Hypothesis Testing
Introduction to Statistics Using R (Psychology 9041B)

Paul Gribble

Winter, 2016

Sampling Distributions & t-tests

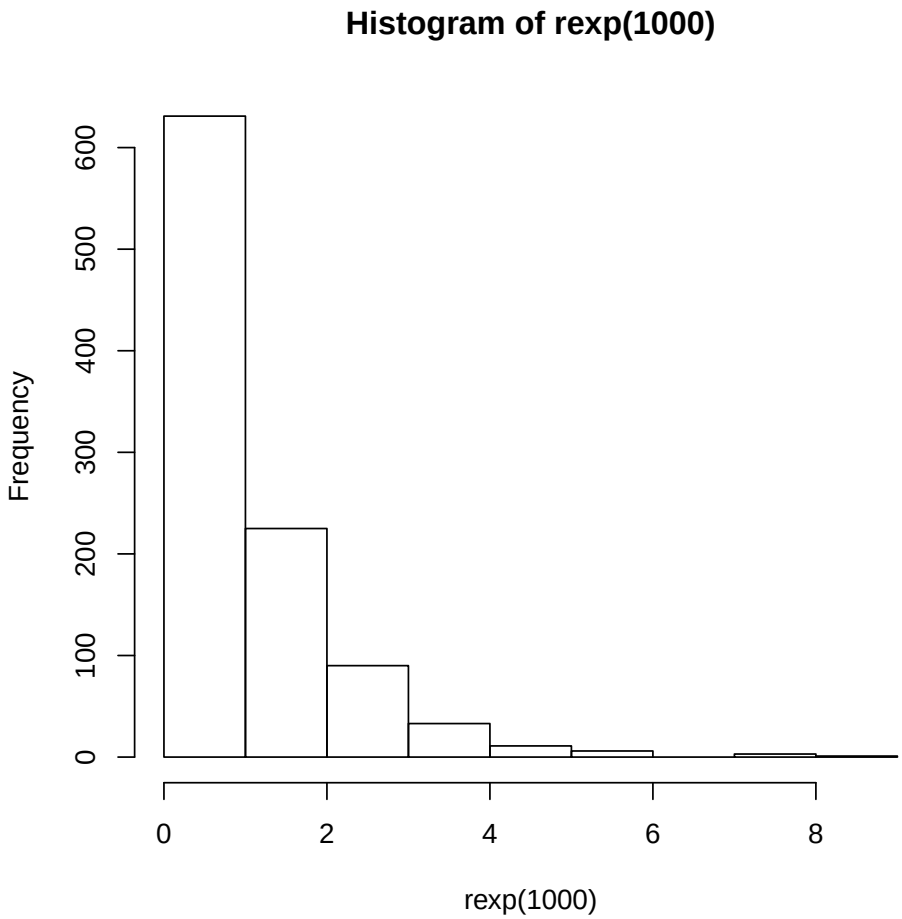
In R it is easy to compute measures of central tendency and dispersion. Let's generate a random sample (taken from a normal distribution with mean zero and standard deviation 1) of size $n = 10$, and compute various measures. Note that when you include round brackets around a statement in R, it prints the output of the statement to the screen.

```
> (x <- rnorm(10, 0, 1))

 [1]  0.7377453  0.6865771  0.2426501 -1.2288901  0.9209672  0.8779727
 [7]  0.7142595  1.4203905 -1.0542188  0.6553102

> (mean(x))

[1] 0.3972764

> (median(x))

[1] 0.7004183

> var(x)

[1] 0.7437691

> sd(x)

[1] 0.8624205
```

