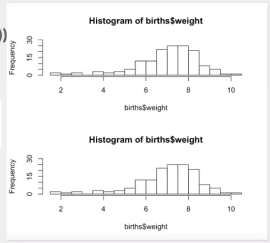
par(mfrow = ...)

The **par() function** with the **mfrow =** argument can be used to visualize two plots in the same window.

par(mfrow = c(# of "rows", # of "columns"))

Specifies that we want two plots, one above the other.

```
> par(mfrow = c(2, 1))
> hist(births$weight, breaks = 25, ylim = c(0, 30))
> hist(births$weight, breaks = 25, ylim = c(0, 30))
```



If we wanted two plots side-by-side, we would use par(mfrow = c(1, 2))

Note: once you set this, it will continue to plot more than one plot in the same window until you change it back to one. You can change it back by using par(mfrow = c(1, 1)).

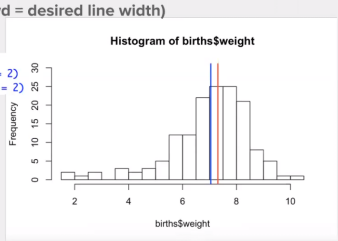
abline()

The **abline()** function can be used to draw a line over a plot, such as a histogram.

abline(v = value, col = "color", lwd = desired line width)

```
hist(births$weight, breaks = 25, ylim = c(0, 30))
abline(v = mean(births$weight), col = "blue", lwd = 2)
abline(v = median(births$weight), col = "red", lwd = 2)
```

In this example, we used abline() to plot a line for the value of the mean on this histogram in blue. Then, we used another abline() to plot the value of the median in red. This allows us to see where the mean and median are (for example) in the distribution.



Just for fun, here are all the colors: http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf

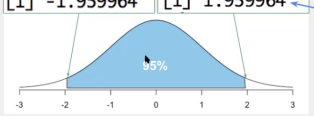
Critical Value and confidence intervals

Confidence should be in decimal form.

95% confidence?

To find lower bound: **mean**(total population *or* sample) - **qnorm**(1 - (1 - *confidence*) / 2) * **standard_error**
To find upper bound: **mean**(total population *or* sample) + **qnorm**(1 - (1 - *confidence*) / 2) * **standard_error**

```
> qnorm(0.025)
[1] -1.959964
> qnorm(0.975)
[1] 1.959964
```



SD sample / N sample size

This number is your **critical value**:
The number of standard errors from the mean within which the given percentage of the data lies.

It is calculated with:
qnorm(1 - (1 - confidence) / 2)

HINT: Look at Lecture 4.2 "Confidence Intervals" slide #8 for examples of calculating the confidence interval.

pnorm()

pnorm(value, mean, sd, lower.tail = TRUE/FALSE)

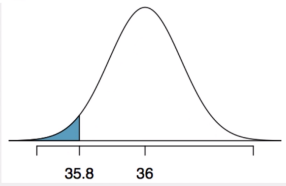
Given a value, what is the probability that another random value is less than or equal to the first, assuming normal distribution.

We can use pnorm to calculate our p-value and the power of the test.

Note: When calculating the p-value from our test statistic, the mean is 0 and the sd is 1 (don't need to use mean and sd arguments)

value | mean | standard error

```
> pnorm(35.8, 36, .11)
[1] 0.03451817
```



Critical Value and confidence intervals

Confidence should be in decimal form.

95% confidence?

0.95

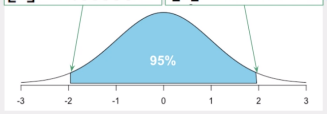
To find lower bound: **mean**(total population *or* sample) - **qnorm**(1 - (1 - *confidence*) / 2) * **standard_error**
To find upper bound: **mean**(total population *or* sample) + **qnorm**(1 - (1 - *confidence*) / 2) * **standard_error**

The confidence interval tells us that 95% (in this case) of our data falls between our lower and upper bounds.

