Confidence Intervals, Hypothesis Tests, and p-values

Rob McCulloch

1. The Audit

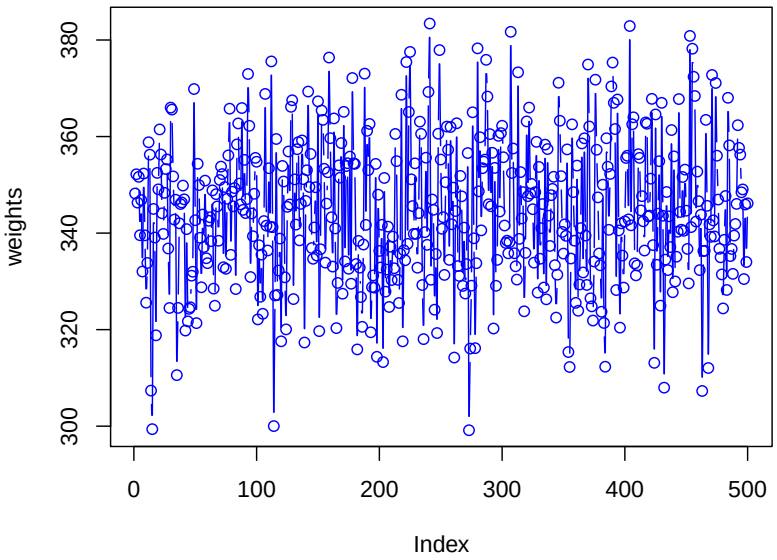
You are in charge of the process which fills cereal boxes.

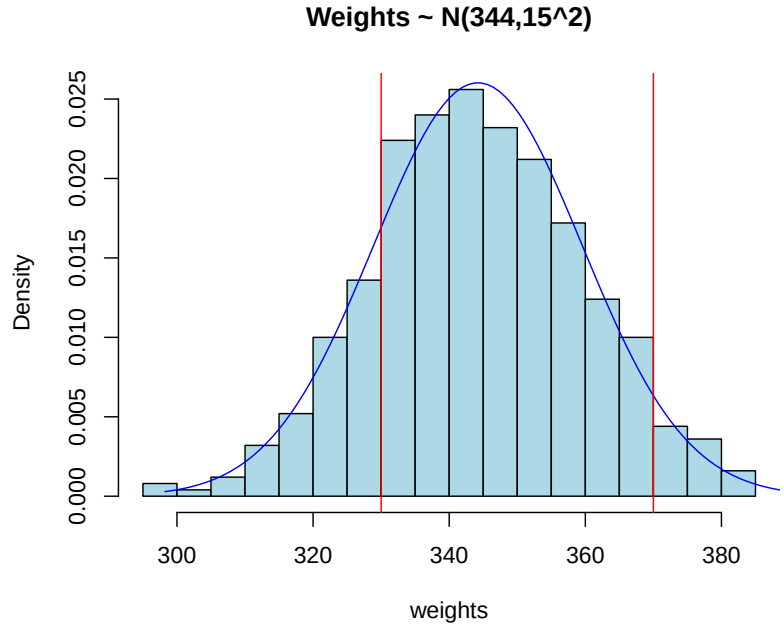
Ideally, the amount of cereal going into a box should weigh 350 grams but a weight in range $350 \pm 20 = (330, 370)$ is considered acceptable.

You have modeled the performance of the current process as iid normal $N(344, 15^2)$.

The weights look like iid draws from a $N(344, 15^2)$.

$Y_i \sim N(344, 15^2)$, where Y_i is the weight of cereal in the i^{th} box.





The probability (under the normal model) that the next weight is in the acceptable range is .78.

It would be better if the distribution was centered at 350 and less variable ($\mu = 350$, smaller σ) !!

Your are working on it, but for now, it is what is it.

At least the process is "under control" so you can work on improving it.

You are about to be audited !!!

They audit you by:

- getting the weights of the cereal in 10 boxes
- computing the average weight from the 10.
- if the average weight is in (330,370) you are ok.

What are your chances??

