Comparing sub-populations

Contents

4.2 Comparing Sub-Populations

4.2.0 Preamble and Intuition

- Oftentimes, interest lies in comparing two or more sub-populations (i.e., $\mathcal{P}_1$ vs. $\mathcal{P}_2$).
  - e.g., treatment vs. control arms of a clinical trial
  - e.g., two different versions of a webpage in online A/B testing
  - e.g., the Shark encounters that occurred in Australian vs. US waters

- Suppose we have the population $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$ and interest lies in comparing some attribute across the two sub-populations:

$$a(\mathcal{P}_1) \text{ vs. } a(\mathcal{P}_2)$$

  - We could compare the two sub-population attributes by way of a difference:

$$a(\mathcal{P}_1) - a(\mathcal{P}_2)$$

  - We could compare the two sub-population attributes by way of a ratio:

$$\frac{a(\mathcal{P}_1)}{a(\mathcal{P}_2)}$$

  - If the attribute is graphical (a histogram or quantile plot, for example) we could compare the two sub-populations by displaying the figures beside one another or overlaying them on top of one another.

Note: that when the attribute is a **measure of location**, sub-population comparisons are typically based on differences and when the attirbute us a **measure of spread**, comparisons are typically based on ratios.

Motivating Example: Comparing Australian vs. American Shark Encounters

- Let's load the data and numerically compare the sharks lengths in the two populations:

```
sharks <- read.csv("/Users/nstevens/Dropbox/Teaching/STAT_341/Lectures/Data/sharks.csv",
                   header = TRUE)

pop <- list(pop1 = subset(sharks, sharks$Australia == 1),
            pop2 = subset(sharks, sharks$USA == 1))

summary(pop[[1]]$Length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    36.0   119.5   164.0   155.9   193.0   240.0
```

```
summary(pop[[2]]$Length)
```

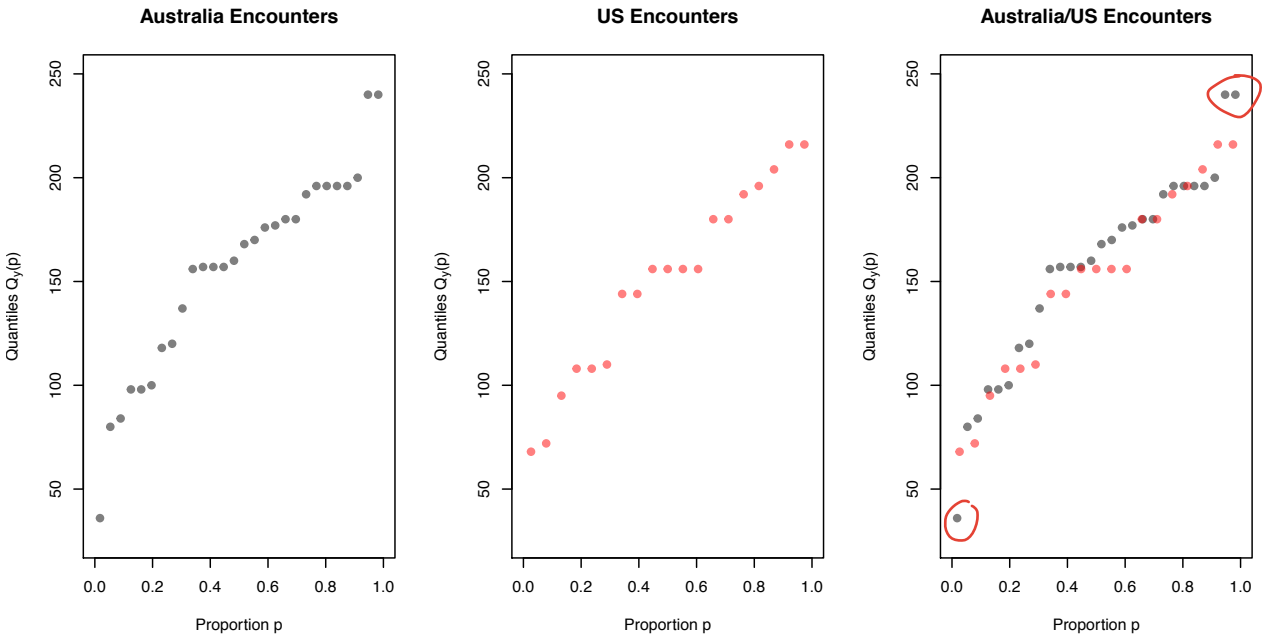
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    68.0   109.0   156.0   150.4   186.0   216.0
```

- Let's also graphically compare the shark lengths in the two populations using a quantile plot:

```r
par(mfrow=c(1,3),oma=c(0,0,2,0))

qvals <- sort(pop[[1]]$Length)
pvals <- ppoints(length(qvals))
plot(pvals, qvals, pch = 19, col=adjustcolor("black", alpha = 0.5),
     xlim=c(0,1), ylim=extendrange(range(sharks$Length)),
     xlab = "Proportion p",
     ylab = bquote("Quantiles Q"["y"]*"(p)"),
     main = "Australia Encounters")
qvals <- sort(pop[[2]]$Length)
pvals <- ppoints(length(qvals))
plot(pvals, qvals, pch = 19, col=adjustcolor("red", alpha = 0.5),
     xlim=c(0,1), ylim=extendrange(range(sharks$Length)),
     xlab = "Proportion p",
     ylab = bquote("Quantiles Q"["y"]*"(p)"),
     main = "US Encounters")

qvals <- sort(pop[[1]]$Length)
pvals <- ppoints(length(qvals))
plot(pvals, qvals, pch = 19, col=adjustcolor("black", alpha = 0.5),
     xlim=c(0,1), ylim=extendrange(range(sharks$Length)),
     xlab = "Proportion p",
     ylab = bquote("Quantiles Q"["y"]*"(p)"),
     main = "Australia/US Encounters")
qvals <- sort(pop[[2]]$Length)
pvals <- ppoints(length(qvals))
points(pvals, qvals, pch = 19, col=adjustcolor("red", alpha = 0.5) )
```



Comparing figures (like those above) tends to require some subjectivity. When comaring sub-populations we most often perform numerical comparisons.

- The difference between means in these sub-populations is:

$$a\left(\mathcal{P}_1\right) - a\left(\mathcal{P}_2\right) = \overline{y}_1 - \overline{y}_2$$

```
mean(pop$pop1[,"Length"]) - mean(pop$pop2[,"Length"])
```

```
## [1] 5.524436
```

- The ratio between standard deviations in these sub-populations is:

$$\frac{a\left(\mathcal{P}_1\right)}{a\left(\mathcal{P}_2\right)} = \frac{SD(\mathcal{P}_1)}{SD(\mathcal{P}_2)}$$

```
sd(pop$pop1[,"Length"]) / sd(pop$pop2[,"Length"])
```

```
## [1] 1.056418
```

Are these differences large?

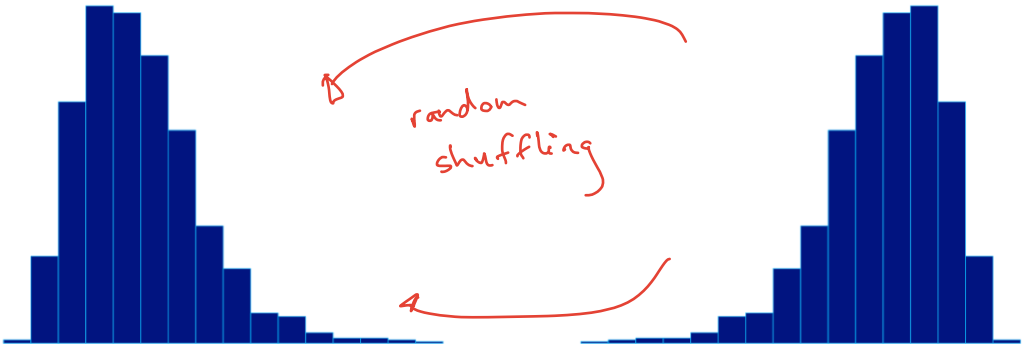
Randomly Mixing Sub-Populations

- If the two sub-populations are essentially the same
  - then the sub-populations observed should not look too different if we were to mix them up with one another

  in other words, swapping units would not dramatically change the features of the resulting sub-populations.

- On the other hand, if the two sub-populations were very different

  then shuffling the units could dramatically change the features of the resulting sub-populations.



- Here we combine the two sub-populations together into one $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$ and then
  - randomly draw two new sub-populations $\mathcal{P}_1^\star$ and $\mathcal{P}_2^\star$
  - ensuring that the sub-population sizes are kept the same

- Here is a function that does this:

```
mixRandomly <- function(pop) {
  pop1 <- pop$pop1
  n_pop1 <- nrow(pop1)

  pop2 <- pop$pop2
  n_pop2 <- nrow(pop2)

  mix <- rbind(pop1,pop2)
  select4pop1 <- sample(1:(n_pop1 + n_pop2),
                        n_pop1,
                        replace = FALSE)

  new_pop1 <- mix[select4pop1,]
  new_pop2 <- mix[-select4pop1,]
  list(pop1=new_pop1, pop2=new_pop2)
}
```

- We then compare the attributes of $\{\mathcal{P}_1, \mathcal{P}_2\}$ with $\{\mathcal{P}_1^\star, \mathcal{P}_2^\star\}$:
  - e.g., $a(\mathcal{P}_1)$ to $a(\mathcal{P}_1^\star)$, $a(\mathcal{P}_2)$ to $a(\mathcal{P}_2^\star)$, or
  - e.g., $a(\mathcal{P}_1) - a(\mathcal{P}_2)$ to $a(\mathcal{P}_1^\star) - a(\mathcal{P}_2^\star)$, or
  - e.g., $\frac{a(\mathcal{P}_1)}{a(\mathcal{P}_2)}$ to $\frac{a(\mathcal{P}_1^\star)}{a(\mathcal{P}_2^\star)}$, or
  - some other measure of difference among the sub-populations.

- If the sub-populations were similar to begin with, there shouldn't be a very large difference between attributes calculated on $\{\mathcal{P}_1, \mathcal{P}_2\}$ versus those calculated on $\{\mathcal{P}_1^\star, \mathcal{P}_2^\star\}$.

Do shark lengths differ in Australia vs. USA?

Let's re-examine the Shark example previously introduced (in which shark length was the variate of interest):

- First let's define some factory functions that will be helpful for calculating differences of means and ratios of standard deviations:

```
getAveDiffsFn <- function(variate) {
  function(pop) {mean(pop$pop1[, variate]) - mean(pop$pop2[,variate])}
}

getSDRatioFn <- function(variate) {
  function(pop) {sd(pop$pop1[, variate])/sd(pop$pop2[, variate])}
}
```

- For shark length we use the factory functions above to define the following:

```
diffAveLengths <- getAveDiffsFn("Length")
ratioSDLengths <- getSDRatioFn("Length")
```

- Recall that the difference in averages and ratio of the standard deviations between the two sub-populations are respectively:

$$\overline{y}_1 - \overline{y}_2 \quad \text{and} \quad \frac{SD(\mathcal{P}_1)}{SD(\mathcal{P}_2)}$$

```
round(c(diffAveLengths(pop), ratioSDLengths(pop)),3)
```

```
## [1] 5.524 1.056
```

- Randomly mixing the two sub-populations (while maintaining $N_1 = 28$ and $N_2 = 19$) and calculating these same summaries yields

$$\overline{y}_1^\star - \overline{y}_2^\star \quad \text{and} \quad \frac{SD(\mathcal{P}_1^\star)}{SD(\mathcal{P}_2^\star)}$$

```
set.seed(341)
mixedPop <- mixRandomly(pop)
round(c(diffAveLengths(mixedPop), ratioSDLengths(mixedPop)),3)
```

