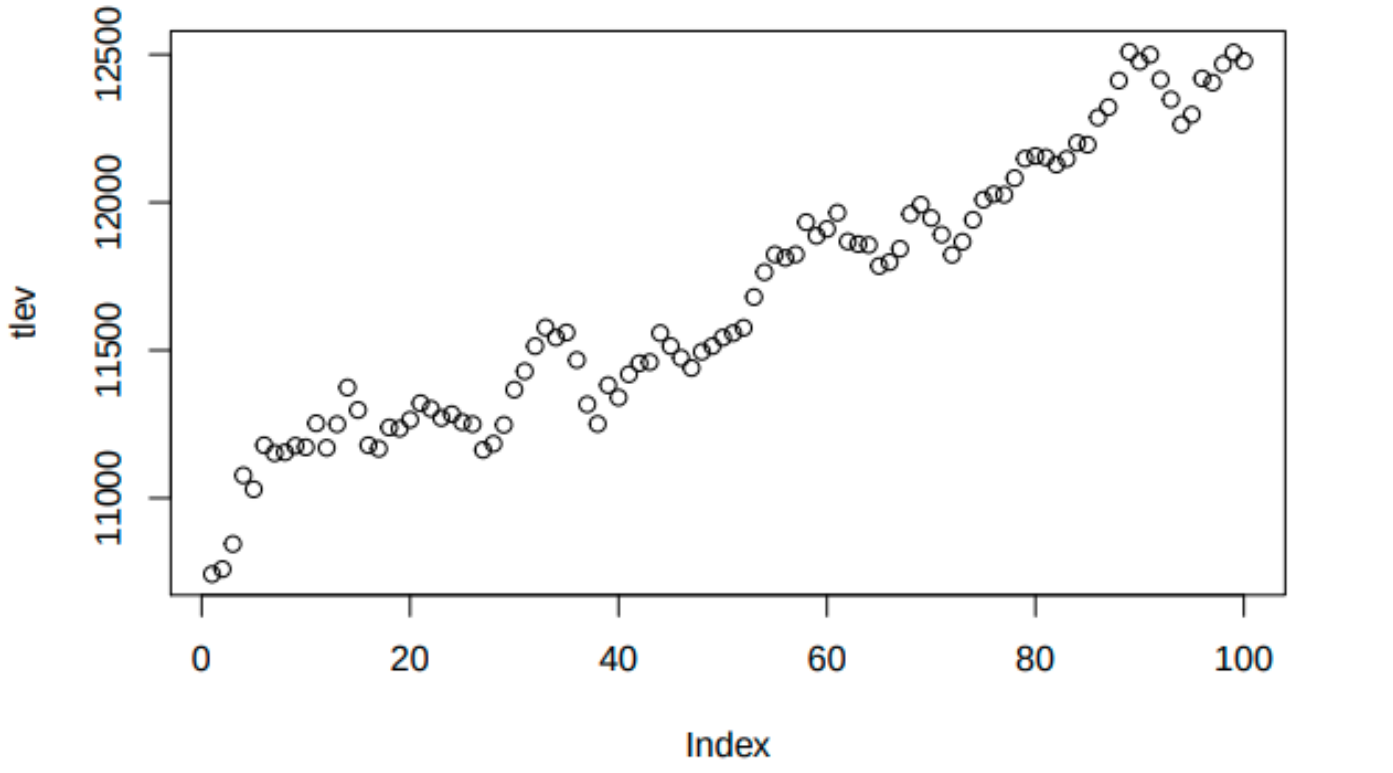
$(0.0017, 0.01650)$



Is this a big interval?

Not everything looks iid normal !!!!!

This data is a time-series of daily levels of a Japanese stock index.