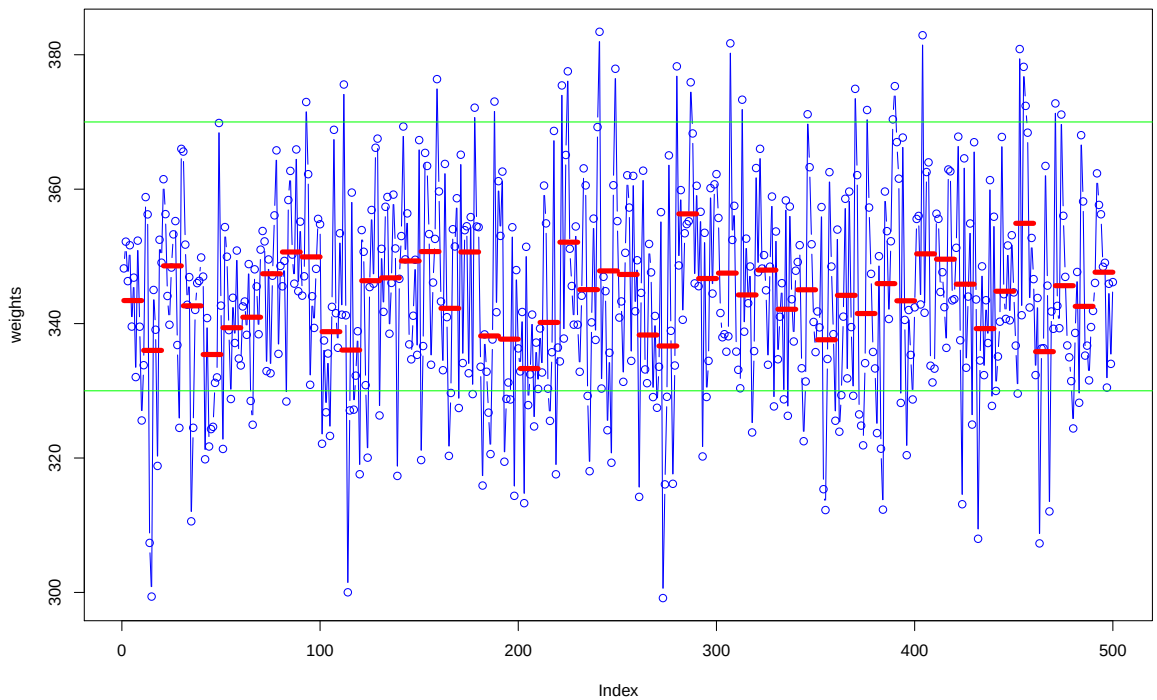
Well, you know how to compute the probability the next weight is in (330,370).

But what about the probability the *average of the next 10* is in the interval???

Is the average more likely to be in the interval than an individual weight?

Hey, why not just take the average of successive batches of 10 (from the 500) and see what they look like??!!

The little (red) bars are at the averages of successive groups of 10 observations.



Phew!! Looks like it is very unlikely that average will be outside the (330,370) range!

You feel a lot better, but there are only 50 means and some are close to the acceptable boundary.

Is there a better way to get a handle on what you really want which is

$$P(330 < \bar{Y} < 370)?$$

Where $\bar{Y}$ is the mean the auditors *will get* when they sample 10.

$\bar{Y}$ is a random variable !!!!

2. The Distribution of the Sample Mean

You find out that:

For,

$$Y_i \sim N(\mu, \sigma^2), \text{ iid},$$

if $\bar{Y}$ is the average of $Y_1, Y_2, \ldots, Y_n$ then,

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

Perfect.

Your model is that the weights are iid $N(344, 15^2)$.

The audit is *about to get* the average of 10 draws.

If $\bar{Y}$ denotes the average of the ten then,

$$\bar{Y} \sim N(344, 15^2/10).$$

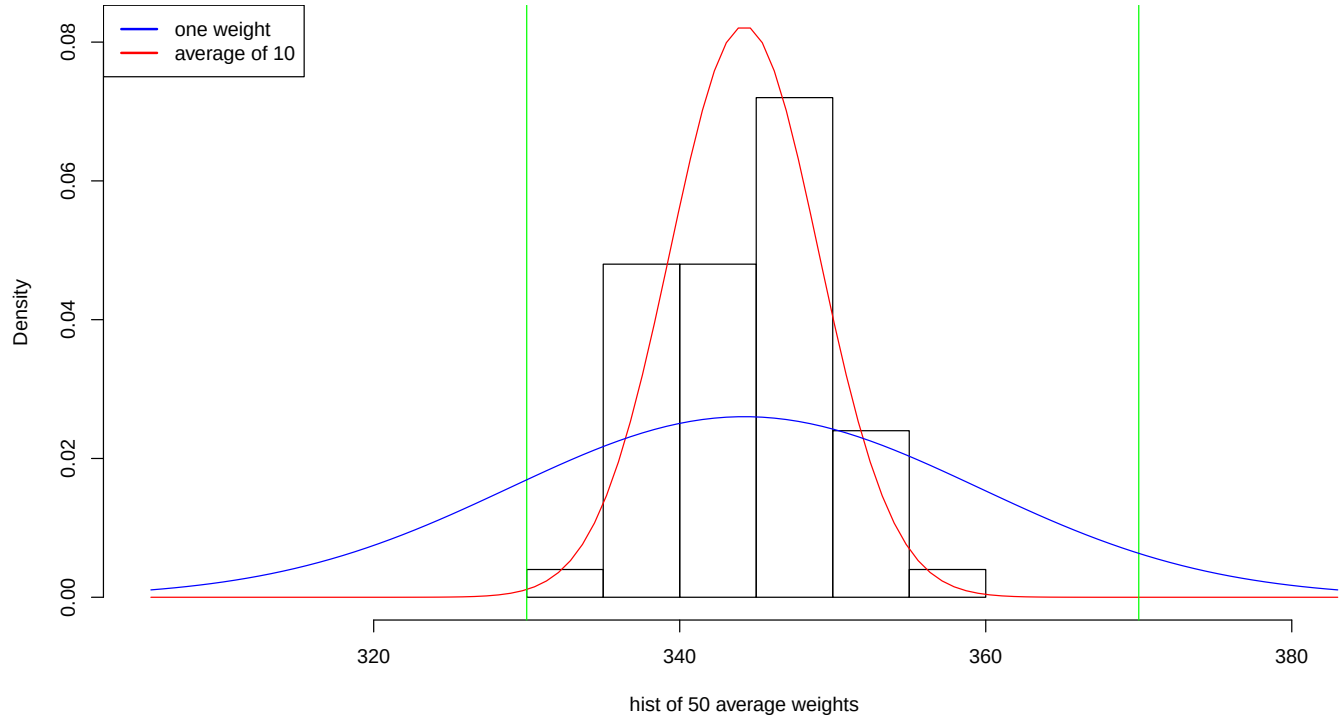
The probability of passing the audit is

$$\begin{aligned} P(330 < \bar{Y} < 370) &= \\ &= F(370) - F(330) \\ &= 1 - 0.000876442 = 0.9991236 \end{aligned}$$

where F is the cdf for $N(344, 15^2/10)$.

