Permutation Tests

Andrew McAdam

14/10/2020

An Overview of Parametric Statistics

In most cases in statistics, we calculate a test statistic and compare this to a theoretical distribution to determine the probability of getting a test statistic as odd as this when the null hypothesis is true. However, the appropriateness of the P-value depends on how appropriate the data are for the assumed theoretical distribution. When we think of "statistical assumptions" these are a list of the key ways in which the data must be like the theoetical distribution in order for the P-value to be correct

What Happens When Assumptions Aren't Met?

When assumptions of a statistical test aren't met we can:

- Ignore the violation of the assumption

- Transform the data to meet the assumptions

- Use an alternative approach

Permutation Test

One alternative approach is to abandon all theoretical distributions and instead use permutations to create an *empirical null distribution* of the test statistic that we are interested in. We can then use this null distribution to calculate the appropriate P-value. This approach is interesting because it doesn't really matter what distribution the data follow, because the null distribution is created empirically.

Steps in a Permutation Test

- Measure the thing you are interested in

- Reshuffle the data, breaking apart the hypothesize relationship to create "null" associations between the variables of interest

- For each 'reshuffle' calculate the measurement of interest

- Repeat this many (e.g. 10,000) times

- Calculate the P-value as the number of iterations from the null distribution that have a calculated value as extreme or more extreme than our original calculated value.

A Worked Example

If you recall we previously used a ficticious dataset that was based on the experiment by Dantzer et al. 2013 in which they exposed some squirrels to playbacks of other squirrels to trick them into thinking they were surrounded by more neighbours. Other squirrels were used as controls. The control squirrels had playbacks of chickadees.

Why did Dantzer et al. 2013 play the calls of chickadees to the control squirrels? What would you call this?

Creating the FCM data

We will just create these fictitious data manually.

```
treatment<-c(rep("control", 9), rep("experimental", 9))
fcm<-c(283, 312, 291, 259, 216, 201, 267, 326, 241, 314, 320, 310, 340, 299, 268, 345, 271, 285)
sq_fcm_data<-data.frame(treatment, fcm)
sq_fcm_data$treatment<-as.factor(sq_fcm_data$treatment)
sq_fcm_data$treatment<-relevel(sq_fcm_data$treatment, ref = "experimental")

print(sq_fcm_data)
```

```
##       treatment fcm
## 1       control 283
## 2       control 312
## 3       control 291
## 4       control 259
## 5       control 216
## 6       control 201
## 7       control 267
## 8       control 326
## 9       control 241
## 10 experimental 314
## 11 experimental 320
## 12 experimental 310
## 13 experimental 340
## 14 experimental 299
## 15 experimental 268
## 16 experimental 345
## 17 experimental 271
## 18 experimental 285
```

Analysing Using a T-test

We previously analyzed these data as an independent two-sample t-test

```
t.test(fcm ~ treatment, data=sq_fcm_data, mu=0, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  fcm by treatment
## t = 2.365, df = 16, p-value = 0.031
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##   4.099377 75.011734
## sample estimates:
## mean in group experimental     mean in group control
##                  305.7778                  266.2222
```