Central limit theorem

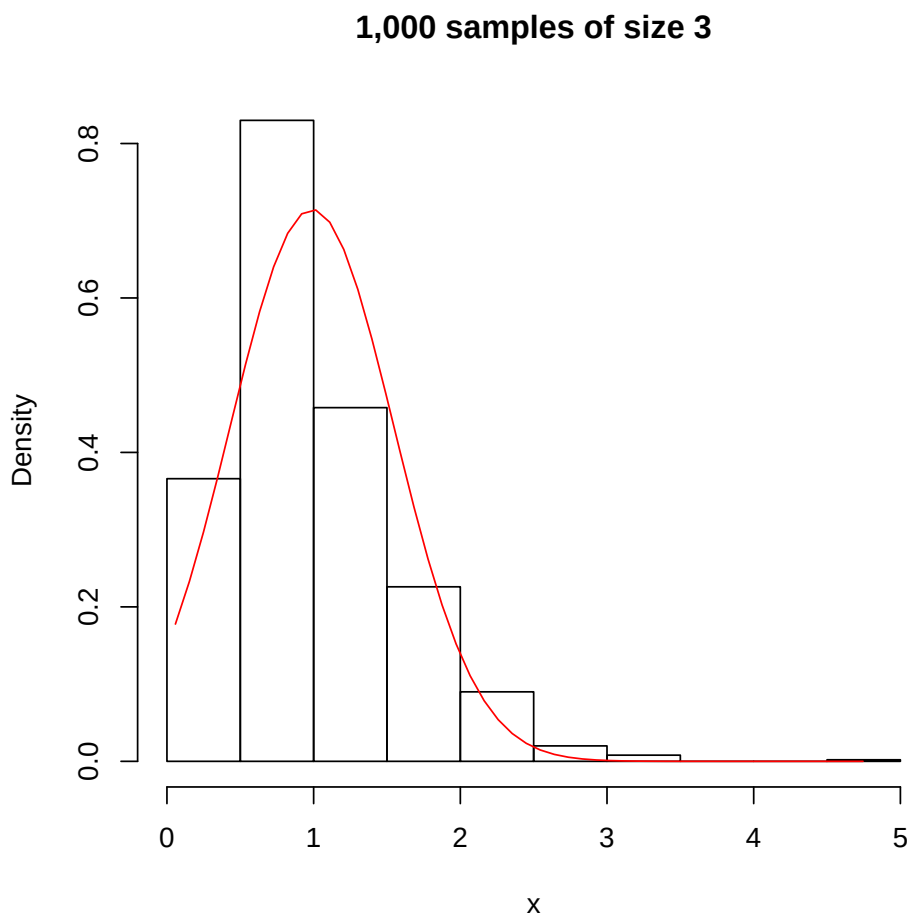
Given random sampling, the sampling distribution of the mean approaches a normal distribution as the size of the sample increases, even if the population distribution of raw scores is *not* normally distributed. The central limit theorem states that the sum of a large number of independent observations from the same distribution has an approximate normal distribution, and this approximation steadily improves as the number of observations increases.

We can demonstrate this relatively easily in R — and along the way learn some useful R functions. First let's choose a non-normal distribution from which we are going to sample. We'll choose an exponential distribution, which is bunched up at small values with a long positive tail:



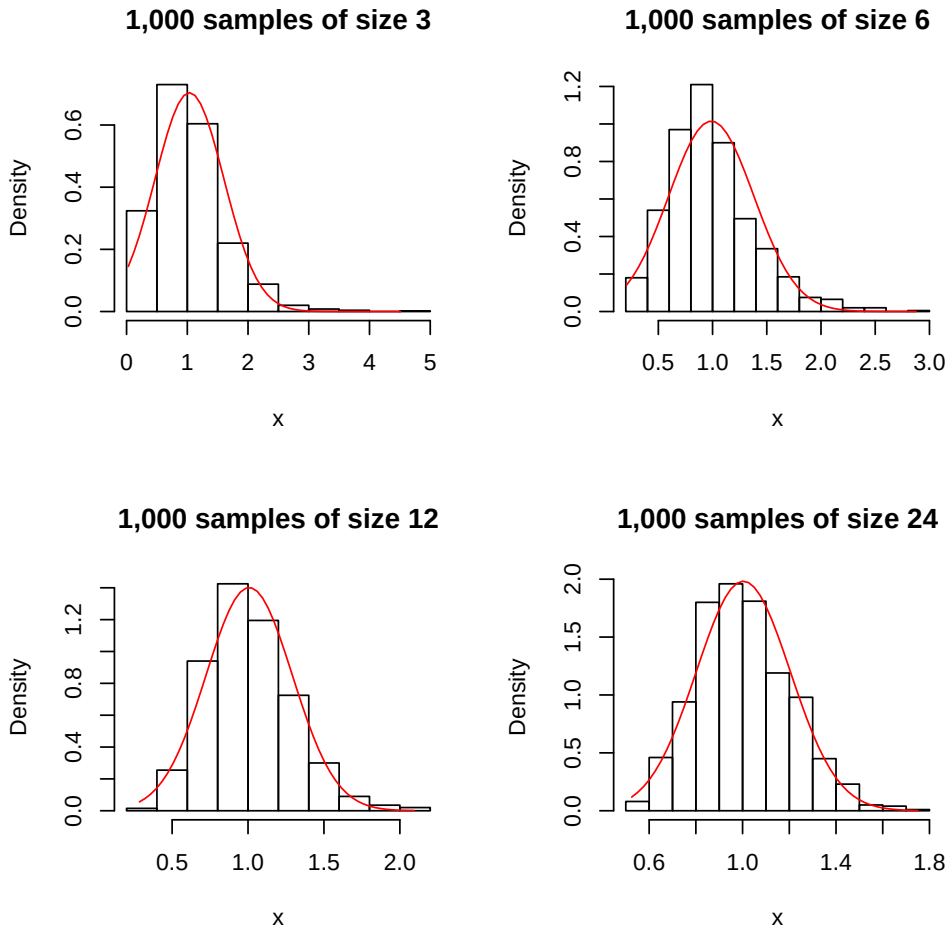
So what we're going to do is repeatedly sample ($n = 10$) from an exponential distribution, each time taking the mean of our 10 samples and storing the means. We'll then vary the sample size and see that as sample size increases, the approximation of the sampling distribution of means to a normal distribution gets better and better. Here is what the sampling distribution of means looks like when 1,000 samples of size 3 are taken:

```
> nsamples <- 1000
> samplesize <- 3
> x <- rep(0,nsamples)
> for (i in 1:1000) {
+         x[i] <- mean(rexp(samplesize))
+ }
> xfit <- seq(min(x), max(x), length=50)
> yfit <- dnorm(xfit, mean=mean(x), sd=sd(x))
> hist(x, prob=T, main=paste( "1,000 samples of size",  samplesize ))
> lines(xfit, yfit, col="red")
```



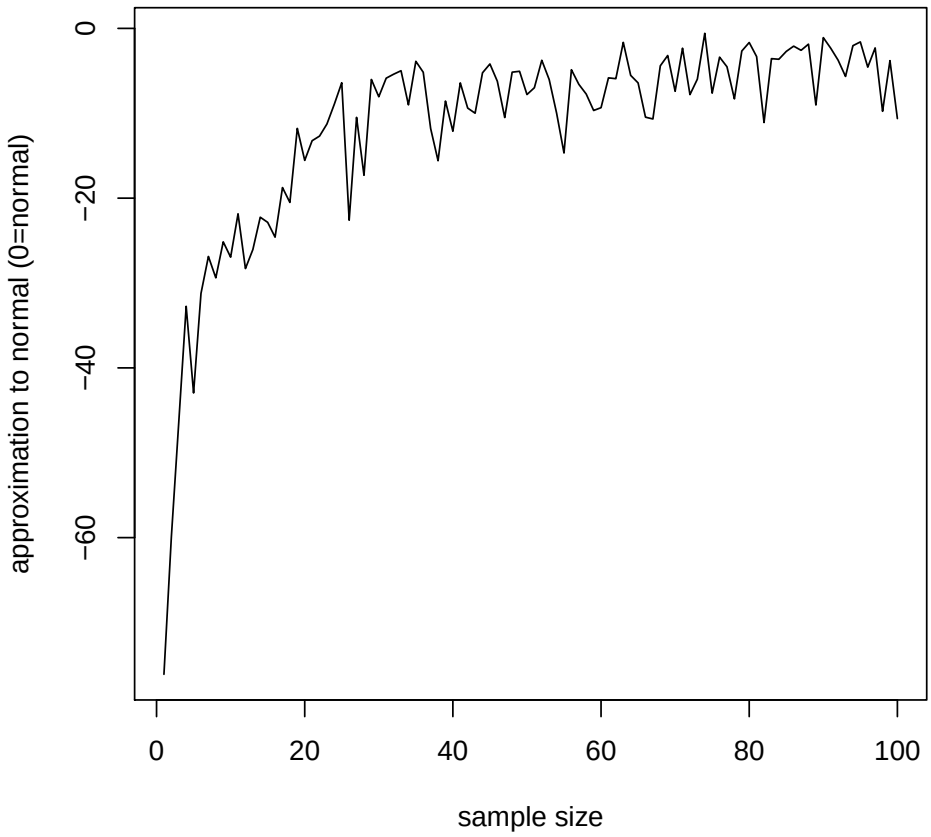
Now let's see what happens when we take 1,000 samples of size 3, 6, 12 and 24:

```
> # instruct the figure to plot sequentially in a 2x2 grid
> par(mfrow=c(2,2))
> nsamples <- 1000
> x <- rep(0,nsamples)
> for (s in c(3,6,12,24)){
+ for (i in 1:1000) {
+         x[i] <- mean(rexp(s))
+ }
+ xfit <- seq(min(x), max(x), length=50)
+ yfit <- dnorm(xfit, mean=mean(x), sd=sd(x))
+ hist(x, prob=T, main=paste( "1,000 samples of size",  s ))
+ lines(xfit, yfit, col="red")
+ }
```