$Y \sim N(344, 15^2)$. $\bar{Y} \sim N(344, (\frac{15^2}{10}) = 4.74^2)$.

Histogram is of the 50 means of 10.



Looks reasonable !!!!

What would the normal pdf for the average of 100 look like?

3. Estimating a Normal Mean

The audit scare is over. You can sleep at night.

While the means of 10 vary, they do not vary enough to make it likely you will fail the audit.

Then it occurs to you:

Hey, I've been using the mean of 500 to pick μ for my normal,

how wrong could that be ????!!!!

What does "wrong" mean?

Once you model the weights as

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

what you really need to know is μ, the "true" center of the normal curve, or, the true long run average.

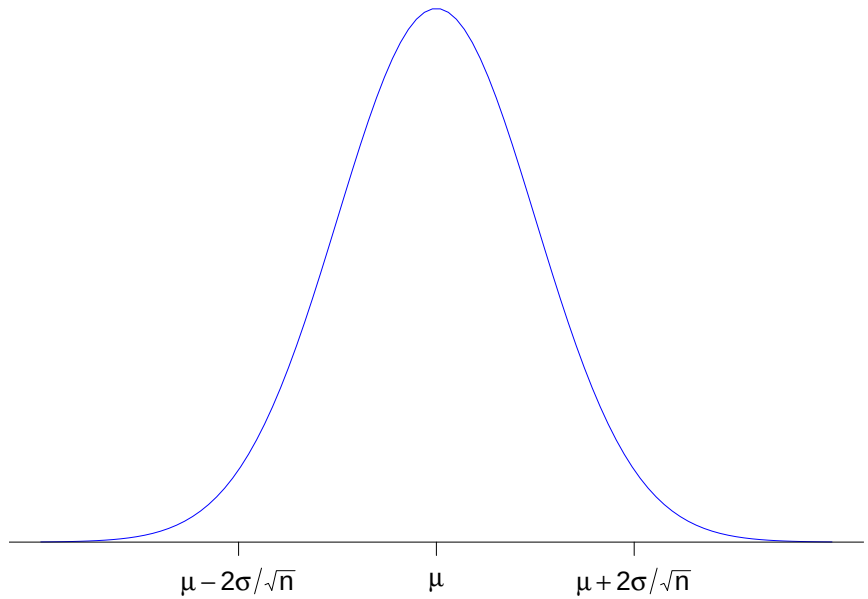
The mean of 344 (using all 500 observations) is still just a guess, or *estimate* of μ.

While we expect the mean of of 500 to be pretty close to the true mean μ (for example, better than the mean of 10) we'd like to know the possible error!!

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?

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

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 $\pm 2\frac{\sigma}{\sqrt{n}}$!!!

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 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!!! Phew again!!!

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

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.

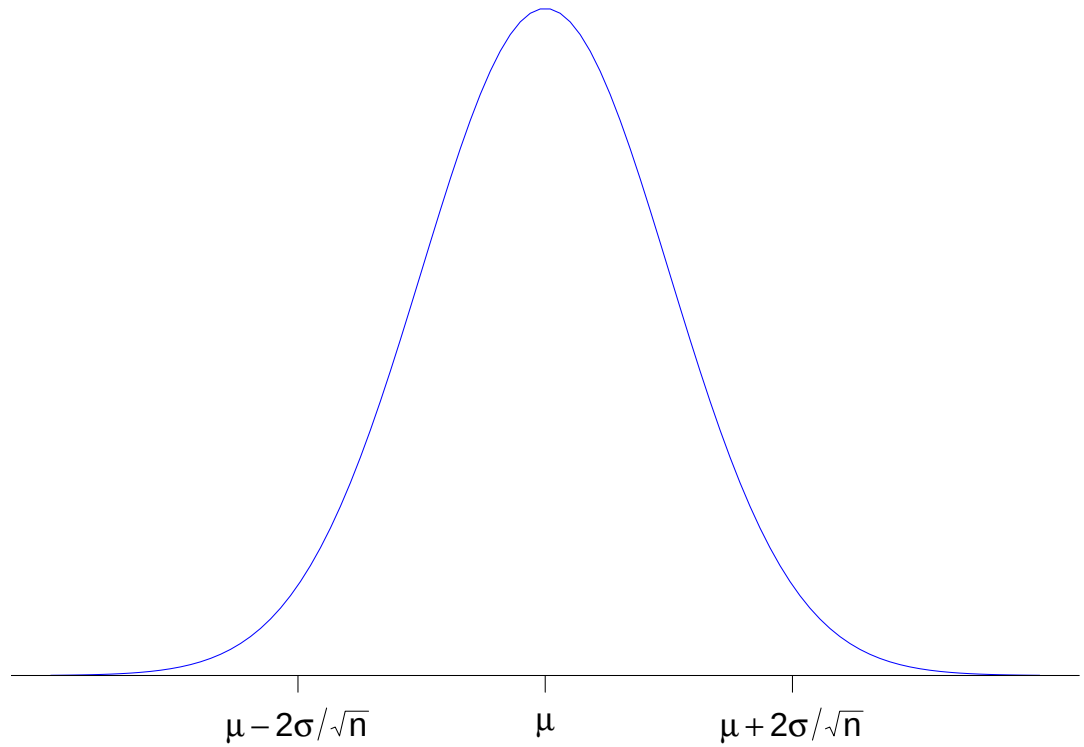
Note:

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

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:

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:

Previously, we modeled the Canadian returns as iid Normal.