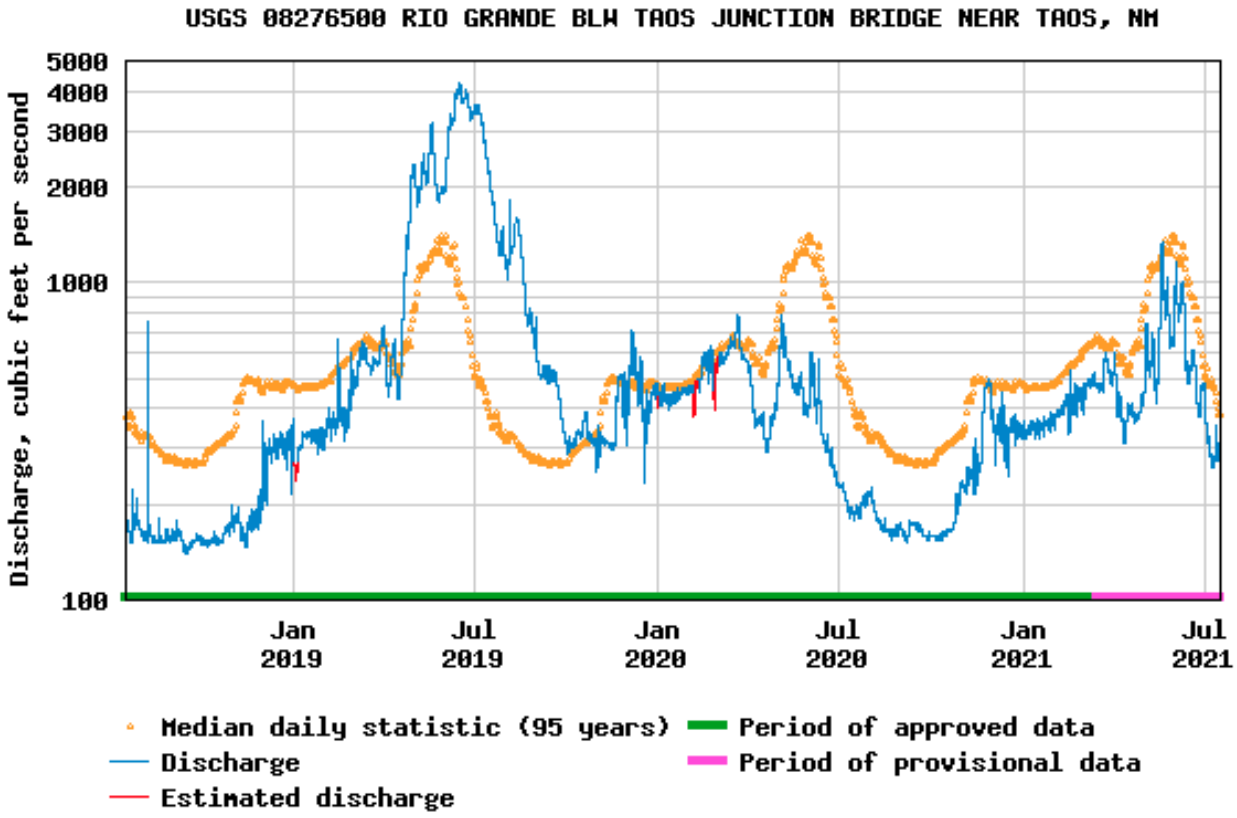


A clear upward trend.

Not everything looks iid normal !!!!!

Daily stream flow for the Rio Grande south of Taos, New Mexico, cfs = cubic feet per second.

Blue is daily cfs, orange is median cfs on that day over 100 past years.



A clear seasonal pattern.

3. The Confidence Interval for a Bernoulli p

For, $Y_1, Y_2, \ldots, Y_n$ iid Bernoulli, we use the sample proportion to estimate the true Bernoulli p. We call this estimate $\hat{p}$. The associated standard error is:

$$se(\hat{p}) = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

The (approximate) 95% confidence interval for a Bernoulli p is:

$$\hat{p} \pm 2se(\hat{p})$$

Example:

A random sample of 1,097 voters were asked how many would vote for candidate A.

We sample from the full population of voters without replacement and let,
Let $Y_i = 1$ if the i^{th} sampled voter would vote for A and 0 otherwise.

What is the distribution of the Y_i?

Let p denote the proportion of voters in the full population that would vote for A.

What is the distribution of Y_1?

What is the distribution of $Y_2 \mid Y_1 = y_1$?

When the population size is large, the Y_i are IID Bernoulli(p), where p is the true population proportion!!

44% responded they would vote for A.

Let p be the probability that a randomly selected voter would vote for A.

$\hat{p} = .44$.

$2 * \sqrt{.44 * (1 - .44)/1097} = 0.0299 \approx .03$.

Confidence Interval: $.44 \pm .03 = (.41,.47)$.

What if n is not way smaller than N ??

In the polling example, our assumption is that we take a sample of size n from a population of size N where $n << N$.

If n is not way smaller than N, the iid assumption may not be reasonable. In that case we need the *finite population correction* to get the right standard error:

$$se(\hat{p}) = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\,\sqrt{\frac{N-n}{N-1}}$$

What happens in this formula if $N >> n$?

Note:

Under the iid Bernoulli model, the sample proportion is an unbiased estimate:

$$E(\hat{p}) = p.$$

In the sampling from a large population example, since everyone has the same chance of being sampled, on average you get it right.

This can fail when we don't have a random sample.