Testing Assumptions

Are the varoiances equal?

```
library(car)
```

```
## Loading required package: carData
```

```
leveneTest(fcm~treatment, data=sq_fcm_data, center=mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##        Df F value Pr(>F)
## group   1  1.3604 0.2606
##        16
```
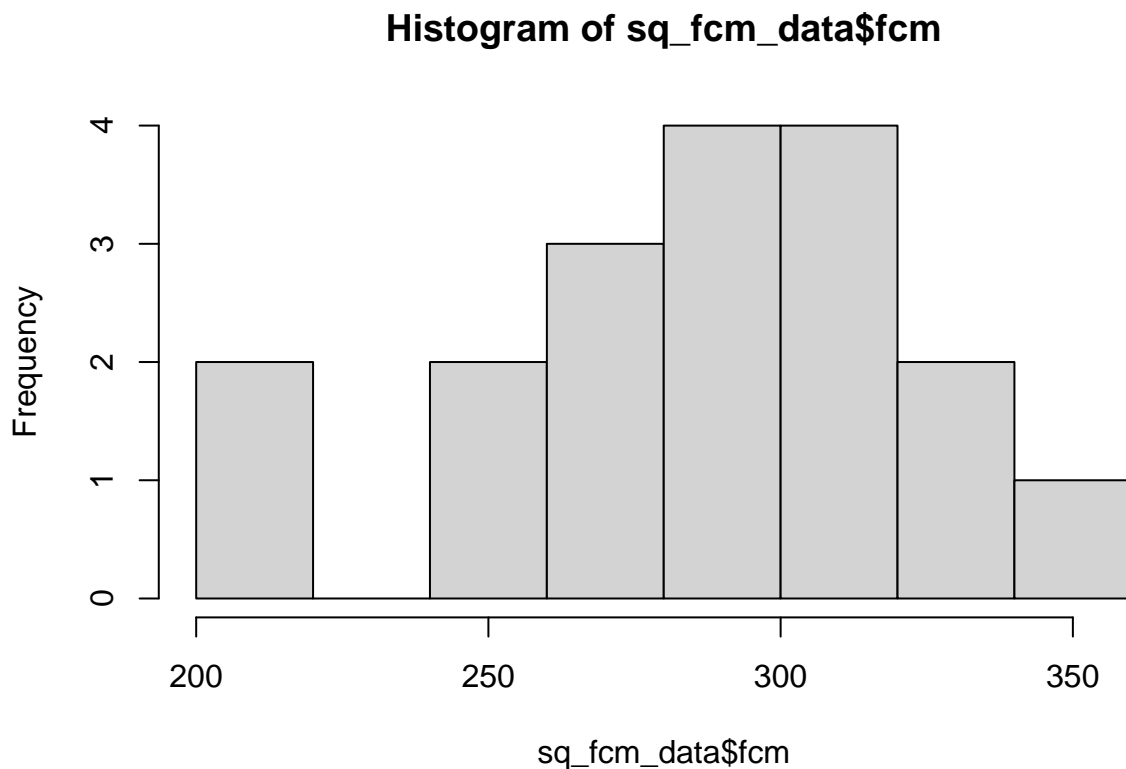
The P-value is not significant so we do not reject the null hypothesis that the variances are equal.

Are the data normally distributed?
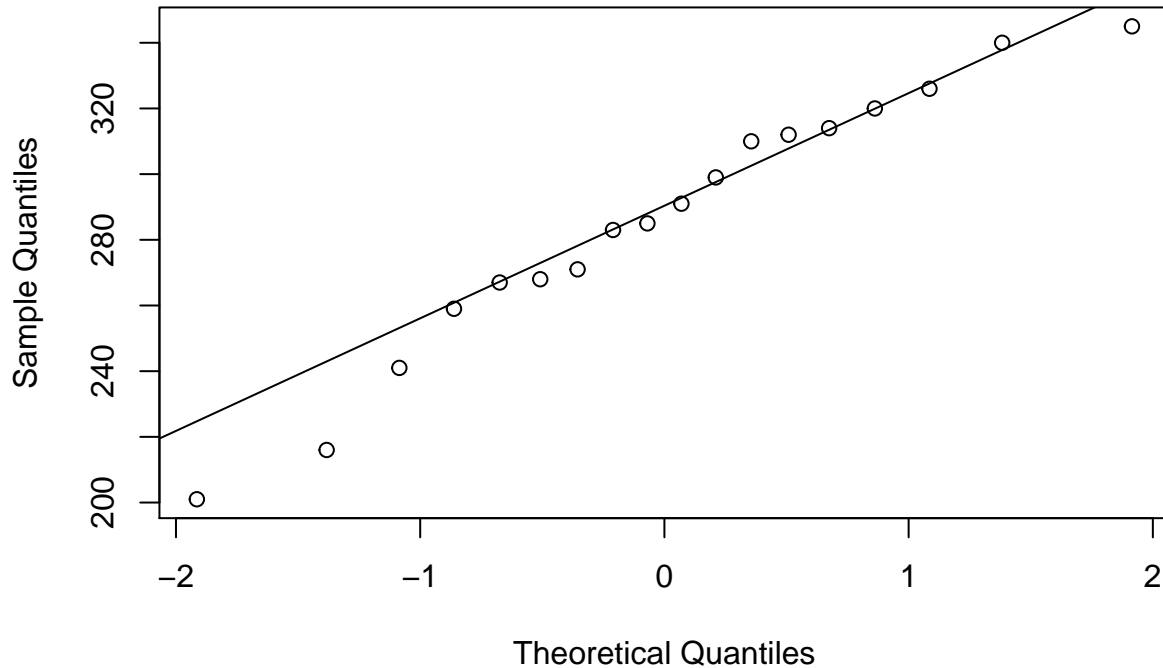
```
hist(sq_fcm_data$fcm)
```