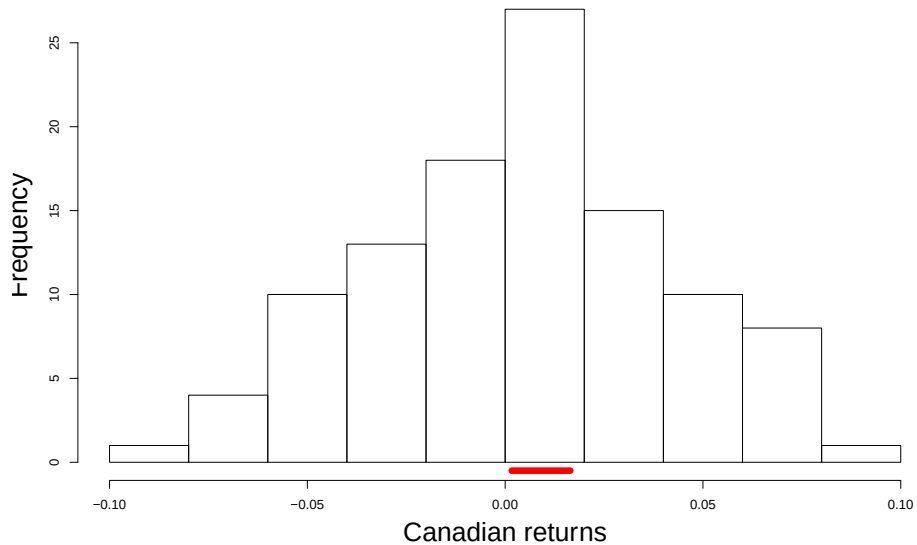
Let's get the 95% confidence interval for μ.

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

ci:

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

$(0.0017, 0.01650)$



Is this a big interval?

Predictive Intervals for IID Normal Data:

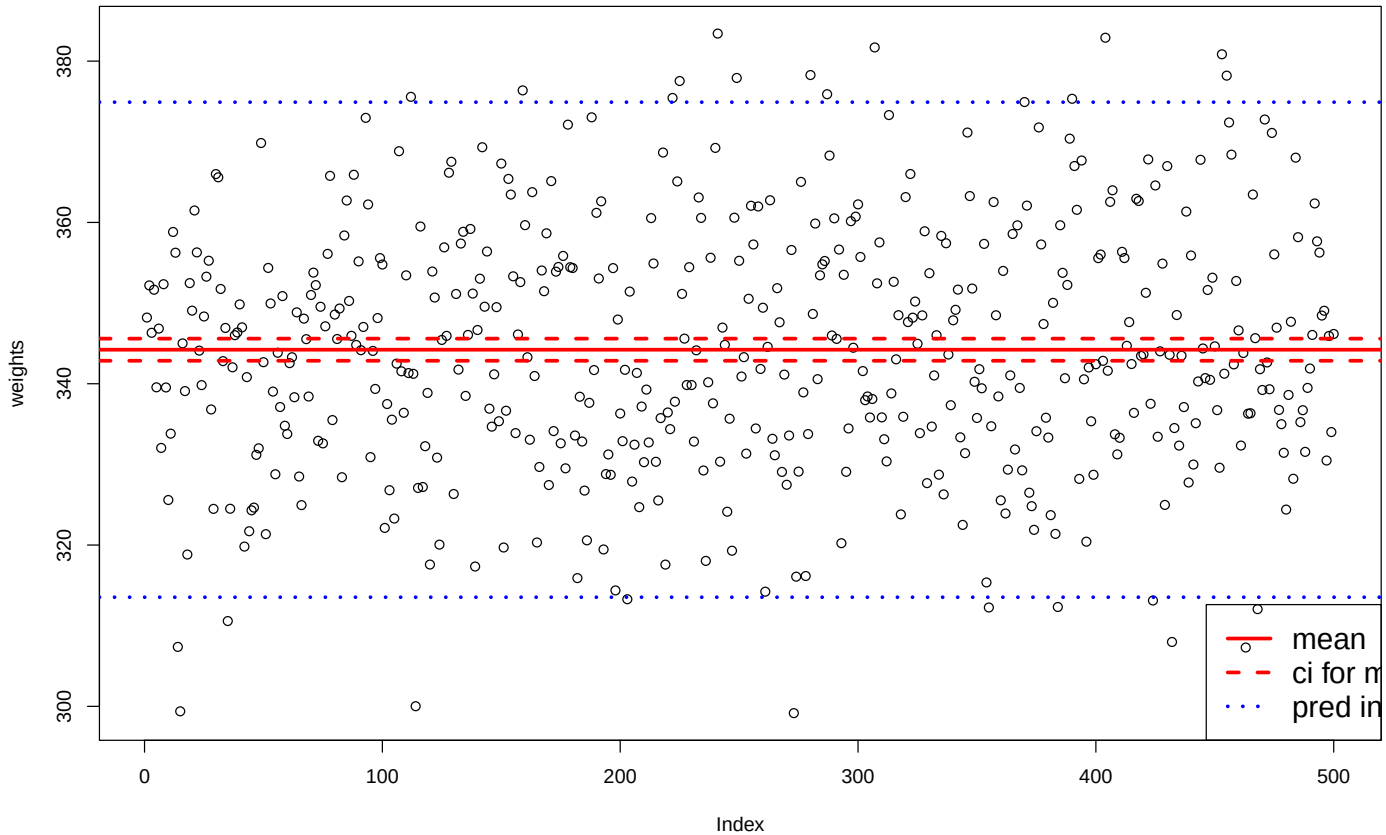
Suppose we would like to predict how much cereal will go in the next box?

