Hypothesis Testing

In the previous set of lectures, we made some casual judgements as to whether a basketball player's play fitted the random model or not by looking at the value of a test statistic in the data, namely the longest run of baskets in the data. We knew the expected value and standard deviation of the test statistic which depended on the sample size and the probability of the player making a basket on each shot. If the longest run of baskets in the data was too long or two short, then we decided that the probability that that could happen due to randomness was small enough that the probability of success on each trial did not remain constant throughout. In this section we will make this method of testing a hypothesis more rigorous. We will give a level of confidence for our results and a measure of the strength of our evidence (the p-value).

The Elements of Hypothesis Testing

Rare Event Concept We have seen that for any set of data, observing data that is four or more standard deviations away from the mean is very very rare and for data with a normal distribution observing data that is more than 2 standard deviations from the mean is rare enough that an observation of such data might lead us to change our beliefs about the underlying distribution of a population.

For Example If someone tells me that they flipped a coin 100 times and the longest run of heads in the outcome had length 50. I know that for such an experiment the longest run of heads has an expected value of $\mu = \frac{-\ln((50)}{\ln(1/2)} \approx 5.64$ and standard deviation $\sigma = \frac{-\pi}{\sqrt{6}\ln(1/2)} \approx 1.85$. This would mean that the observed value of the statistic "the longest run of heads" in the data has a z-score of $z = \frac{50 - 5.64}{1.85} \approx 23.98$ and is roughly 23.98 standard deviations above the mean. Although possible, this event would be so rare for a fair coin flipped 100 times, it would lead me to reject the hypothesis that the data was the result of flipping a fair coin 100 times.

This reasoning is the essence of Hypothesis testing. Since this is such a simple, yet powerful tool, we make the method more precise so that we can apply it to any situation where we have knowledge of the distribution of a sample statistic. This type of Statistical inference is called **Hypothesis Testing**

Example 1: How Predictable is the Notre Dame Offense on third down with 5 to go? In the 2013 football season, the Notre Dame football team made 87 plays on third down with 5 yards or less to go to make first down. Below we list the sequence of play types they made in those situations in the order in which they happened. Is this the result of random choices for the play

(we make no assumption about probability here)?

RUSH RUSH PASS PASS PASS RUSH PASS RUSH RUSH RUSH PASS RUSH PASS PASS RUSH PASS PASS PASS PASS RUSH PASS RUSH PASS RUSH RUSH RUSH PASS PASS RUSH RUSH RUSH RUSH PASS PASS PASS RUSH PASS RUSH RUSH PASS RUSH RUSH RUSH PASS PASS RUSH PASS PASS RUSH RUSH RUSH PASS RUSH PASS PASS RUSH RUSH PASS PASS RUSH PASS PASS RUSH RUSH RUSH PASS RUSH PASS PASS RUSH PASS RUSH PASS PASS RUSH PASS RUSH PASS PASS RUSH PASS PASS RUSH PASS RUSH PASS PASS RUSH PASS RUSH PASS PASS RUSH PASS PASS RUSH PASS RUSH PASS PASS PASS

Example 2: A Hypothesis about a Population Proportion: Many people believe that the London Olympics is cursed, since a seemingly large number (18) of those who participated have had died in the four years since the olympics. `http://www.bbc.com/news/magazine-36055238`. In other words they think that this number is large enough to make them reject the hypothesis that the death rate among the olympians who participated in London is the same as that for the population in general in favor of one that says it is significantly higher. We will assume that the sample comes from the general population. Here our test statistic will be the proportion of olympians who participated in the London games who have passed away since. This proportion varies from sample to sample, however we know that since the sample size is large that its distribution is approximately normal with mean equal to the population proportion and corresponding standard deviation. We can then look at the z-score of the proportion from the sample and if it is too large (or too small) we will reject the hypothesis that this is a typical sample from the general population and thus it must come from a different population. If the evidence leads us believe the olympians are from a different population, we can put forward some further hypotheses (supernatural or otherwise) as to what might be causing the elevated risk for these athletes. Of course, any further hypothesis as to the risk factor involved would require testing.

Example 3: A Hypothesis about a population mean (A hypothetical example) (a) A football team had (supposedly) inflated 11 footballs before a game to 12.5 psi (pounds per square inch), the minimum inflation required by the NFL. At half time, the footballs were measured with the same gauge and the following psi measurements were recorded:

$$11.8, 11.2, 11.5, 11.0, 11.45, 11.95, 12.3, 11.55, 11.35, 10.9, 11.35$$

Suppose the average drop in pressure recorded for such footballs under similar atmospheric and environmental conditions for this type of gauge is 0.999, would you suspect tampering with the pressure of the footballs?

(b) If on the other hand, the measurements at half time were

$$11.5, 10.85, 11.15, 10.7, 11.1, 11.6, 11.85, 11.1, 10.95, 10.5, 10.9$$

would you suspect that the balls were tampered with?