Histogram of sq_fcm_data$fcm

```
qqnorm(sq_fcm_data$fcm)
qqline(sq_fcm_data$fcm)
```

Normal Q–Q Plot

The data don't look too bad. We can also test for normality using a Shapiro-Wilk Test

```
shapiro.test(sq_fcm_data$fcm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sq_fcm_data$fcm
## W = 0.96126, p-value = 0.6261
```

So the P-value is again not significant, which means that we do not reject the null hypothesis that the data are normally distributed.

Permutation Test

OK so maybe these data aren't the best example then for a permutation test... Oh well. In any case, let's move on and pretend as though one of the assumptions was not met and we wanted to pursue a permutation test as an alternative approach.

Step 1 Calculate the Test Statistic of Interest

In this case, what we care about is the difference between means in our treatment and control samples. So we will measure that directly.

```
mean(subset(sq_fcm_data$fcm, sq_fcm_data$treatment=="experimental"))-mean(subset(sq_fcm_data$fcm, sq_fc
```

```
## [1] 39.55556
```

So the observed difference in fecal cortisol metabolite levels between experimental and control squirrels was 39.56 ng/g.

Step 2 Permute the Data To Break The Hypothesized Relationship

Recall that our null hypothesis was that there was no difference in means between the experimental and control squirrels. If this were true, then it shouldn't matter if we reshuffled the experimental and control labels in the 'treatment' variable. This is what I mean by 'breaking' the hypothesized relationship. We can use the *sample()* command to do this.

```r
# I am first going to save a copy of the data as a new name that we will use for our reshuffling
null_fcm<-sq_fcm_data

null_fcm$treatment<-sample(null_fcm$treatment)
print(null_fcm)
```

```
##          treatment fcm
## 1   experimental 283
## 2   experimental 312
## 3   experimental 291
## 4   experimental 259
## 5        control 216
## 6        control 201
## 7   experimental 267
## 8        control 326
## 9   experimental 241
## 10       control 314
## 11 experimental 320
## 12       control 310
## 13 experimental 340
## 14       control 299
## 15       control 268
## 16       control 345
## 17       control 271
## 18 experimental 285
```

Notice how the fcm data have remained the same, but we have now randomly reshuffled the treatment labels. We can do this again to convince ourselves that this was a random reshuffle.

```r
null_fcm$treatment<-sample(null_fcm$treatment)
print(null_fcm)
```

```
##          treatment fcm
## 1        control 283
## 2        control 312
## 3   experimental 291
## 4        control 259
## 5        control 216
## 6   experimental 201
## 7   experimental 267
## 8   experimental 326
## 9   experimental 241
## 10       control 314
## 11       control 320
## 12 experimental 310
## 13       control 340
## 14 experimental 299
## 15       control 268
## 16 experimental 345
```

```
## 17      control 271
## 18 experimental 285
```

In this dataset the key relationship between treatment and fcm is now 'broken'

We could then recalculate our difference between means (now that we have forced the null hypothesis to be true).

```r
exp_mean<-mean(subset(null_fcm$fcm, null_fcm$treatment=="experimental"))
control_mean<-mean(subset(null_fcm$fcm, null_fcm$treatment=="control"))

exp_mean-control_mean
```

```
## [1] -2
```

Step 3 Repeat this Permutation Many Times to Create a Null Distribution

I will use a for loop to create 10,000 iterations of this reshuffling process.

```r
permuted_fcm<-1:10000

for (i in 1:10000){
  null_fcm<-sq_fcm_data
  null_fcm$treatment<-sample(null_fcm$treatment)
  exp_mean<-mean(subset(null_fcm$fcm, null_fcm$treatment=="experimental"))
  control_mean<-mean(subset(null_fcm$fcm, null_fcm$treatment=="control"))
  permuted_fcm[i]<-exp_mean - control_mean

}
```

We can visualize this empirical distribution of null differences between means using a histogram

```r
hist(permuted_fcm, main="Permuted distribution of\n difference between means", xlab="permuted difference
abline(v= 39.556)
abline(v= -39.556)
```