For iid normal data,
The *predictive* 95% interval is:

$$\bar{y} \pm 2\, s_y \sqrt{1 + \frac{1}{n}}$$

Note: for n "big" $\bar{y} \approx \mu$, $s_y \approx \sigma$ and $\sqrt{1 + \frac{1}{n}} \approx 1$ so that the interval is like $\bar{y} \pm 2s_y \approx \mu \pm 2\sigma$.

For the 500 cereal observations:



4. 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:

Previously, we modeled defects as iid Bernoulli.

We had $n = 300$ and $\hat{p} = .18$.

Now we can assess the accuracy of this estimate!!!

se: $\sqrt{\frac{.18*.82}{300}} = .0222.$
2*se: .0444
Confidence interval: $.18 \pm .0444 = (0.14, 0.22)$.

BIG!!

Note: Probability next 10 are good is?

If p were .14, we would get: $(1 - .14)^{10} = .22$
If p were .22, we would get: $(1 - .22)^{10} = .08$

Example:

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

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.*

5. The Improved Cereal Process

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

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

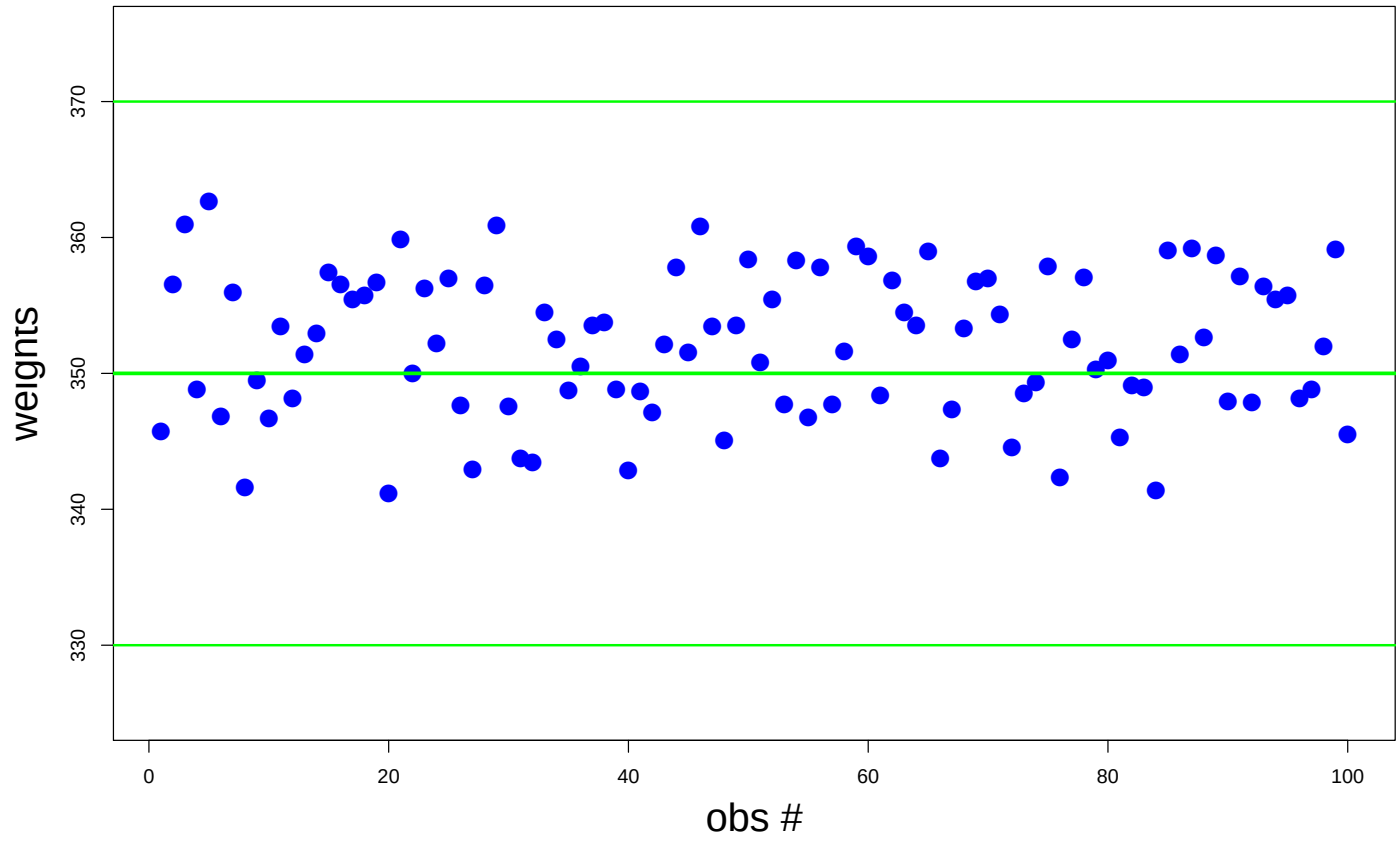
The process was supposed to be centered at 350 and $350 \pm 20 = (330, 370)$ is the range of acceptable weights.