You can see as the size of the sample increases, the approximation of the sampling distribution of means to a normal distribution gets better and better. Just to get even more quantitative, we can quantify the closeness to a normal distribution by performing a Shapiro-Wilk normality test (`shapiro.test()`) for a range of sample sizes, and plotting the relationship between the sample size and the p-value (probability that the null hypothesis — that the data are sampled from a non-normal distribution — is true). For visual clarity we will plot the log(p-value)

```
> nsamples <- 1000
> srange <- 1:100
> x <- rep(0,nsamples)
> p <- rep(0,max(srange))
> for (s in srange){
+ for (i in 1:1000) {
+         x[i] <- mean(rexp(s))
+ }
+ p[s] <- shapiro.test(x)$p.value
+ }
> par(mfrow=c(1,1))
> plot(srange, log(p), type="l", xlab="sample size",
+          ylab = "approximation to normal (0=normal)")
```



Hypothesis testing: a single case

Let's say you know in advance that the mean IQ score in the population is 100 and the standard deviation is 15 (as in the Wechsler Adult Intelligence Scale). You have an idea that listening to classical music increases intelligence — even listening to a single piece of music. You get your sister to listen to Beethoven's Ninth and then measure her IQ [1]. Her IQ score after listening to Beethoven's Ninth is 140. Did the treatment work?

Hypothesis testing operates by answering the question:

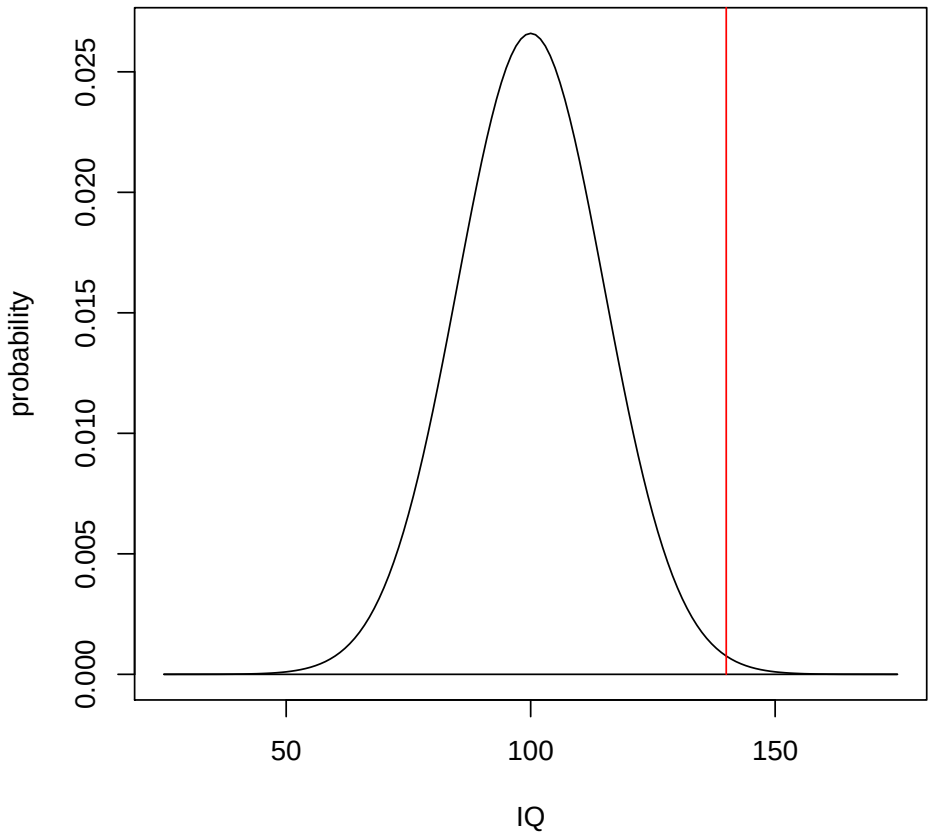
> What is the probability that we could have observed a score that high under the null hypothesis that the treatment had no effect on IQ?

Said in a different way, what is the probability that we could have observed a score that high by randomly sampling from the population of individuals who have not undergone the treatment [2]?

The probability equals the area under the normal curve to the right of the $IQ = 140$ point:

[1] Note that this is not an example of a good experiment. Obviously it would be helpful to know her IQ before listening to the music and then see if it changed.

[2] Again, this is a bad example of an experiment, obviously in the general population some individuals have listened to classical music before



To determine the probability in `R` we can use the `pnorm()` command:

```
> (p <- 1-pnorm(140, mean=100, sd=15))

[1] 0.003830381
```

The reason we have to use `1-pnorm()` is because by default `pnorm()` reports the lower tail of the distribution, and we want the upper tail.

So the chances we would have observed an IQ in an individual randomly sampled from a population who was *not* affected by the treatment is only 3.8 in 1,000. Under the logic of hypothesis testing, we have to decide if this probability is small *enough* in order for us to reject the null hypothesis. If we decide to use an α -level of 0.05 (5 %) then we would be forced to reject the null hypothesis and conclude that the treatment indeed had an effect.

If we didn't have a hypothesis *a priori* about the direction of the effect of the treatment (whether it should increase or decrease IQ) then we would need to divide our α -level by two - because we would be doing a *two-tailed* test. The above test was a *one-tailed* test because we had a hypothesis in advance that the treatment should raise IQ.

Hypothesis testing: a single group

Now let's say we decide to test 20 people instead of just one. We randomly sample them from the population, apply our treatment, observe that the mean IQ of our sample was 115, and we ask the

same question:

> What is the probability of observing a *mean* IQ score as high as we observed given the *null hypothesis* that there was no effect of the treatment?

We can transform our sample mean into a *z-score*, and then compute a probability using the `pnorm()` function. Remember the definition of a z-score:

$$z = \frac{\bar{X}_1 - \mu}{\sigma/\sqrt{N}} \tag{1}$$

```
> (z <- (115-100)/(15/sqrt(20)))

[1] 4.472136

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

[1] 3.872108e-06
```

So there is a 3.8 in one million chance that we could have observed a mean IQ as large as we did if we had sampled 20 people from a population of individuals who were not affected by the treatment. Again, we compare the observed probability to our α -level, and make a decision to reject (or not reject) the null hypothesis that the treatment had no effect. In this case if our α -level was 0.05, we would reject the null hypothesis and conclude the treatment indeed had an effect.

Note that the logic of hypothesis testing fundamentally depends on the assumption that our sample is a true random sample from the population. If the mean IQ in the population is 100 but we take a sample of 20 from first year university students, this would represent a biased sample. The IQ scores of first year university students are likely not representative of the population of humans as a whole.