Permuted distribution of difference between means
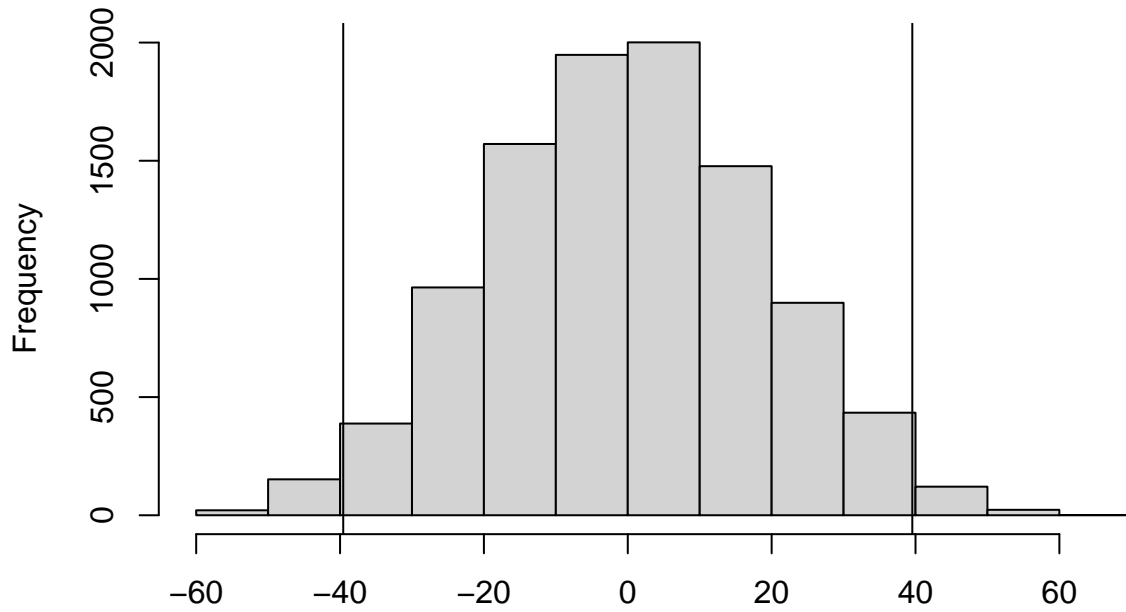
permuted difference between means

I have added vertical lines corresponding to the observed difference between sample means at both the positive and negative tail of this distribution.

Step 4 Calculate the P-Value as the Number of Iterations that Exceed the Observed Difference

```
length(subset(permuted_fcm, permuted_fcm>=39.556|permuted_fcm<=-39.556))/10000
```

```
## [1] 0.0339
```

You can see that this P-value that we calculated based on our permutation test is quite similar to what we calculated when we assumed normality and equal variance and a t-distribution of T test statistics.

An Example for You to Try - Resistance to TTX in Garter Snakes

The skin of rough-skinned newts stores an extremely potent neurotoxin called tetrodotoxin (TTX) deters many predators. In some geographic areas, garter snakes have evolved resistance to this toxin, which is measured as the degree of reduction in their crawl speed following an injected dose of TTX. Here we seek to compare whether resistance in snakes differs between two populations in Oregon.

H0: The TTX resistance for snakes from Benton is the same as for snakes from Warrenton.

HA: The TTX resistance for snakes from Benton is different from snakes from Warrenton.

TTX

This variable is known to be non-normally distributed.

There are a total of

Create the Data

```
locality<-c(rep("Benton", 7), rep("Warrenton", 5))
resistance<-c(0.29, 0.77, 0.96, 0.64, 0.70, 0.99, 0.34, 0.17, 0.28, 0.20, 0.20, 0.37)
ttx<-data.frame(locality, resistance)

print (ttx)
```

```
##       locality resistance
## 1       Benton       0.29
## 2       Benton       0.77
## 3       Benton       0.96
## 4       Benton       0.64
## 5       Benton       0.70
## 6       Benton       0.99
## 7       Benton       0.34
## 8    Warrenton       0.17
## 9    Warrenton       0.28
## 10   Warrenton       0.20
## 11   Warrenton       0.20
## 12   Warrenton       0.37
```

There are a total of 12 observations of TTX resistance from two localities (Benton and Warrenton)