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 of 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.

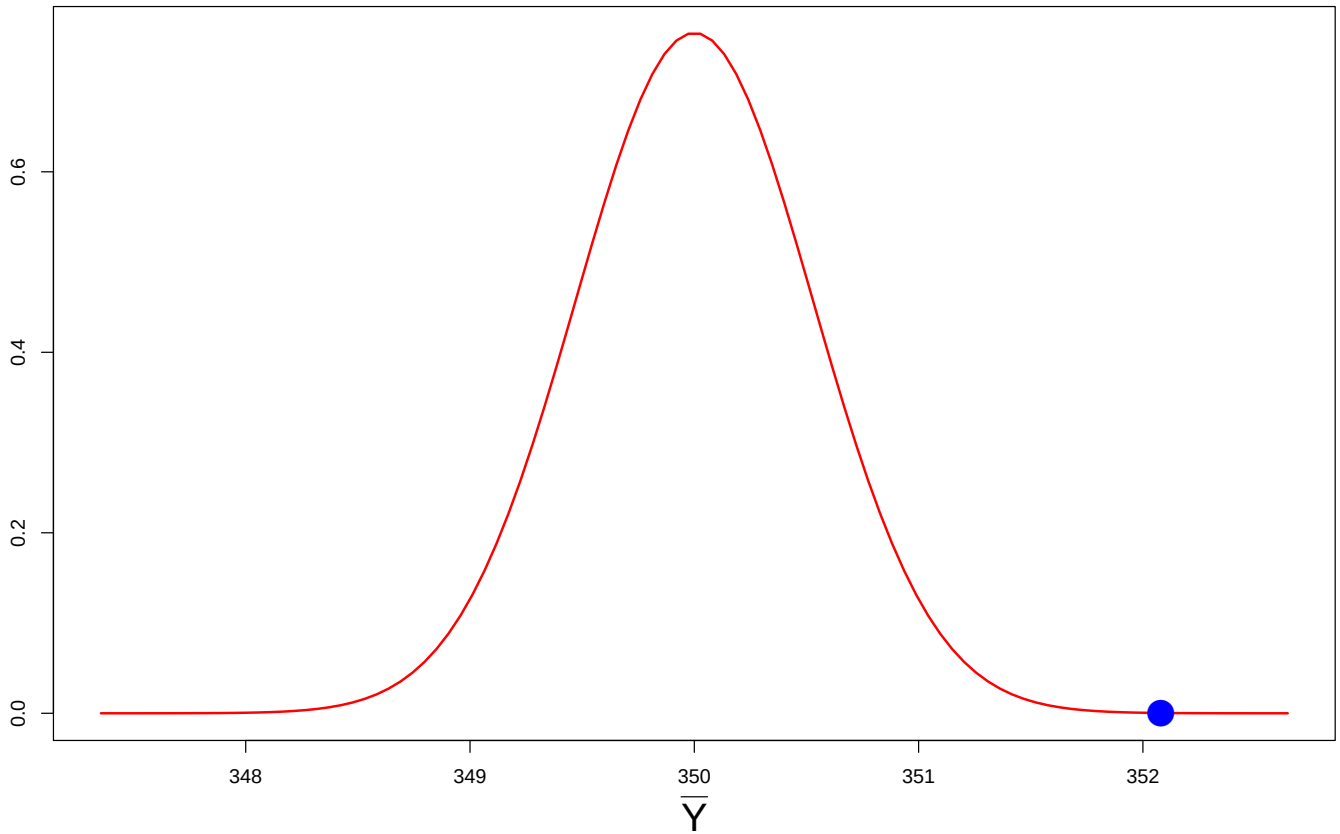
However for n greater than about 20 (sound familiar?) it turns out you can plug in the sample standard deviation without making too much of an error.

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.

6. 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 ($\mu = 350$).