> *These internet ratings are worthless, there are always a few people who are pissed off and those are the ones that go online and enter a rating.*
>
> *The new restaurant has 10 ratings and they are all 5 out of 5. All that tells me is that the owner has exactly 10 friends.*

Sampling is tricky!!!!

https://www.nytimes.com/2020/07/16/upshot/polls-biden-trump-how-accurate.html?action=click&module=Well&pgtype=Homepage&section=The%20Upshot

Perhaps most important, many pollsters now weight their sample to properly represent voters without a college degree.
The failure of many state pollsters to do so in 2016 is widely considered one of the major reasons the polls underestimated Mr. Trump's support. Voters without a four-year college degree are far less likely to respond to telephone surveys |
and far likelier to support Mr. Trump.
By our estimates, weighting by education might move the typical poll by as much as four points in Mr. Trump's direction.

Voters without a four-year college degree are far less likely to respond to telephone surveys — and far likelier to support Mr. Trump.

4. The Improved Cereal Process

Remember our cereal box filling process was off center and too variable.

Definitely **not** $6\,\sigma$ quality!!

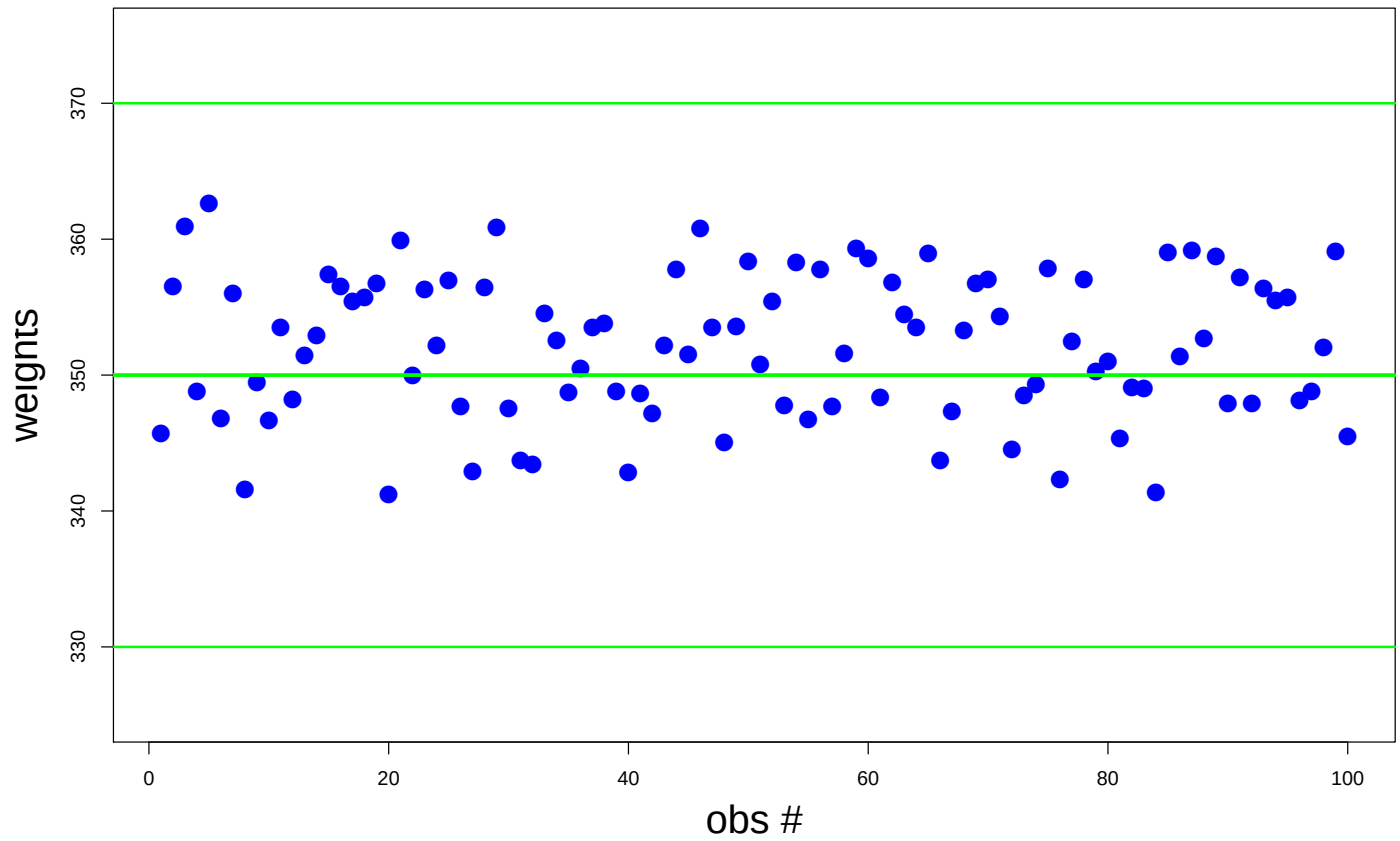
The process was supposed to be centered at 350 and $350 \pm 20 = (330, 370)$ is the range of acceptable weights.