Hypothesis Testing : Stating your Hypotheses

In formulating your theory for a test of hypothesis, it is important to state what the prevailing viewpoint is and if your data leads you to reject the prevailing viewpoint, what the alternative viewpoint will be, i.e. you should state a null and an alternative hypothesis.

Definition The **Null Hypothesis,** denoted H_0, is the prevailing viewpoint. It represents the status quo to the party performing the experiment. It is the hypothesis that will be supported unless the data provides convincing evidence that it is false.

Definition The **Alternative Hypothesis,** (or research Hypothesis) denoted H_1, is the hypothesis which will be accepted if the data provides convincing evidence of its truth.

Examples In Example 1 above, The Null Hypothesis can be stated as:

H_0: The experiment is a result of randomly choosing each play (using a spinner or die or some such device). or
H_0: The number of runs of PASS's and RUSH's in the data is consistent with a random choice on each play.

The Alternative Hypothesis can be stated as:
H_1: The experiment is not consistent with each play being chosen randomly or the number of runs in the data is too high or too low to be consistent with random choices.

In Example 2 above, The Null Hypothesis can be stated as:

H_0: The population of olympians who competed at the London games comes from a population of young people with an average age of 26 where the mortality rate is approximately $p = 0.00066$.
or
H_0: $p = 0.00066$ where p represents the mortality rate per annum for the population of London olympians.

The Alternative Hypothesis can be stated as:
H_1: population of olympians who competed at the London games comes from a population where $p > 0.00066$, this is called a one sided alternative hypothesis since we will reject the hypothesis only if the sample proportion is too large.

In Example 3 above, The Null Hypothesis can be stated as:

H_0: The footballs measured at half time had decreased pressure due to the normal environmental factors such as ambient temperature and gauge accuracy.
or
H_0: $\mu = 0.999$ where μ represents the average rate of deflation .

The Alternative Hypothesis can be stated as:
H_1: The average deflation was higher that would be expected under normal conditions, or $\mu > 0.999$.

The form of the Alternative Hypothesis depends on the conditions under which you are willing to reject the Null hypothesis (perhaps incurring a cost). Sometimes you are only willing to reject the Null hypothesis if the numbers from the data are too large (resp. too small), in which case your Alternative hypothesis will be one sided to the right or (resp. one sided to the left) as for Examples 2 and 3. Sometimes the Alternative hypothesis is two sided as for Example 1 . Also sometimes the Null hypothesis covers an interval.

Test Statistic We will decide, as a result of our research, that there is or is not sufficient evidence to reject the null hypothesis. Before we can decide, we must make a **decision rule**. Our null and alternative hypotheses are usually based on the value of a population parameter. We usually use the value of the corresponding sample statistic with a known distribution(under the assumption that the null hypothesis is true) to decide whether to reject the Null Hypothesis or not.

A Test Statistic is a sample statistic used to decide whether to reject the null hypothesis or not.

Example 3 Our test statistic will be the $T = \frac{\bar{x}-0.999}{s/\sqrt{11}}$ where $\bar{x}$ is average of the decrease in pressure (12.5 minus the observed pressure at half time) for all 11 balls in the sample and s is the sample standard deviation. If the null hypothesis H_0 is true and the average decrease due to environmental factors is $\mu = 0.999$, then the distribution of this test statistic T is approximately a t distribution with 10 degrees of freedom (because it is a small sample). Our decision rule will be of the form: Reject H_0 if the observed value of T is too big. Exactly how big will depend on the level of significance we will want for our results.

Reliability of the results: Level of Significance Note that when we reject the hypothesis that the coin is unbiased in our introductory example, we might be making an error. There is a small chance that the coin is unbiased and the results are just and instance of a rare event.

Definition If we reject the null hypothesis when it is true we might be making

a **Type I Error**.

Example: In example 3, if our decision rule is: "Reject H_0 if $|T| = \frac{\bar{x}-0.999}{s/\sqrt{11}} > 1.812461$", and we end up rejecting the null hypothesis, then the probability of making a Type I error is approximately 0.05. (Here 1.812461 is the 95th percentile of the t distribution with 10 degrees of freedom, found using the command `qt(.95,10)` in R). In this case we say that our test has a 95% level of confidence.

Note that we have control over the probability
of a type I error when we are formulating our decision rule.

If I wish to reduce the chances of making a Type I error to 0.01 (increase the level of confidence to 99%) we would change our decision rule to :
"Reject H_0 if $|T| = \frac{\bar{x}-0.999}{s/\sqrt{11}} > 2.763769$" where `qt(0.99,df=10) = 2.763769`.

The probability of type I error for our test is usually denoted by α and is called the **level of significance** of the test.

For our initial decision rule above, the level of significance of our test is $\alpha = 0.05$. For the decision rule "reject H_0 if $|z| > 2.763769$" the level of significance is $\alpha = 0.01$.