When we calculate a confidence interval with a sample of our total population, what we are often looking for is whether or not the true population mean falls within the sample confidence interval.

What does a confidence interval look like?

```
> qnorm(0.025)
[1] -1.959964
> qnorm(0.975)
[1] 1.959964
```



This number is your **critical value**:
The number of standard errors from the mean within which the given percentage of the data lies.

It is calculated with:
qnorm(1 - (1 - confidence) / 2)

HINT: Look at Lecture 4.2 "Confidence Intervals" slide #8 for examples of calculating the confidence interval.

Power of the Test:

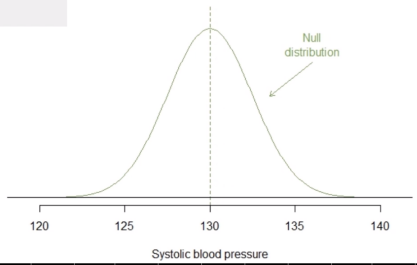
Power level tells you the probability that we would correctly reject the null hypothesis (H_0) given an interest in detecting a certain difference.

	Fail to reject H_0	Reject H_0
H_0 is true	✓	Type 1 Error (α)
H_A is true	Type 2 Error (β)	✓

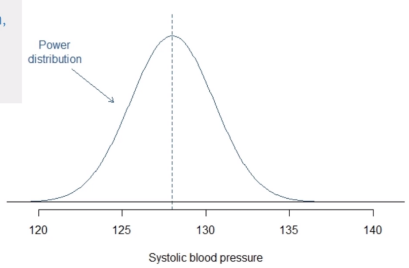
We are looking for this

null hypo always = or ≤ sth

Consider a Null hypothesis:
The mean systolic blood pressure is 130.
Mu = 130

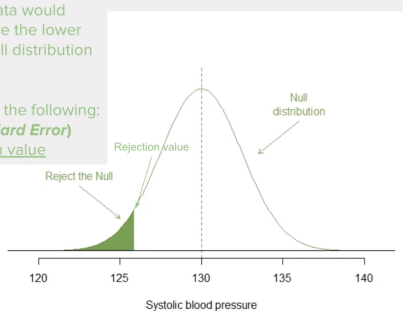


Say we are interested in a systolic blood pressure difference of 2. We want to know, if 128 is the true mean, what is the likelihood of us detecting that the null hypothesis is incorrect?

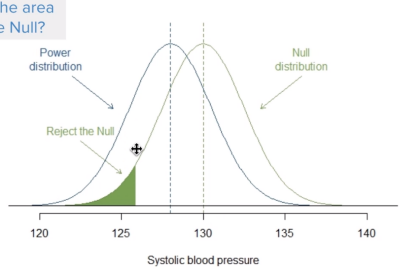


This means that to reject the Null hypothesis, our observed data would need to fall in the area where the lower .05 (alpha) portion of the Null distribution lies

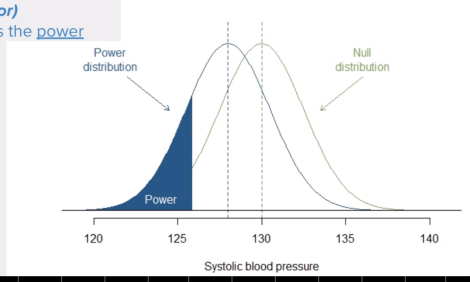
We can find this value using the following:
qnorm(α, *Null Value*, *Standard Error*)
We will call this our rejection value



So the question really becomes:
What proportion of the Power (or sample) distribution falls in the area in which we would reject the Null?