You go on vacation and while you are away, your assistant works on the process.

When you get back, the assistant **claims** the process is much tighter and it is correctly centered, that is, $\mu = 350$ **!!!**

There is data on the weights from 100 boxes from the new process.

Wow, it does look much better !!!!!



The average weight is 352.08.

The sample standard deviation is 5.3.

You say the mean is a little high, but your assistant says, "hey, it is just a sample of 100, I could still be right that the true mean is 350!!"

When someone *claims* they know the true parameter value we can *test the hypothesis* that the claim is true.

Your assistant *claims* $\mu = 350$.

We will test the hypothesis that $\mu = 350$.

The basic reasoning behind testing is:

If the claim were true, what would the data look like?

If the data looks like something you could get
if the claim were true,
you cannot reject it.

But, if you get something that *would be* unlikely
if the claim were true,
you can reject it.

For a hypothesis about μ given the $N(\mu, \sigma^2)$ model, we "look at the data" by looking at the sample mean.

We ask **if** the claim were true, what kind of sample mean would we get?

We got a mean of 352.08.
Is that likely if $\mu = 350$???

If the claim is true ($\mu = 350$) then

$$\bar{Y} \sim N(350, \sigma^2/100).$$

$$\bar{Y} \sim N(350, \sigma^2/100).$$

Well, we have a problem since we don't know σ.

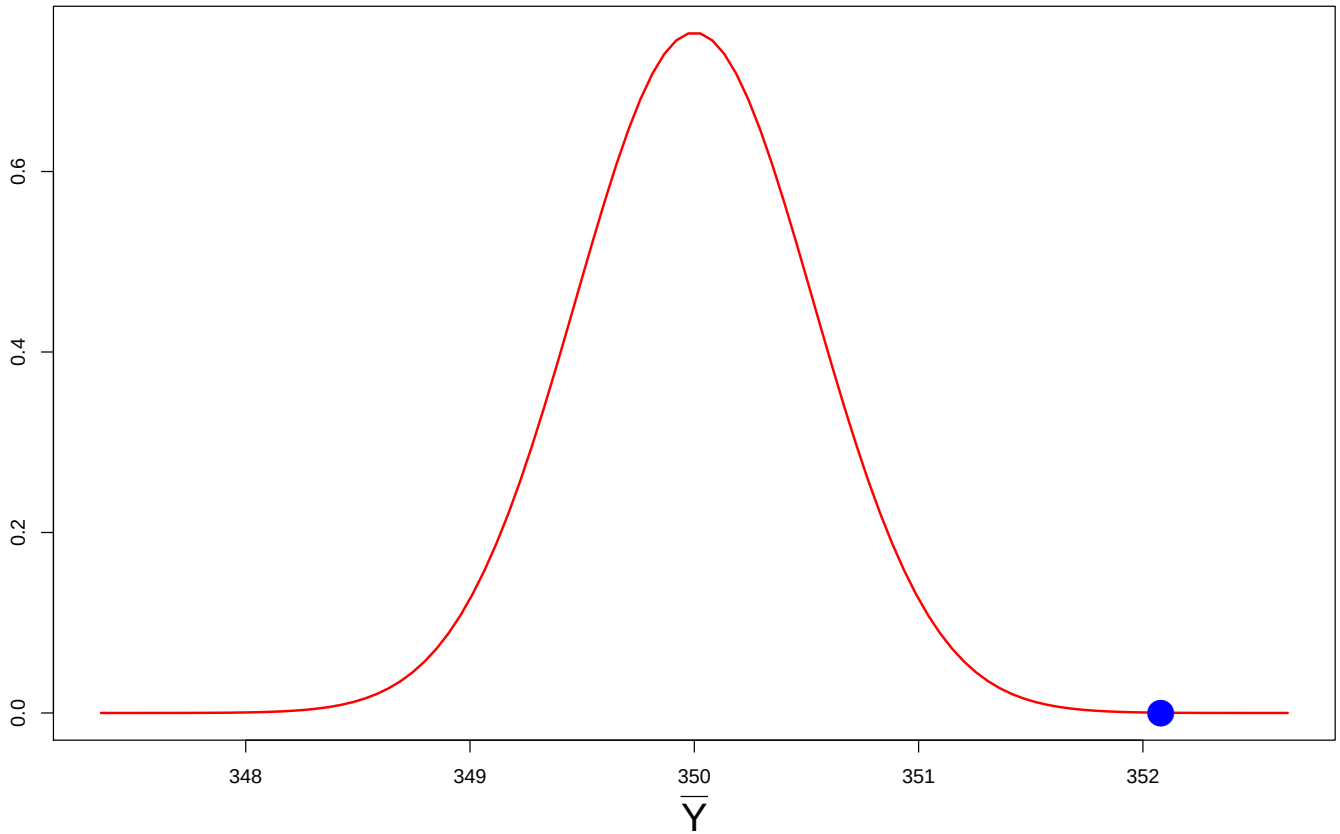
However for n greater than about 20 (sound familiar?) it turns out you can plug in the sample standard deviation without making too much of an error.