The Elements of a Hypothesis Test

In general, the method of hypothesis testing follows a similar path to the one above.

1. We identify the Null and Alternative hypotheses.

2. We identify a test statistic and its probability distribution (For us this will just be a Normal Distribution).

3. We formulate our decision rule for the test statistic (this depends on what level of significance we want for our test).

4. We calculate the value of our test statistic from the data collected and decide whether or not to reject H_0.

5. We state our conclusions, (reject H_0 or not). We include a measure of reliability, (level of significance) and we state any assumption we made in the process of testing (In the above example we reject H_0 at a 5% level of significance.)

Type I and Type II Error

	H_0 is true	H_0 is false
reject H_0	Type I error	correct
fail to reject H_0	correct	Type II error

There are two types of error we can make when we decide whether to accept the Null Hypothesis or reject the Null hypothesis.

TYPE I Error = Reject H_0 when H_0 is true.

The Level of Significance of the test, denoted by α is the probability that we make a type I error.

α = The probability that we reject H_0 when H_0 is in fact true.

TYPE II Error = Fail to reject H_0 when H_0 is false.

We let β denote the probability that we make a type II error. The value of β is usually unknown, because the alternative hypothesis is usually less specific than the null hypothesis.

β = the probability that we fail to reject the null hypothesis when the null hypothesis is in fact false.

p-Values The **p-value** of our data gives us the probability that we will observe the value of the test statistic we get from our data or something more extreme (calculated from the probability distribution of our test statistic).

Important Decision rule when testing with R If the p-value of our data is less than the desired level of significance, we reject the null hypothesis. If the p-value of our data is not less than the desired level of significance, we do not reject the null hypothesis. When running our test in R, the test will return the p value of the data upon which we will base our decision.

Hypothesis test for a sample mean

In Summary, our hypothesis test for a sample mean involves the following **steps**:

	One Tailed		
	Left Tailed	**Right Tailed**	**Two Tailed Test**
Null Hypothesis	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
Alternative Hypothesis	$H_a : \mu < \mu_0$	$H_a : \mu > \mu_0$	$H_a : \mu \neq \mu_0$
Level of Significance	α	α	α
Level of Confidence	$(1-\alpha)100\%$	$(1-\alpha)100\%$	$(1-\alpha)100\%$
Test Statistic :	$t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$	$t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$	$t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$
Rejection Region	$t < -t_\alpha$	$t > t_\alpha$	$t < -t_{\frac{\alpha}{2}}$ or $t > t_{\frac{\alpha}{2}}$
	where t_α is chosen so that $P(t > t_\alpha) = \alpha$		where $t_{\frac{\alpha}{2}}$ is chosen so that $P(t > t_{\frac{\alpha}{2}} = \frac{\alpha}{2})$

where s denotes the sample standard deviation and the t-distribution used has $n-1$ degrees of freedom.

Example 3(a): Lets test our Hypothesis with a 95% level of confidence:

- **Identify Null and Alternative Hypotheses:** H_0: $\mu = 0.999$ where μ represents the average rate of deflation .

  H_1: The average deflation was higher that would be expected under normal conditions, or $\mu > 0.999$.

- **Test Statistic and its distribution:** $T = \frac{\bar{x}-0.999}{s/\sqrt{11}}$, t distribution with 10 degrees of freedom.

- **Decision rule** `qt(95/100, df=10)` $= 1.812461$.

```
> qt(95/100, df=10)

[1] 1.812461
```

  Reject H_0 if $T = \frac{\bar{x}-0.999}{s/\sqrt{11}} > 1.812461$

- **Calculate value of test statistic and decide** (we use the first set of data here).

$$T = \frac{\bar{x} - 0.999}{s/\sqrt{11}} = \frac{1.013636 - 0.999}{0.4099335/\sqrt{11}} = 0.1184176.$$

```
> x<-rep(12.5,11)
> p<-c(11.8,11.2,11.5,11,11.45,11.95,12.3,11.55,11.35,10.9,11.35)
> y<-x-p
> m<-mean(y)
> m

[1] 1.013636

> s<-sd(y)
> s

[1] 0.4099335

> t<-(m - 0.999)/(s/sqrt(11))
> t

[1] 0.1184176
```

In this case we do not reject the Null Hypothesis at a 95% level of significance. the p value of the data is $P(T > 0.1184176) = 0.4539358$.

```
> pvalue<-1-pt(t,df=10)
> pvalue

[1] 0.4540409
```

- In this case we do not have sufficient evidence to reject the Null Hypothesis.

