A significance test involves the following steps

1. Specify null and alternative hypotheses, and a test statistic to discriminate between the two
2. Compute the observed value of the test statistic
3. Compute the p-value which is the probability the test statistic will be more extreme than the observed one as defined by the alternative hypothesis
4. Make a decision to accept or reject the null hypothesis based on the significance level and the p-value.

In `R` to perform a significance test you need to identify the following:

1. The type of test (Test of proportion? Test of mean?)
2. The null hypothesis (Either a specification of p or μ)
3. The alternative as `less`, `greater` or `two.sided`

0.1 Test of proportion

For example, suppose a presidential approval rating was 50%. A new poll is taken and 575 of 1,200 respond favorably. Does this indicated a smaller approval rating?

This is a test of proportion with hypotheses

$$H_0 : p = 0.5, \quad H_A : p < 0.5$$

Or an alternative of `less`.

The p-value is returned by the `prop.test()` function which is used as follows

```
> prop.test(575, 1200, p = 0.5, alternative = "less")

        1-sample proportions test with continuity correction

data:  575 out of 1200, null probability 0.5
X-squared = 2.001, df = 1, p-value = 0.0786
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000 0.5033
sample estimates:
     p
0.4792
```

Notice the p-value is calculated as 0.0786. We would accept the null hypothesis at the 0.05 significance level.

0.2 Test of mean

To test for the mean is no different than above, except instead of summarizing the data we need all of the data.

For example, consumer reports tests the braking distance of a HUMMER. GM reports that it can stop from a given speed with a 150 foot average. The test finds values of 165, 158, 174 and 160. Test the following

$$H_0 : \mu = 150, \quad H_A : \mu > 150$$

This is done by entering the data and then using `t.test()`

```
> distances = c(165, 158, 174, 160)
> t.test(distances, mu = 150, alternative = "greater")

        One Sample t-test

data:  distances
t = 3.994, df = 3, p-value = 0.01406
alternative hypothesis: true mean is greater than 150
95 percent confidence interval:
 155.9   Inf
sample estimates:
mean of x
    164.2
```

The p-value of 0.01406 would lead us to reject at the 0.05 significance level.

0.3 Watch the assumptions

In order to make statistical inference, an underlying probability model must be correct. For the test of proportions, the assumptions are that n is large and the sample is a random sample from the target population.

For the test of means, the assumptions are that the data are a random sample from a normally distributed population with unknown mean μ (with an assumed value under H_0) and unknown variance σ^2. You should check these before applying the tests.

To check for normality, we can do one of the following. Assume the data is stored in `dataset`

- A quantile-normal plot using `qqnorm(dataset)`. The graph should be more or less along a line
- A boxplot using `boxplot(dataset)`. The boxplot should be more or less symmetric, and any tails should not be too long.

- A histogram. The following commands will draw a reference normal distribution for you.

```
> hist(scale(dataset), prob = TRUE)
> curve(dnorm(x), add = TRUE)
```

  The density should be a decent fit to the histogram.

0.4 Exercises

For each of the following, find the p-value. Additionally indicate if you would accept or reject at the 0.075 significance level.

 Question 1: After NY state implemented the no-handheld-cell-phone-while-driving law, 95% of drivers were in compliance. Has this changed for the worse? A sample of 175 drivers using cell-phones, found 150 in compliance.

 Question 2: The poverty rate in 2000 was found by the census to be 11.3 percent. To see if the rate has increased, in the year 2001 a random sample of 50,000 was conducted and the sample rate was found to be 11.7 percent. Has the rate increased? What is the target population of the random sample?

 Question 3: A *New York Times* article on screw tops replacing corks for wine bottles had a wine-producers claim that 5% of corks fail to protect the wine. The cork-producing industry, claims it is much lower. What do we think? In Professor Verzani's experience, he has consumed, say, 250 bottle of wine and can recall only 10 times when the wine was off due to faulty corking. Can this data be used to perform a significance test of the above? Is not why not? If so, state the hyptheses, and compute the p-value.

 Question 4: A popular add claims that 4 out of 5 dentists surveyed prefer trident. To test this, 100 dentists are surveyed and only 70 recommend chewing Trident. Is this data consistent with the claim or indicative that it is too high?

 Question 5: A study of a weightloss drug finds that a cohort of 10 patients lost the following amounts

```
10  12  6 6 -2 8 9 8 7 10
```

Without the weight loss drug a person is expected to average six pounds of weight loss. Perform a significance test of

$$H_0 : \mu = 6, \quad H_A : \mu > 6$$

based on this data. How would you check that the data is normally distributed? Why is this important?

0.4.1 Built-in data sets

In R there are several built-in data sets. These can be univariate or multivariate. A data set is loaded into your session using the command `data()`, as in `data(dataset.name)`. For most all data sets, this adds a variable `dataset.name` to your environment.

If the data set is univariate, typing `dataset.name` will show all the data. If the data is multivariate, typing `edit(dataset.name)` will show the data. Close the spreadsheet window to continue, otherwise you won't get the prompt back. Multivariate data sets have columns which are named that contain the variables. These variable names can be conveniently "attached" prior to their usage using the command `attach(dataset.name)`. Now you can use the variables by their column names.

 Question 6: The data set `nhtemp` contains mean temperatures in New Haven. Perform a significance test of

$$H_0 : \mu = 50, \quad H_A : \mu \neq 50$$

for this data. (You need to load it in with the command `data(nhtemp)`.

Do you accept or reject? Is the data normally distributed? How can you tell? Do you need to know that for this problem? Why or why not?

 Question 7: The data set `mtcars` contains data on cars from 1974. Have cars changed since then? (Not mine perhaps).

First load the data set and attach the names

```
> data(mtcars)
> attach(mtcars)
```

1. Do a significance test to see if the mean weight is 3 thousand pounds The variable `wt` contains the data. Use a two-sided alternative.

2. Perform a significance test to see if the mean horsepower is 225. Use a two-sided alternative.

 Question 8: The data set `survey` in the MASS package can be loaded using

```
> data(survey, package = "MASS")
```

Attach the data frame.

1. The variable `Pulse` contains information about the pulse of the students. Perform a significance test of the hypotheses

$$H_0 : \mu = 72, \quad H_A : \mu > 72$$

   What is the p-value? Is the data appropriate for the t-test? Why or why not?

2. The variable `Wr.Hnd` is a measurement of the span of the writing hand. Perform a significance test of the hypotheses

$$H_0 : \mu = 18.5, \quad H_A : \mu \neq 18.5$$

   What is the p-value? Is the data appropriate for the t-test? Why or why not?

 Question 9: The data set `galaxies` in the MASS package is loaded with

```
> data(galaxies, package = "MASS")
```

These numbers measure velocities of galaxies.

If appropriate, use a t-test to perform a significance test of

$$H_0 : \mu = 20,500, \quad H_A : \mu \neq 20,500$$

What is the p-value? Is the data appropriate for the t-test? Why or why not?

 Question 10: The built-in data set `trees` contains three measurements: `Girth`, `Height`, and `Volume`.

For each question, find the p-value using a t-test is appropriate. If not, explain why the t-test is not appropriate.

1. For the variable `Girth` test the hypotheses

$$H_0 : \mu = 13, H_A : \mu > 13$$

2. For the variable `Height` test the hypotheses

$$H_0 : \mu = 75, H_A : \mu \neq 75$$

3. For the variable `Volume` test the hypotheses

$$H_0 : \mu = 33, H_A : \mu < 33$$