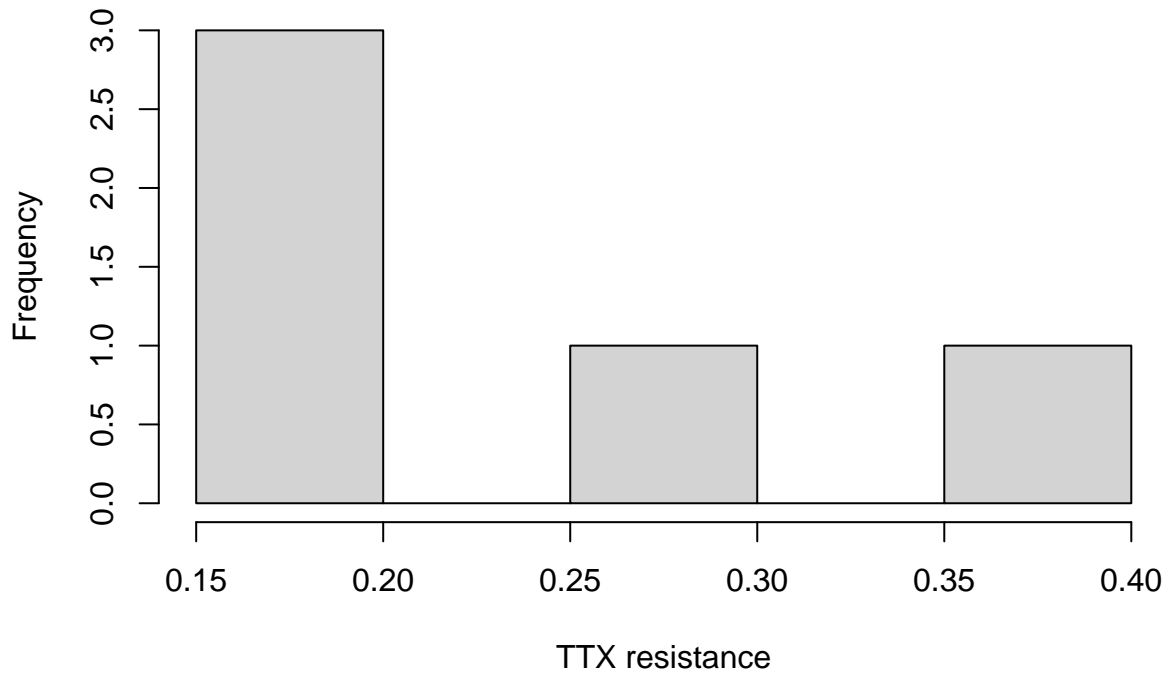## Graphically Assess the Assumption of Normality