Example 3(b) On the other hand if the readings at half time were given by the second set of data, our test statistic would be $T = 3.230324$ and we would reject the null hypothesis with a 95% level of confidence. The pvalue of the data here is 0.004005306.

```
> p1<-c(11.5,10.85,11.15,10.7,11.1,11.6,11.85,11.1,10.95,10.5,10.9)
> y1<-x-p1
> m1<-mean(y1)
> m1

[1] 1.390909

> s1<-sd(y1)
> s1

[1] 0.4023793
```

```
> t1<-(m1 - 0.999)/(s1/sqrt(11))
> t1

[1] 3.230324

> pvalue1<-1-pt(t1,df=10)
> pvalue1

[1] 0.004508002
```

Running a t-test in R: R has a built in function to run the above test called `t.test()`. The arguments are the name of the data set, mu=population mean if the null hypothesis were true and alternative = "greater", "less" or "two.sided".

Example 3(a) :

```
> x<-rep(12.5,11)
> p<-c(11.8,11.2,11.5,11,11.45,11.95,12.3,11.55,11.35,10.9,11.35)
> t.test(x-p, mu= 0.999, alternative="greater")

        One Sample t-test

data:  x - p
t = 0.11842, df = 10, p-value = 0.454
alternative hypothesis: true mean is greater than 0.999
95 percent confidence interval:
 0.7896169       Inf
sample estimates:
mean of x
 1.013636
```

Example 3(b) :

```
> x<-rep(12.5,11)
> p1<-c(11.5,10.85,11.15,10.7,11.1,11.6,11.85,11.1,10.95,10.5,10.9)
> t.test(x-p1, mu= 0.999, alternative="greater")

        One Sample t-test

data:  x - p1
t = 3.2303, df = 10, p-value = 0.004508
alternative hypothesis: true mean is greater than 0.999
95 percent confidence interval:
 1.171018      Inf
sample estimates:
mean of x
 1.390909
```

Hypothesis test for a sample proportion

In Summary, our hypothesis test for a sample proportion involves the following **steps**:

	One Tailed		
	Left Tailed	**Right Tailed**	**Two Tailed Test**
Null Hypothesis	$H_0 : p = p_0$	$H_0 : p = p_0$	$H_0 : p = p_0$
Alternative Hypothesis	$H_a : p < p_0$	$H_a : p > p_0$	$H_a : p \neq p_0$
Level of Significance	α	α	α
Level of Confidence	$(1-\alpha)100\%$	$(1-\alpha)100\%$	$(1-\alpha)100\%$
Test Statistic :	$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
Rejection Region	$z < -z_\alpha$	$z > z_\alpha$	$z < -z_{\frac{\alpha}{2}}$ or $z > z_{\frac{\alpha}{2}}$
	where z_α is chosen so that $P(z > z_\alpha) = \alpha$		where $z_{\frac{\alpha}{2}}$ is chosen so that $P(z > z_{\frac{\alpha}{2}} = \frac{\alpha}{2})$

Here z has an approximate standard normal distribution.

Example 2: Lets test our Hypothesis with a 99% level of confidence:

- **Identify Null and Alternative Hypotheses:** H_0: $p = 0.00066$ where p represents the mortality rate per annum for the population of London olympians.

  H_1: $p > 0.00066$,

- **Test Statistic and its distribution:** $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, standard normal distribution.

- **Decision rule**

  ```
  > qnorm(99/100)

  [1] 2.326348
  ```

  Reject H_0 if $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > 2.326348$

- **Calculate value of test statistic and decide** $n = 10,568$, $\hat{p} = 18/(4 * 10,568) = 0.001703255$ (number deaths per year is approx 18/4).

```
> n<-10568
> hatp<-18/(4*n)
> hatp

[1] 0.0004258138

> p0<-0.00066
> z<-(hatp-p0)/(sqrt((p0*(1-p0))/(n)))
> z

[1] -0.937409

> pvalue<-1-pnorm(z)
> pvalue

[1] 0.8257259
```

$$Z = \frac{0.0004258138 - 0.00066}{\sqrt{0.00066(1-0.00066)/10,568}} = \frac{-0.0002341862}{0.0002498229} = -0.9374089.$$

In this case we do not reject the Null Hypothesis at a 99% level of significance. the p value of the data is $P(Z > -0.9374089) = 0.8257259$.

- In this case we do not have sufficient evidence to reject the Null Hypothesis in fact the sample proportion is less than the overall population proportion.

Running a test about a population proportion in R: R has a built in function to run the above test called `prop.test()`. The arguments are x = sample frequency, n=sample size, p = proportion when null hypothesis is true and alternative = "greater", "less" or "two.sided".

Example 3 :

```
> prop.test(x=18/4, n=10568, p=0.00066,  alternative="greater",correct=FALSE)

        1-sample proportions test without continuity correction

data:  18/4 out of 10568, null probability 0.00066
X-squared = 0.87874, df = 1, p-value = 0.8257
alternative hypothesis: true p is greater than 0.00066
95 percent confidence interval:
 0.0001997168 1.0000000000
sample estimates:
           p
0.0004258138
```

If we do not set `correct` to FALSE here, R will make a continuity correction and give us a slightly different p-value.