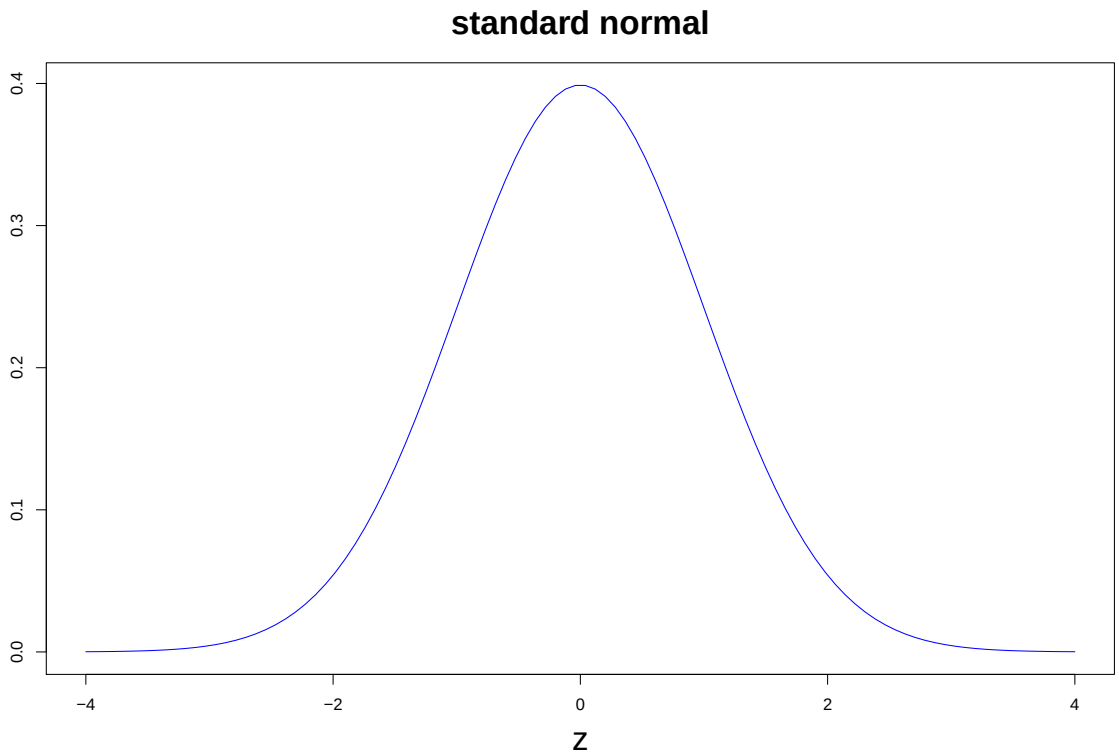
```
> t.test(weights,mu=350)

        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).

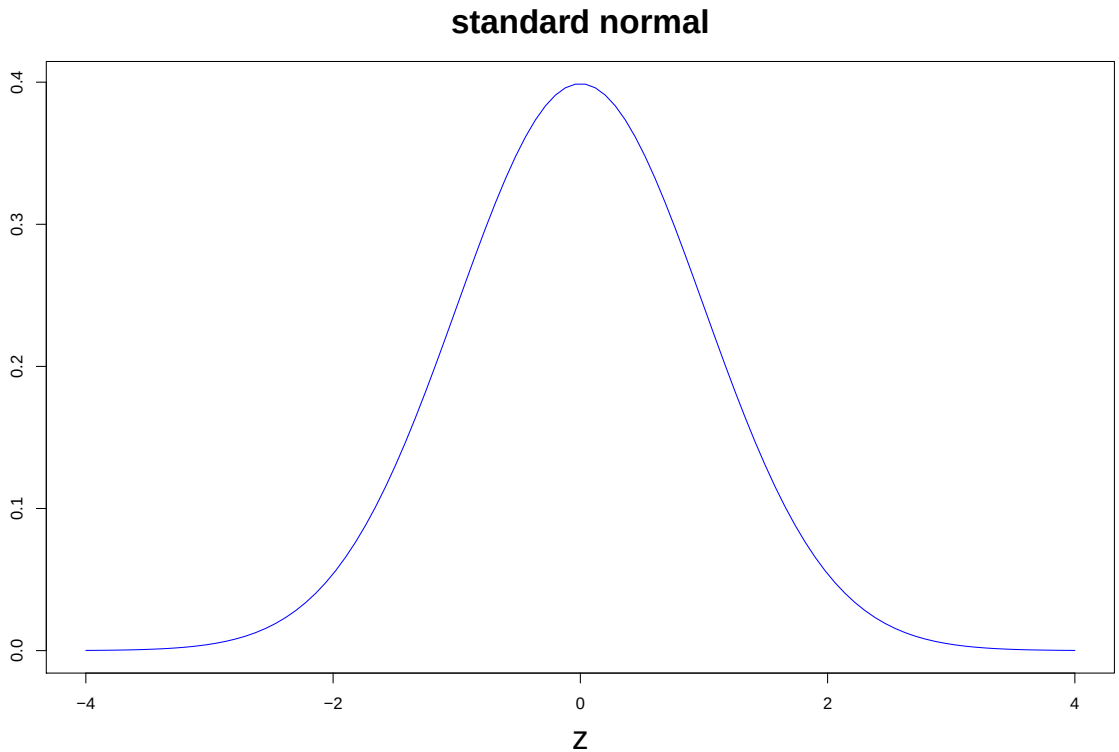
Your assistant says "Ok, maybe it is not perfectly centered, but it looks like we are pretty sure it is darn close!!"

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!!

7. 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.

- t=4, strong evidence against the null.
- t=2.01, some evidence against the null.
- t=1.99, some evidence against the null.
- t=1, no evidence against the null.