Typically we do not know the mean and standard deviation of the population, we have to estimate them from our sample. The best estimate of the population mean is the sample mean. The best estimate of the population standard deviation is *the standard error of the sampling distribution of the mean*. For very large samples (e.g. $N > 100$) this is fairly accurate. For small samples it is not. Another theoretical sampling distribution exists that is appropriate for these situations: the t-distribution.

The t-distribution

The t-distribution is similar to the z-distribution, however there is a different shape for each sample size N.

Let's do the same example as above: We sample 20 subjects at random from the population. Let's assume we don't know the population standard deviation. We assume the population is normally distributed. Let's compute the probability of observing a mean IQ of 115 (15 points higher than the supposed population mean) or higher given a sample of size $N = 20$ and $sd = 30$.

$$t = \frac{\bar{X}_1 - \mu}{sd/\sqrt{N}} \tag{2}$$

```
> (tobs <- (115-100)/(30/sqrt(20)))
```

```
[1] 2.236068
```

```
> (p <- 1-pt(tobs, 19))
```

```
[1] 0.01877027
```

So there is a 1.9 % chance of observing such a mean given the null hypothesis. If our α -level is 0.05 then we would reject the null hypothesis and conclude that the treatment had an effect.

Confidence intervals for the mean

The confidence interval for the mean at a probability of $1 - \alpha$ is:

$$\bar{X} \pm t_\alpha \left(\frac{sd}{\sqrt{N}} \right) \tag{3}$$

So for our sample of size $N = 20$ with $\bar{X} = 115$ and $sd = 30$, the 95 %confidence intervals can be calculated as:

```
> n <- 20
> tcrit <- qt(1-(0.05/2), df=n-1)
> sxbar <- 30/sqrt(20)
> (cim <- c(115-(tcrit*sxbar), 115+(tcrit*sxbar)))
```

```
[1] 100.9596 129.0404
```

The degrees of freedom for the t-statistic are N-1. Note also we divide our α by two in order to get both tails of the t-distribution.

t-test for difference between groups

Assume we have two random samples, and we want to test the hypotheses that these samples have been drawn from the same population (null hypothesis) or different populations (alternate hypothesis). We compute the t-statistic as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} \tag{4}$$

Again, the denominator of this fraction can be estimated from the sample data. The term actually depends on whether scores in the two samples are *correlated* or *independent*.

independent groups t-test

When the two groups are independent (e.g. separate subjects), then:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left[\frac{((N_1-1)s_1^2)+((N_2-1)s_2^2)}{N_1+N_2-2}\right]\left[\frac{1}{N_1}+\frac{1}{N_2}\right]}} \tag{5}$$

$$df = N_1 + N_2 - 2 \tag{6}$$

In `R` it is dead easy to run a t-test, using the `t.test()` function:

```
> g1 <- c(5,4,4,6,5)
> g2 <- c(6,7,5,8,7)
> t.test(g1, g2, alternative="two.sided", paired=FALSE, var.equal=TRUE)

        Two Sample t-test

data:  g1 and g2
t = -2.846, df = 8, p-value = 0.02161
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.2584451 -0.3415549
sample estimates:
mean of x mean of y
      4.8       6.6
```

Note that we have to tell `t.test()` whether we want a one-tailed or two-tailed test, whether the groups are correlated (paired) or independent, and whether or not we want to assume that group variances are equal or not. The homogeneity of variances is an underlying assumption of the t-test. If it is violated you can simply tell `t.test()` that `var.equal=FALSE` and it will run a corrected version of the test. You can test the homogeneity of variances assumption using `bartlett.test()`:

```
> bartlett.test(c(g1,g2), c(rep(1,5), rep(2,5)))

        Bartlett test of homogeneity of variances

data:  c(g1, g2) and c(rep(1, 5), rep(2, 5))
Bartlett's K-squared = 0.33533, df = 1, p-value = 0.5625
```

In this case `p=0.5625` so we do not reject the null hypothesis that the variances are equal — in other words homogeneity of variance has not been violated.