The Wald Wolfowitz Test

The Wald Wolfowitz test is a test for randomness in data with two values success (S) and failure (F). The test statistic is the number of runs of Ss and Fs in the data and unlike the test using the longest run, it does not make any assumptions about the probability of success or failure on any trial.

Given a sequence with two values, success (S) and failure (F), with N_s success' and N_f failures, let X denote the number of runs (of both S's and F's). Wald and Wolfowitz determined that for a random sequence of length N with N_s success' and N_f failures (note that $N = N_s + N_f$), the number of runs has mean and standard deviation given by

$$E(X) = \mu = \frac{2N_sN_f}{N} + 1, \qquad \sigma(X) = \sqrt{\frac{(\mu-1)(\mu-2)}{N-1}}.$$

The distribution of X is approximately normal if N_s and N_f are both bigger than 10. Therefore the Z- value:

$$Z = \frac{x-\mu}{\sigma} = \frac{x - \left(\frac{2N_sN_f}{N} + 1\right)}{\sqrt{\frac{(\mu-1)(\mu-2)}{N-1}}}$$

has a standard normal distribution.
(see http://www.itl.nist.gov/div898/handbook/eda/section3/eda35d.htm for more details).

Now we can use this information to test if our sequence of PASSs and RUSHs in Example 1 is likely to have been generated randomly (as a sequence of independent identical Bernoulli trial) or not at a 95% level of confidence.

- **Null Hypothesis** H_0 :, the sequence was generated randomly.
- **Alternative Hypothesis** H_A : The sequence was not generated randomly.
- **Test Statistic and its distribution:** $Z = \frac{x-\mu}{\sigma} = \frac{x - \left(\frac{2N_sN_f}{N} + 1\right)}{\sqrt{\frac{(\mu-1)(\mu-2)}{N-1}}}$, standard normal, where x is the observed number of runs in the data.
- **Decision rule:** Reject H_0 if observed value of Z is greater than `qnorm(.975)`= 1.96 or if the observed value of Z is less than `qnorm(.025)`= -1.96.

- **Calculate the value of the statistic and decide:** Lets S denote PASS (1 in the data vector below) and F denote RUSH (0 in data vector below). $N_s = 45$, $N_f = 87 - 45 = 42$, $N = 87$ and $x = 52$.

$$\mu = \frac{2(45)(42)}{87} + 1 = 44.44828, \quad \sigma \approx \sqrt{\frac{(43.44828)(42.44828)}{86}} \approx 4.630918.$$

$$z = (x - \mu)/\sigma = 1.630719$$

We do not have enough evidence to reject the null hypothesis here.

```
> ndrp<-c( 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1,
+ 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0,
+  0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0,
+ 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0,
+ 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1
+ )
> n<-length(ndrp)
> n

[1] 87

> ns<-sum(ndrp)
> ns

[1] 45

> nf=n - ns
> nf

[1] 42

> mu<-((2*ns*nf)/n)+1
> mu

[1] 44.44828

> sigma<-sqrt(((mu-1)*(mu-2))/(n-1))
> sigma

[1] 4.630918

> t<-rle(ndrp)
> t
```

```
Run Length Encoding
  lengths: int [1:52] 2 3 1 1 3 1 1 2 1 4 ...
  values : num [1:52] 0 1 0 1 0 1 0 1 0 1 ...

> #runs
> x<-length(t$values)
> x

[1] 52

> z<-(x-mu)/sigma
> z

[1] 1.630719

> #p-value
> 2*pnorm(-z)

[1] 0.1029497
```

- **p value** $P(z > 1.630719) + P(z < -1.630719) =$ `2*pnorm(-1.630719)` $= 0.1029497$.

Example: Use the Wald Wolfowitz Runs Test to test the following sequence of consecutive baskets and misses for basketball player J R Smith for randomness at a 95% level of confidence:

MMBBMMMMBMMMBBMMBMBMMBMMBMBMMBBMMMMBBMBMBBBBMMBMBMMBBBBMMM BMMMBMM BBMB-MMMBBMMB BMBMBMMBMMMB MMMMMMMBBMM MMMBMMBMBBMMMM BBBMMMMMMBBBB MMBMMB-MMBMBMMMMB BMBMBBMBBMB BMMBBBMM BBMBMMMMBMBMM BMMBBMBMMBBMMBBBBMM BMMBM-MMMM BMBMBBBMBMBMMBBMMBMBMMM MMMMM