So, now we have, **if** the claim is true,

$$\bar{Y} \sim N(350, 5.3^2/100) = N(350, .53^2).$$

Notice that .53 is just $se(\bar{y}) = \frac{s_y}{\sqrt{n}} = \frac{5.3}{10}$.

Here is the density of $\bar{Y}$ (**if** the claim is true) with the observed $\bar{y}$ (the big blue dot).



If the claim $\mu = 350$ were true, it would be quite unlikely to observe $\bar{y} = 352.08$, so we reject the claim.

To further get a sense of how unusual $\bar{y} = 352.03$ would be **if** the claim were true, we can "z it".

If the claim were true the right way to z it would be:

$$z = \frac{\bar{y} - 350}{\sigma/\sqrt{n}} \approx \frac{\bar{y} - 350}{se(\bar{y})} = \frac{352.08 - 350}{.53} = 3.92.$$

If the claim were true, getting $\bar{y} = 352.08$ would be just like getting 3.92 from a standard normal - *not too likely*.

We reject the claim.

5. Testing a Normal Mean

Here is the formal summary and jargon for what we just did.

To test the *null hypothesis* (the claim)

$$H_o : \mu = \mu^o$$

against the *alternative hypothesis*

$$H_A : \mu \neq \mu^o$$

We compute the *test statistic*

$$t = \frac{\bar{y} - \mu^o}{se(\bar{y})}$$

We reject at level .05 if $|t| > 2$.

The test statistic is called a " t statistic".

For small n is should look like a draw from the t distribution - we are skipping this.

For larger n (≥ 20) the t should look like a z!!.

If the null hypothesis is true, the t statistic should look like a draw from the standard normal.

Example:

Here is the R output for the test we have done ($\mu = 350$).

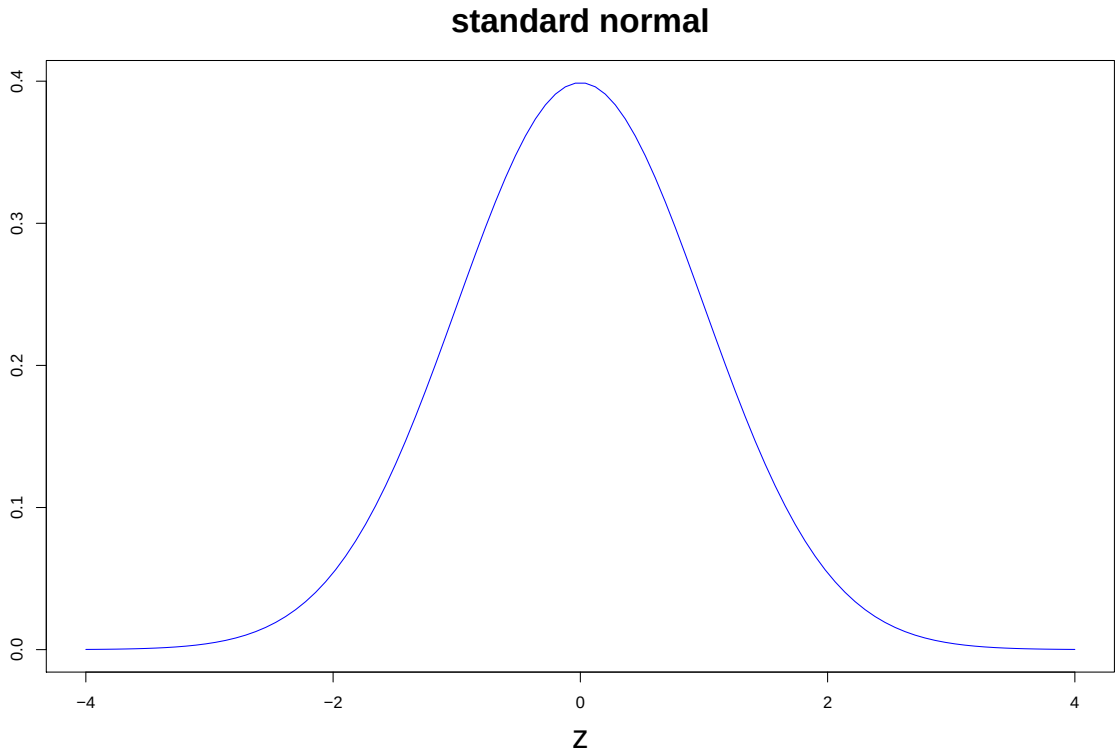
```
> t.test(weights,mu=350)

        One Sample t-test

data:  weights
t = 3.9323, df = 99, p-value = 0.0001562
alternative hypothesis: true mean is not equal to 350
95 percent confidence interval:
 351.0308 353.1306
sample estimates:
mean of x
 352.0807
```