correlated groups t-test

For correlated groups,

$$\bar{D} = X_{i1} - X_{i2} \tag{7}$$

$$t = \frac{\sum D_i}{\sqrt{\frac{N\sum D_i^2 - (\sum D_i)^2}{N-1}}} \tag{8}$$

$$df = N - 1 \tag{9}$$

When the two groups are correlated, as in when the same subject contributes a score in each group, we simply pass `t.test()` the argument `paired=TRUE`:

```
> t.test(g1, g2, alternative="two.sided", paired=TRUE, var.equal=TRUE)

        Paired t-test

data:  g1 and g2
```

```
t = -4.8107, df = 4, p-value = 0.008581
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.8388506 -0.7611494
sample estimates:
mean of the differences
                  -1.8
```

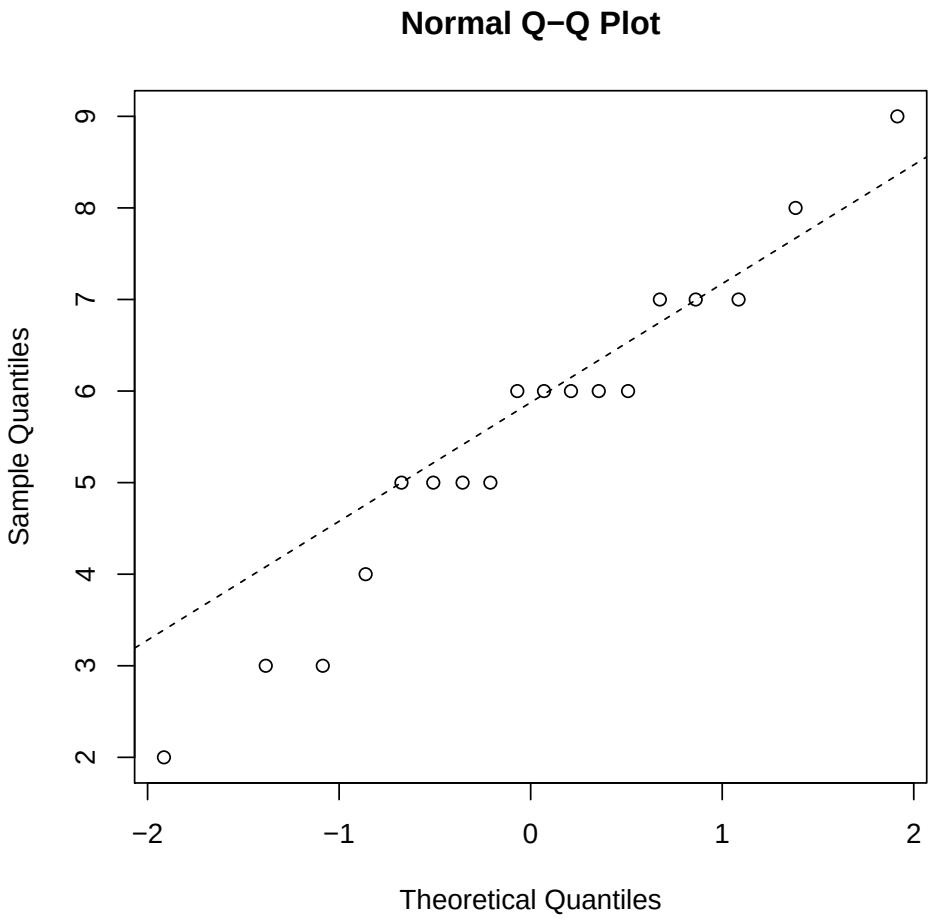
Testing the normality assumption

In R we can test the normality assumption using `shapiro.test()`:

```
> g3 <- c(5,6,8,7,6,3,4,5,6,7,6,5,5,6,7,9,3,2)
> shapiro.test(g3)

        Shapiro-Wilk normality test

data:  g3
W = 0.96427, p-value = 0.6856
```

The p-value is 0.6856 which is greater than our α-level of 0.05, so we fail to reject the null hypothesis that the sample was drawn from a normal population.

Another visual method is to generate a normal quantile-quantile plot:



If the sample is normally distributed, the data points should all fall along the dashed line. If the data are not normally distributed, then a non-parametric test is more appropriate than the t-test (e.g the `wilcox.test()`).