We can just run our code from above with the data encoded in `JR` below to get a p-value.

```
> JR<-c(0,0,1,1,0,0,0,0,1,0,0,0,1,1,0,0,1,0,1,
+ 0,0,1,0,0,1,0,1,0,0,1,1,0,0,0,0,1,1,0,1,0,1,1,1,
+ 1,0,0,1,0,1,0,0,1,1,1,1,0,0,0,
+ 1,0,0,0,1,0,0,
+ 1,1,0,1,0,0,0,1,1,0,0,1,
+ 1,0,1,0,1,0,0,1,0,0,0,1,
+ 0,0,0,0,0,0,0,1,1,0,0,
+ 0,0,0,1,0,0,1,0,1,1,0,0,0,0,
```

```
+ 1,1,1,0,0,0,0,0,0,1,1,1,1,
+ 0,0,1,0,0,1,0,0,1,0,1,0,0,0,0,1,
+ 1,0,1,0,1,1,0,1,1,0,1,
+ 1,0,0,1,1,1,0,0,
+ 1,1,0,1,0,0,0,0,1,0,1,0,0,
+ 1,0,0,1,1,0,1,0,0,1,1,0,0,1,1,1,1,0,0,
+ 1,0,0,1,0,0,0,0,0,0,
+ 1,0,1,0,1,1,1,0,1,0,1,0,0,1,1,0,0,1,0,1,0,0,0,
+ 0,0,0,0,0)
> n<-length(JR)
> n

[1] 232

> ns<-sum(JR)
> ns

[1] 95

> nf=n - ns
> nf

[1] 137

> mu<-((2*ns*nf)/n)+1
> mu

[1] 113.1983

> sigma<-sqrt(((mu-1)*(mu-2))/(n-1))
> sigma

[1] 7.349133

> t<-rle(JR)
> t

Run Length Encoding
  lengths: int [1:121] 2 2 4 1 3 2 2 1 1 1 ...
  values : num [1:121] 0 1 0 1 0 1 0 1 0 1 ...

> #runs
> x<-length(t$values)
> x
```

```
[1] 121

> z<-(x-mu)/sigma
> z

[1] 1.061584

> #p-value
> 2*pnorm(-z)

[1] 0.2884245
```

With a p-value of 0.2884245, we would not reject the Null hypothesis that the data was generated randomly, in other words; we do not have sufficient evidence of streakiness in the data.

Testing two population proportions ($\mathbf{p_1 - p_2}$) from independent samples for equality

Assumptions:

1. We randomly select two independent samples from Population 1 and Population 2 of sizes n_1 and n_2 respectively. The proportion of interest in the sample from population 1 is $\hat{p}_1$ and the proportion of interest in the sample from population 2 is $\hat{p}_2$.

2. The true proportion in population 1 is p_1 and the true proportion in population 2 is p_2.

3. n_1 and n_2 are large in that $n_1\hat{p}_1 \geq 15$ and $n_1(1-\hat{p}_1) \geq 15$ and $n_2\hat{p}_2 \geq 15$ and $n_2(1-\hat{p}_2) \geq 15$

If the above assumptions are satisfied, then the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal with mean $p_1 - p_2$ and standard deviation

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \approx \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

This means that the sampling distribution of

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

is approximately standard normal (note: the value of $p_1 - p_2$ is unknown here, however if we are testing for equality of the population proportions we set it equal to 0).

Thus we know the sampling distribution of $\hat{p}_1 - \hat{p}_2$ which allows us to test hypotheses about $p_1 - p_2$.

There are other variations of the estimate of the variance used and the distribution used depending on sample size etc... You will learn about the details in higher level statistics courses. Our Test goes as follows:

Out test goes as follows:

	Left Tailed	**Left Tailed**	**Two Tailed Test**
Null Hypothesis	$H_0 : p_1 - p_2 = 0$	$H_0 : p_1 - p_2 = 0$	$H_0 : p_1 - p_2 = 0$
Alternative Hypothesis	$H_a : p_1 - p_2 < 0$	$H_a : p_1 - p_2 > 0$	$H_a : p_1 - p_2 \neq 0$
Level of Significance	α	α	α
Level of Confidence	$(1-\alpha)100\%$	$(1-\alpha)100\%$	$(1-\alpha)100\%$
Test Statistic :	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sigma_{\hat{p}_1 - \hat{p}_2}}$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sigma_{\hat{p}_1 - \hat{p}_2}}$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sigma_{\hat{p}_1 - \hat{p}_2}}$
$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$\approx \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$		
Distribution	St. Normal	St. Normal	St. Normal
Rejection Region	$z < -z_\alpha$	$z > -z_\alpha$	$z < -z_{\frac{\alpha}{2}}$ or $z > z_{\frac{\alpha}{2}}$
	where z_α is chosen so that $P(z > z_\alpha) = \alpha$	where z_α is chosen so that $P(z > z_\alpha) = \alpha$	where $z_{\frac{\alpha}{2}}$ is chosen so that $P(z > z_{\frac{\alpha}{2}} = \frac{\alpha}{2})$
P Value	$P(z <$ observed value$)$	$P(z >$ observed value$)$	$2P(z >$ |observed value|$)$

Example : Football : Peyton Manning's had a (regular season) career completion percentage of 65.3% (6125 completions out of 9380 attempts) and Tom Brady has a (regular season) career completion percentage of 63.8% (5244 completions out of 8224 attempts) could these players have the same level of skill and the difference in the numbers be due to random variation or did Peyton Manning have an edge when it came to completing passes or is the difference in these percentages large enough to be considered so unlikely to happen if both had equal abilities that we would reject that hypothesis?