This is best done by assessing the normality within each sample separately

Warrenton

```
#Warrenton
hist(subset(ttx$resistance, ttx$locality=="Warrenton"), main="Warrenton", xlab="TTX resistance")
```
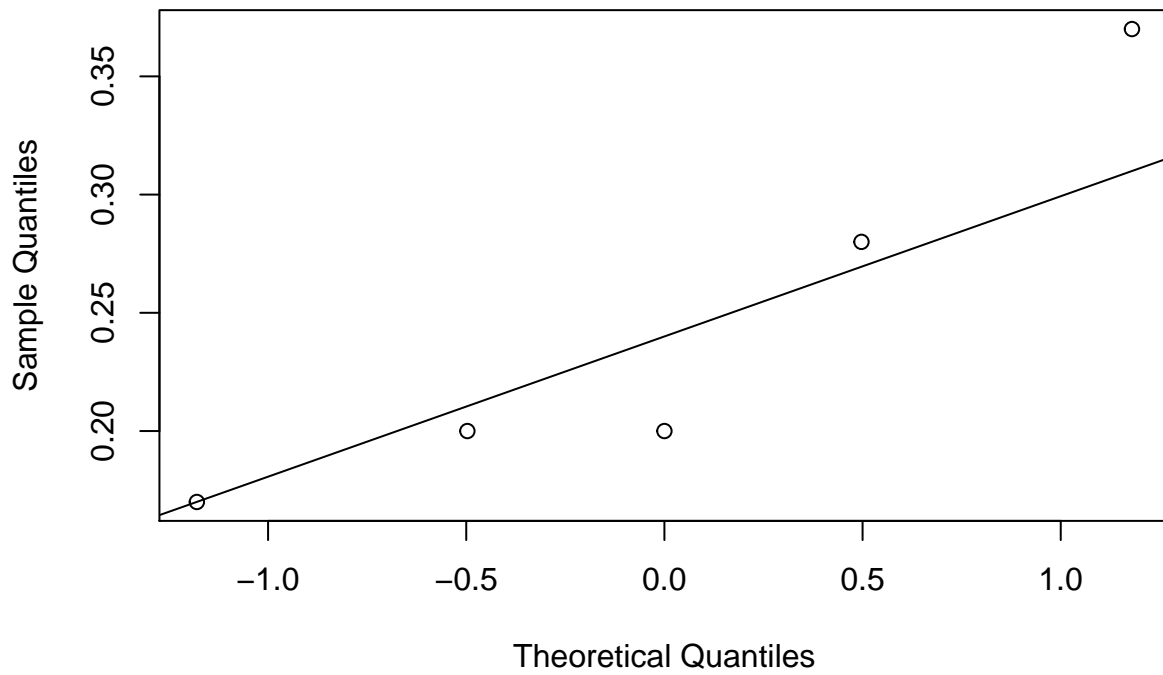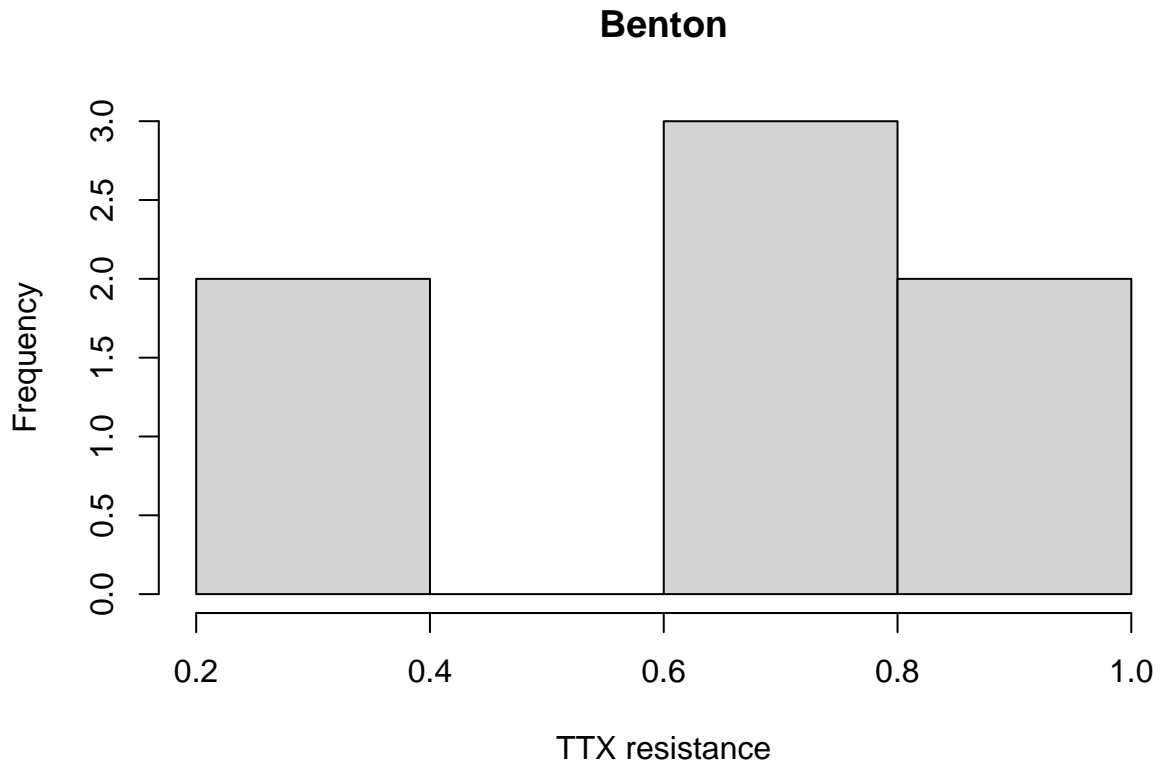
Warrenton