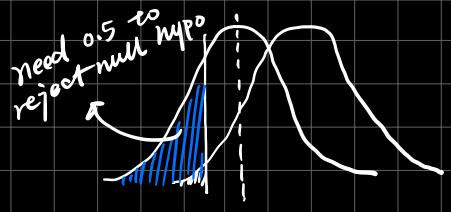
We can find this proportion using the following:
pnorm(*Rejection Value*, *Observed Mean*, *Standard Error)*
The value returned is the power



Power of the Test Continued

qnorm(**alpha**, **null value**, **standardError**) => **rejectionValue**
The rejection value refers to the value where we can reject the Null
pnorm(**rejectionValue**, **observedValue (mean observed)**, **standardError**)
This gives us the proportion of the observed data that would fall beyond the rejection value. (Also known as the **power**)
If the power level is really low, it is likely we would not be satisfied by that level of power because we won't be able to correctly reject the null hypothesis a majority of the time.

If the power level is really high, we would likely be satisfied with that level of power because we would correctly reject the null hypothesis the majority of the time.
Note: alpha is **arbitrary**, it is not dependent on what your confidence interval is. If you aren't given what alpha is, you should assume 0.05 because that is the significance level most commonly used.



Normal Hypothesis Testing: Conditions

Conditions that need to be met in order to conduct a Normal Hypothesis Test:

- Independence
  - Variables must be independent of each other.
  - We can assume independence with random sampling.
- No Skew/Large sample size
  - Your distribution should be normally distributed (without skew) in order to conduct normal hypothesis testing.
  - However, if you have a **large sample size** (which is really what we want anyway) and there is skew in the distribution, it doesn't really matter. You can still conduct normal hypothesis testing. If your sample size is small and you have skew, this testing might not be appropriate because it will likely be affected by the skew.
  - The best case scenario for this type of testing: no skew and large sample size.

by()

The **by()** function performs a function for a numerical variable (y) for each of the categories in categorical variable (x).
by(numericVariable, categoricalVariable, functionDesired)
Value desired could be, for example: 'hist', 'mean', 'min', or 'max'

```
> View(nc)
> by(nc$weight, nc$gender, mean)
nc$gender: female
[1] 6.902883
------------------------------------------------------------
nc$gender: male
[1] 7.301509
```
mean weight of female

t.test()

This function is used to display information on hypothesis testing for a specific variable or variables.

There are two formats we will be using:
For difference of two means--
t.test(*numeric variable*~*categorical variable*, conf.level =)
For single mean--
t.test(*numeric variable*, conf.level =)

The variables that are being input into t.test() are in the form of vectors.
The argument conf.level refers to the level of confidence you want to use for your confidence interval. It is optional and if you do not specify a confidence, R will default to a value of .95

t.test() Examples

The question: Was the average home in Ames, Iowa built in 1971?

"mu =" is another optional argument. It sets the null value. It defaults to 0.

```
> t.test(ames$Year.Built, mu=1971)

        One Sample t-test

data:  ames$Year.Built
t = 0.63769, df = 2929, p-value = 0.5237
alternative hypothesis: true mean is not equal to 1971
95 percent confidence interval:
 1970.261 1972.452
sample estimates:
mean of x
 1971.356
```

We are given the p-value, alternative hypothesis, and confidence interval.
The null hypothesis is implied.

no, within

Based on the information shown here, would you be surprised if the true mean was 1972?
Would you be able to **reject** the null hypothesis? no, p > 0.05 fail to reject

t.test() Examples

The question: Do homes on gravel roads differ in overall quality from homes on paved roads?

We use "~" when we want to simulate hypothesis testing using a difference of means

```
> t.test(ames$Overall.Qual~ames$Street)
        Welch Two Sample t-test

data:  ames$Overall.Qual by ames$Street
t = -4.2111, df = 11.104, p-value = 0.001428
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.437491 -0.765388
sample estimates:
mean in group Grvl mean in group Pave
          4.500000           6.101439
```

numerical quality ~ categorical

We are given the p-value, alternative hypothesis, and confidence interval.
The null hypothesis is implied.

There is a diff in qual (sanity check)

not included

Based on the information shown here, would you be surprised if the true difference of means was -2.5?
Would you be able to **reject** the null hypothesis? yes, p < 0.05