If the null were true, the sample mean we got would be like getting 3.9 from the standard normal; the t stat is bigger than 2 $\Rightarrow$ reject.

Note:

The level of the test is the probability of rejecting a null hypothesis that is true.

If the null is true, the t should look like a z, a standard normal draw.



If we reject when $|t| > 2$ then the chance of rejection is .05, (**if** the null is true).

Note:

If $|t| < 2$, *we do not accept the null hypothesis* -
we "fail to reject it" !!!

What the....

This is because the t stat can be small for two very different reasons:

$$t = \frac{\bar{y} - \mu^o}{se(\bar{y})}$$

(a) You could have the top is very, very small and the bottom is small, in this case you might accept.

(b) *But*, you could also have a small t just because the bottom is very big, in which case your data is not informative and you should not accept the null since that would imply you decided it is true.

Example:

Let's test whether the true mean return for Finland is 0, using the conret.csv data set.

Here is the R output:

```
> t.test(finland)

        One Sample t-test

data:  finland
t = 1.3138, df = 106, p-value = 0.1918
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.005043264  0.024856348
sample estimates:
  mean of x
0.009906542
```

Here we cannot reject, but we sure do not want to accept, the confidence interval is huge!!

We *fail to reject the null hypothesis*.

Confidence Intervals are less confusing !!!

In the Finland example we could see what was going on by looking at the confidence interval - there was a lot of uncertainty!!!

In the weights example the confidence interval is (351.0308, 353.1306).

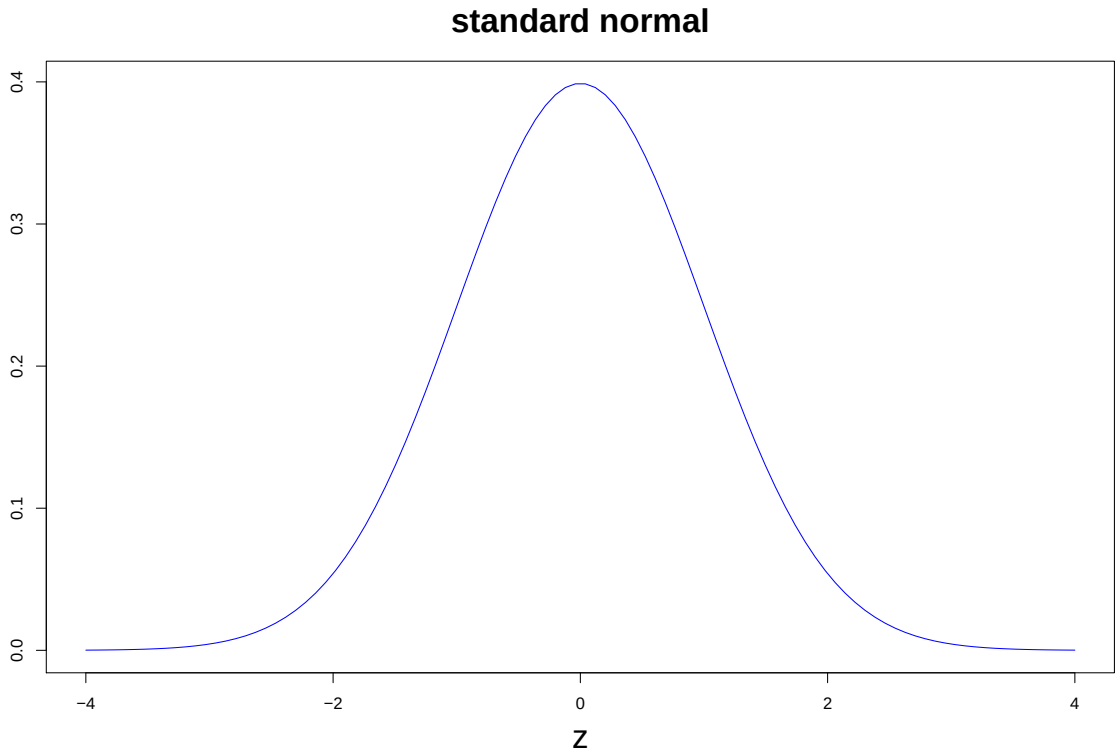
Your assistant says "Ok, maybe it is not perfectly centered, but it looks like we are pretty sure it is darn close!!"