TTX resistance

```r
qqnorm(subset(ttx$resistance, ttx$locality=="Warrenton"), main="Q-Q plot TTX for Warrenton")
qqline(subset(ttx$resistance, ttx$locality=="Warrenton"))
```
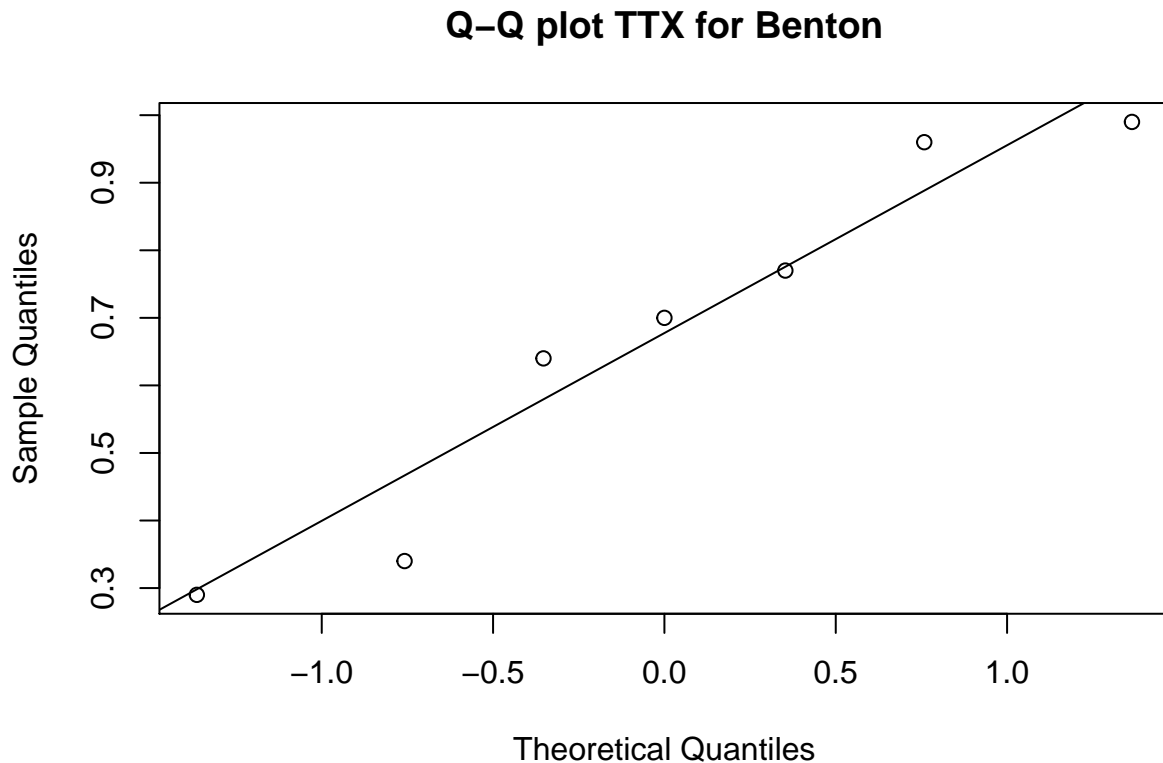
Q–Q plot TTX for Warrenton

Theoretical Quantiles

Benton

```
#Benton
hist(subset(ttx$resistance, ttx$locality=="Benton"), main="Benton", xlab="TTX resistance")
```

Benton


```
qqnorm(subset(ttx$resistance, ttx$locality=="Benton"), main="Q-Q plot TTX for Benton")
qqline(subset(ttx$resistance, ttx$locality=="Benton"))
```

Q–Q plot TTX for Benton

Shapiro-Wilk Tests

```
shapiro.test(subset(ttx$resistance, ttx$locality=="Warrenton"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(ttx$resistance, ttx$locality == "Warrenton")
## W = 0.87391, p-value = 0.2826
```

```
shapiro.test(subset(ttx$resistance, ttx$locality=="Benton"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(ttx$resistance, ttx$locality == "Benton")
## W = 0.9138, p-value = 0.4228
```

So neither are 'significantly' non-normal, but they really don't look very normally distributed. This is a case where we likely lack reasonable power to detect non-normality.

So lets proceed with a Permutation Test with these data.

Your Task

Write a permutation test to create a null distribution of differences between the mean TTX resistance in Warrenton and Benton snakes and use this null distribution to calculate a P-value from this permutation test to be able to test our hypothesis above.