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

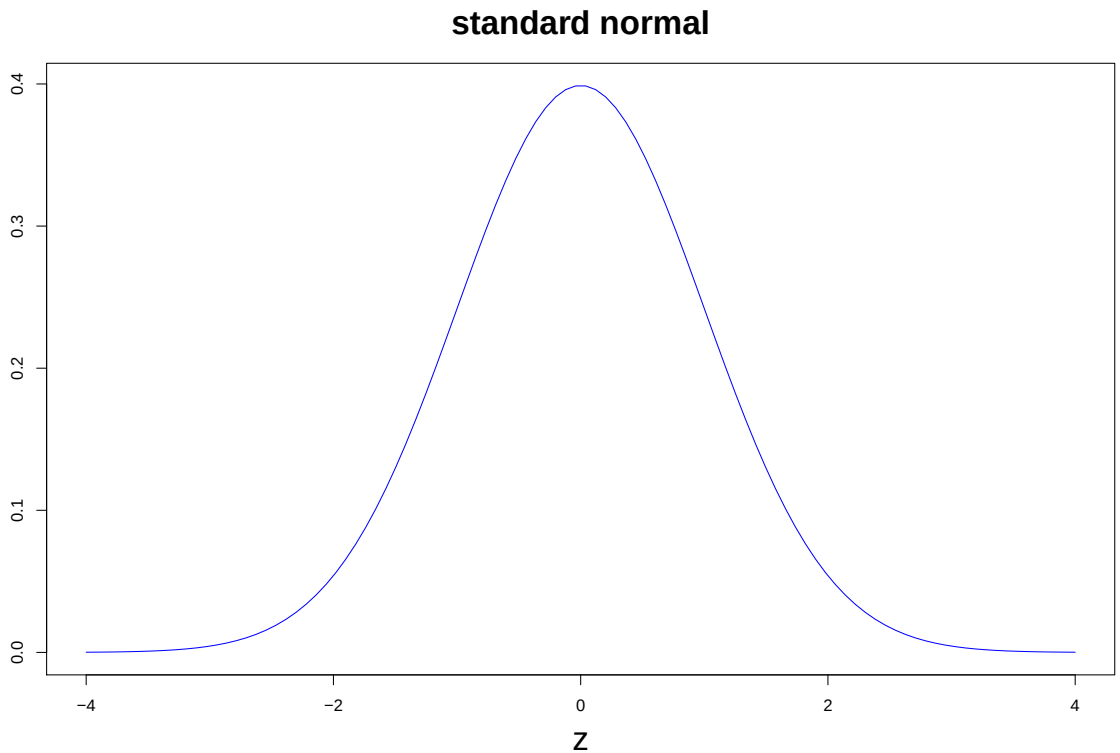
The p-value is the probability of getting a t test-statistic as far out or farther, **if** the null is true.

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.

Example:

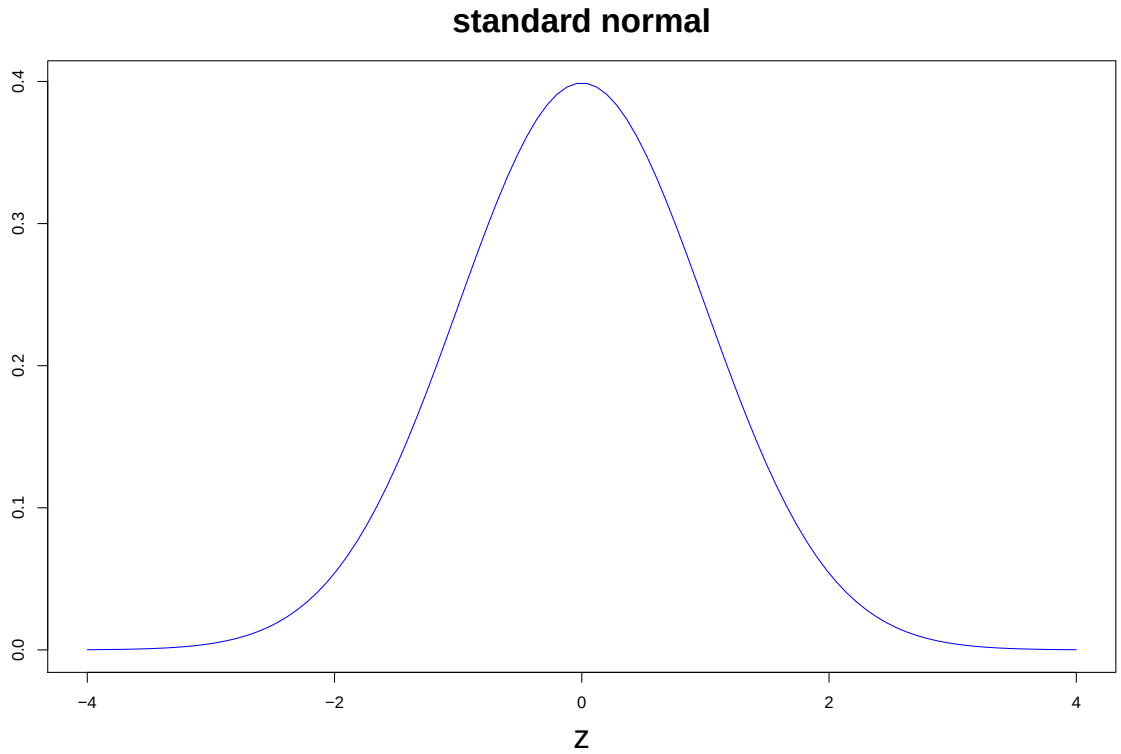
Recall that we tested whether the true mean of the Finnish returns is equal to 0.

```
> 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:

$$2 * F(-1.3131) = .1889.$$

8. 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 ⇒ 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!!