In both cases, the confidence interval seems much more useful than the test!!

The only catch with the confidence interval is that you have to understand what your problem is when you decide if it is big or small but that is a good thing!!

6. p-values

If the null is true, our t test-statistic should look like a draw from the standard normal, *our t should look like a z*.



If we reject when $|t| > 2$, then, $P(\textit{reject} \mid H_o \textit{ true}) \approx .05$.

But sometimes we don't have to make a decision right away.

Rather than just rejecting/(fail to reject) we want to simply report *how far out in the tail* the t-statistic is.

The further out it is, the "more evidence" there is against the null.

The p-value is just a way of measuring "how far out in the tail" the t-statistic is.

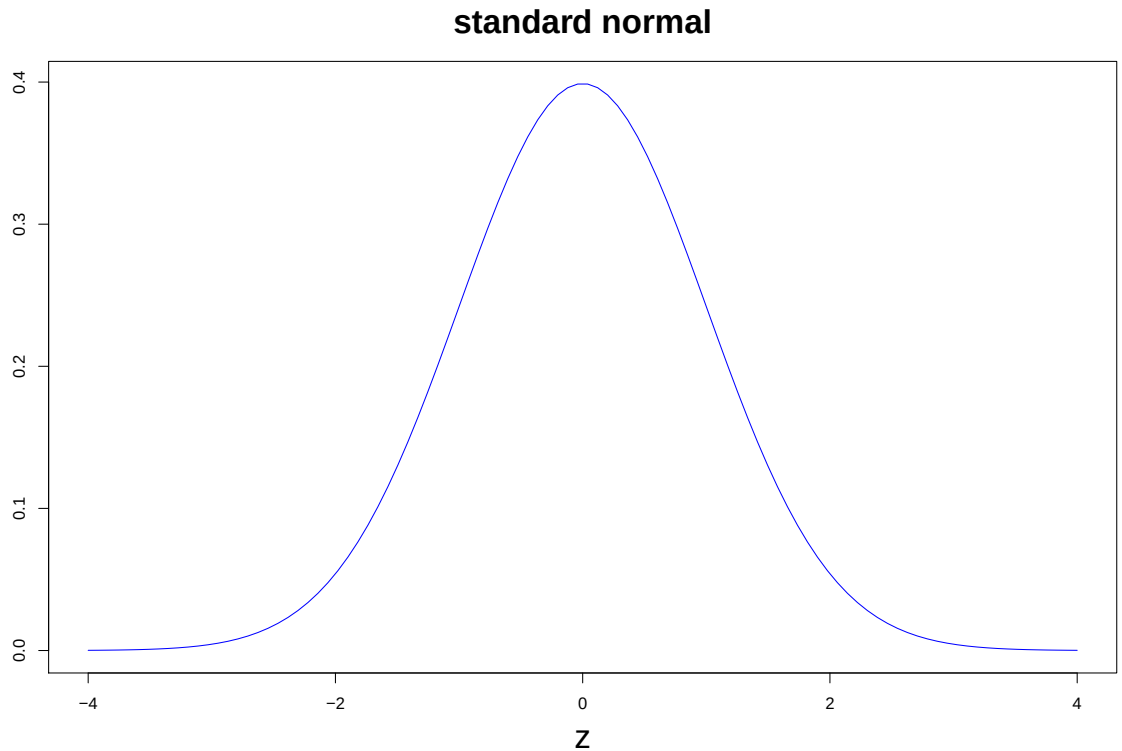
The p-value is the probability of getting a t test-statistic as far out or farther, **if** the null is true.

t = 1.
p-val is prob of greater than 1 or less than -1 = .32.

t = -2.
p-val is prob of greater than 2 or less than -2 = .05.

t = 3.
p-val is prob of greater than 3 or less than -3 = .0027.

t = 4.
p-val is prob of greater than 4 or less than -4 = .00006.



Note: p-value $= 2 * F(-|t|)$, where F is the standard normal CDF.
CDF: $F(z) = P(Z < z), Z \sim N(0, 1)$.