Here we have $x_1 = 6125$, $n_1 = 9380$, $x_2 = 5244$, $n_2 = 8224$, $\hat{p}_1 = \dfrac{x_1}{n_1}$, $\hat{p}_2 = \dfrac{x_2}{n_2}$.

- **Null and Alt. Hyp.:** Let's suppose that we would like to test the hypothesis $H_0 : \ p_1 - p_2 = 0$ against the alternative two sided hypothesis $H_1 : \ p_1 - p_2 \neq 0$ at a 5% level of significance.
- **Test Statistic and distribution:** $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sigma_{\hat{p}_1 - \hat{p}_2}}$, standard normal distribution.
- **Decision Rule:** Notice that since the distribution is approximately normal and our desired level of significance is $\alpha = 0.05$, our decision rule should be: Reject H_0 if $|z| > 1.96 =$ `qnorm(.975)` here.

- **Value of test statistic and decision:** We have

$$\sigma_{\hat{p}_1-\hat{p}_2} \approx \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \approx 0.007$$

and the value of our test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sigma_{\hat{p}_1-\hat{p}_2}} = 2.122016$$

```
> x1<-6125
> x2<-5244
> n1<-9380
> n2<-8224
> p1<-x1/n1
> p2<-x2/n2
> sigma<-sqrt((p1*(1-p1)/n1)+(p2*(1-p2)/n2))
> sigma

[1] 0.007228579

> z<-(p1-p2)/sigma
> z

[1] 2.122016

> 2*(1-pnorm(z))

[1] 0.03383639
```

So we reject H_0 in this case and conclude that the difference between the completion percentages is significant using a 5% level of significance. (Of course statistical significance is not always the same as real life significance).

- **p value** The p value of our data is $P(z > 2.122016) + P(z < -2.122016)$ = `2*pnorm(-2.122016)` ≈ 0.034.

Testing in R We can use the `prop.test()` function in R to test our hypothesis that two proportions are equal. We use the command `prop.test(x, n, alternative = "two.sided")` where $x = (x_1, x_2)$ the counts in both samples, $n = (n_1, n_2)$ the sample sizes and `alternative` gives the nature of the test.

```
> prop.test(c(6125, 5244), c(9380, 8224), alternative="two.sided", correct=FALSE)
```

```
        2-sample test for equality of proportions without continuity
        correction

data:  c(6125, 5244) out of c(9380, 8224)
X-squared = 4.5076, df = 1, p-value = 0.03375
alternative hypothesis: two.sided
95 percent confidence interval:
 0.001171406 0.029506914
sample estimates:
   prop 1    prop 2
0.6529851 0.6376459
```

Testing for equality for two population means

In this section we will consider two independent random samples from different populations, Population 1 and Population 2. Let μ_1 be the population mean for a particular variable for Pop. 1 and let μ_2 be the the population mean for the same variable for Pop. 2. We wish to compare the means μ_1 and μ_2. We will test the hypothesis that the difference between the means is 0 i.e. $\mu_1 - \mu_2 = 0$. We will use the difference between the sample means, $\bar{x_1} - \bar{x_2}$ to make a decision.

Sampling Distribution of $\bar{x_1} - \bar{x_2}$.

Assumptions:

- We have chosen two random samples in an independent manner of size n_1 and n_2, from Populations 1 and 2 respectively.
- The populations are both normally distributed.
- We will assume that the variances in both populations are equal (one would use a different estimate of variance otherwise).

Variations in these assumptions require variations in the nature of the test, however for large samples sizes, the test shown below can almost always be used.

Under these assumptions, we can show that the test statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

(where $\bar{x}_1$ and $\bar{x}_2$ are the means from the respective samples, $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ and S_1^2 and S_2^2 are the sample variances of the respective samples) has a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

	OneTailed Left Tailed	Left Tailed	Two Tailed Test
Null Hypothesis	$H_0 : \mu_1 - \mu_2 = 0$	$H_0 : \mu_1 - \mu_2 = 0$	$H_0 : \mu_1 - \mu_2 = 0$
Alternative Hypothesis	$H_a : \mu_1 - \mu_2 < 0$	$H_a : \mu_1 - \mu_2 > 0$	$H_a : \mu_1 - \mu_2 \neq 0$
Level of Significance	α	α	α
Level of Confidence	$(1-\alpha)100\%$	$(1-\alpha)100\%$	$(1-\alpha)100\%$
Test Statistic :	$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$	$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$	$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
	$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$	$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$	$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$
Distribution	$t(n_1+n_2-2)$	$t(n_1+n_2-2)$	$t(n_1+n_2-2)$
Rejection Region	$t < -t_\alpha$	$t > -t_\alpha$	$t < -t_{\frac{\alpha}{2}}$ or $t > t_{\frac{\alpha}{2}}$
	where t_α is chosen so that $P(t > t_\alpha) = \alpha$	where t_α is chosen so that $P(t > t_\alpha) = \alpha$	where $t_{\frac{\alpha}{2}}$ is chosen so that $P(t > t_{\frac{\alpha}{2}} = \frac{\alpha}{2})$
P Value	$P(t <$ observed value$)$	$P(t >$ observed value$)$	$2P(t >$ |observed value|$)$