Example:

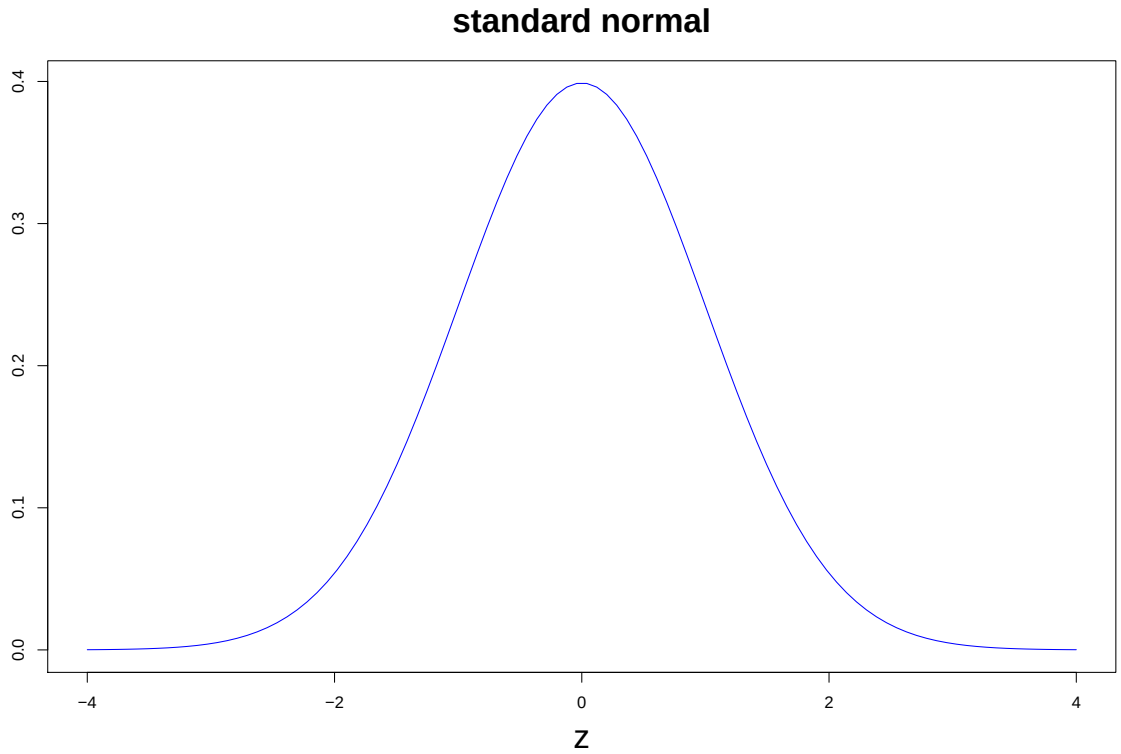
Recall that we tested whether the true mean of the Finnish returns is equal to 0.

```
> t.test(finland)

        One Sample t-test

data:  finland
t = 1.3138, df = 106, p-value = 0.1918
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.005043264  0.024856348
sample estimates:
  mean of x
0.009906542
```

The reported p-value is about:

$$2 * F(-1.3138) = .1889.$$

7. p-values and testing



If t is a little less than 2, the p-value will be a little bigger than .05.

If t is a little greater than 2, the p-value will be a little smaller than .05.

If you want to test at level .05, you can reject if the p-value is less than .05 .

This works generally,

to test at level α,

reject if p-value $< \alpha$ *!!!*

SMALL p $\Rightarrow$ REJECT.

We use this testing/p-value setup for all kinds of Hypotheses !!!

Example:

Previously, we modeled the returns on "Canada" as iid normal.

We did this by eye-balling the time-series plot and the histogram.

We can test the null-hypothesis that the returns are normal, assuming they are iid.

```
Shapiro-Wilk normality test

data:  can
W = 0.98607, p-value = 0.3307
```

big p-value $\Rightarrow$ Fail to reject.

We can test if the Canadian returns are iid:

```
Runs Test

data:  can
statistic = 0.31954, runs = 50, n1 = 49, n2 = 46, n = 95, p-value =
0.7493
alternative hypothesis: nonrandomness
```

big p-value $\Rightarrow$ Fail to reject.

We can test if the true mean of the Canadian returns is 0:

```
One Sample t-test

data:  can
t = 2.4467, df = 106, p-value = 0.01606
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.001719553 0.016411288
sample estimates:
  mean of x
0.009065421
```

small p-value $\Rightarrow$ reject.

Warning:

The tests are not infallible.

Inevitably, for complex hypotheses, the tests will be more sensitive to some alternatives than others.

The best test is the intra-ocular test !! (look at your data, it should hit you right between the eyes !!)

Fama:

> *With formal statistics, you say something - a hypothesis - and then you test it. Harry always said that your criterion should be not whether or not you can reject or accept the hypothesis, but what you can learn from the data. The best thing you can do is use the data to enhance your description of the world. That has been the guiding light of my research. You should use market data to understand markets better, not to say this or that hypothesis is literally true or false. No model is ever strictly true. The real criterion should be: Do I know more about markets when I'm finished than I did when I started?*

For example, look at the CI and interpret it, rather than blindly accept the test!!