Example: Did my 12 week training program make a difference to my race time? The following data show my times for 30 one mile races prior to my 12 week training program and the times for 30 one mile races after my 12 week training program.

Before training program (mean $\bar{x}_1 = 4.98303$):
{4.994, 5.08, 5.145, 5.066, 5.032, 4.906, 4.786, 5.245, 4.871, 4.909, 4.934, 4.761, 4.787, 4.818, 4.995, 4.837, 4.947, 5.028, 4.760, 5.078, 5.167, 4.962, 5.036, 4.989, 4.871, 5.262, 5.096, 5.150, 4.963, 5.016}

After training program (mean $\bar{x}_2 = 4.60493$):
{4.490, 4.584, 4.466, 4.517, 4.462, 4.535, 4.520, 4.355, 4.427, 4.676, 4.901, 4.551, 4.548, 4.46, 4.648, 4.488, 4.480, 4.650, 4.783, 4.755, 4.566, 4.741, 4.873, 4.947, 4.721, 4.520, 4.580, 4.470, 4.584, 4.850}

Is the difference in these means significant at a level of significance equal to .01?

- **Null and Alt. Hyp.** H_0: $\mu_1 - \mu_2 = 0$, H_1: $\mu_1 - \mu_2 > 0$.

- **Test Statistic and distribution:**

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where} \quad S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}.$$

  Distribution: t distribution with n_1+n_2-2 degrees of freedom, $n_1 = n_2 = 30$.

- **Decision Rule:** Reject H_0 the observed value of t is greater than `qt(.99, df= 58)` ≈ 2.392377.

- **Calculate statistic and make decision:** $S_1^2 = 0.0188621$, $S_2^2 = 0.02356931$,
  $S_p^2 = \frac{(29)0.0188621+(29)0.02356931}{58} = 0.0212157$
  $t = (4.98303 - 4.60493)/\sqrt{0.02121571(\frac{2}{30})} = 10.05364$.
  Here we reject the null hypothesis and conclude that the 12 week training program made a difference with a 99% level of confidence.

- **p value** `1-pt(10.05365, df=58)` $= 1.265654/10^{14}$.

Using R to test a hypotheisis about difference between two means: Here we use `t.test(x, y, alternative="greater", var.equal=TRUE)`, where `x` and `y` are the data sets in question and we set our alternative hypothesis appropriately. We can set `var.equal` to either `TRUE` or `FALSE` depending on whether we believe the variances in the populations are equal or not respectively.

```
> bf<-c(4.994, 5.08, 5.145, 5.066, 5.032, 4.906, 4.786,
+      5.245, 4.871, 4.909, 4.934, 4.761, 4.787, 4.818,
+       4.995, 4.837, 4.947, 5.028, 4.760, 5.078, 5.167, 4.962,
+       5.036, 4.989, 4.871, 5.262, 5.096, 5.150, 4.963,
+       5.016)
>  af<-c(4.490, 4.584, 4.466, 4.517, 4.462, 4.535, 4.520,
+       4.355, 4.427, 4.676, 4.901, 4.551, 4.548, 4.46, 4.648,
+       4.488, 4.480, 4.650, 4.783, 4.755, 4.566,
+       4.741, 4.873, 4.947, 4.721, 4.520, 4.580, 4.470,
+       4.584, 4.850)
> qt(.99, df= 58)

[1] 2.392377

> n1<-length(bf)
> n2<-length(af)
> barx1<-mean(bf)
> barx1
```

```
[1] 4.983033

> barx2<-mean(af)
> barx2

[1] 4.604933

> spsquare<-((n1-1)*((sd(bf))^2) + (n2-1)*((sd(af))^2))/(n1+n2-2)
> spsquare

[1] 0.0212157

> s<-sqrt(spsquare*((1/n1)+(1/n2)))
> t<-(barx1-barx2)/s
> t

[1] 10.05365

> 1-pt(10.05365, df=58)

[1] 1.265654e-14

> t.test(bf,af,alternative="greater",var.equal=TRUE, correct=FALSE)

        Two Sample t-test

data:  bf and af
t = 10.054, df = 58, p-value = 1.272e-14
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.3152358       Inf
sample estimates:
mean of x mean of y
 4.983033  4.604933
```