```
## [1] -18.152   0.928
```

- It seems that the standard deviation does not change much under shuffling, but the mean can change quite a lot.

- Are these differences unusual?
  - What would have happened if our random mixing was slightly different (i.e., we obtained a *different* random partition)?
  - To make this determination we need a more rigorous statistical analysis... we could perform a hypothesis test...

- We will discuss such tests from a computational perspective:
  - Let's look at many such shufflings $\{\mathcal{P}_1^\star, \mathcal{P}_2^\star\}$

– And each time calculate $\overline{y}_1^\star - \overline{y}_2^\star$ and $\frac{SD(\mathcal{P}_1^\star)}{SD(\mathcal{P}_2^\star)}$

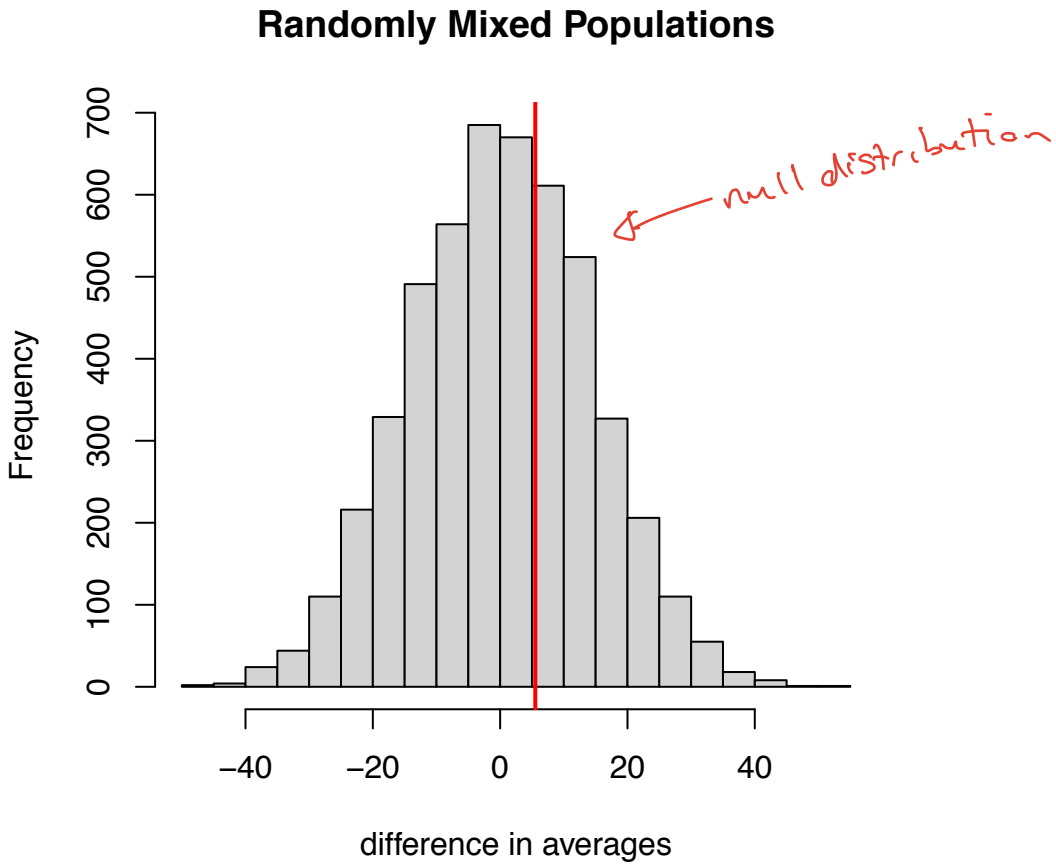
– Then let's make a histogram for each of these statistics and observe where $\overline{y}_1 - \overline{y}_2$ and $\frac{SD(\mathcal{P}_1)}{SD(\mathcal{P}_2)}$ lie

– There are $\binom{N_1+N_2}{N_1} = \binom{28+19}{19} = 6.97 \times 10^{12}$ potential rearrangments, but this is too many to consider. We'll use 5,000 rearrangement instead.

- First for the **averages**:

```
set.seed(341)
diffLengths <- sapply(1:5000,
                      FUN = function(...){diffAveLengths(mixRandomly(pop))})

hist(diffLengths, breaks=20,
     main = "Randomly Mixed Populations", xlab="difference in averages",
     col="lightgrey")
abline(v=diffAveLengths(pop), col = "red", lwd=2)
```



- The red line represents the difference between the average shark lengths in each of the sub-populations

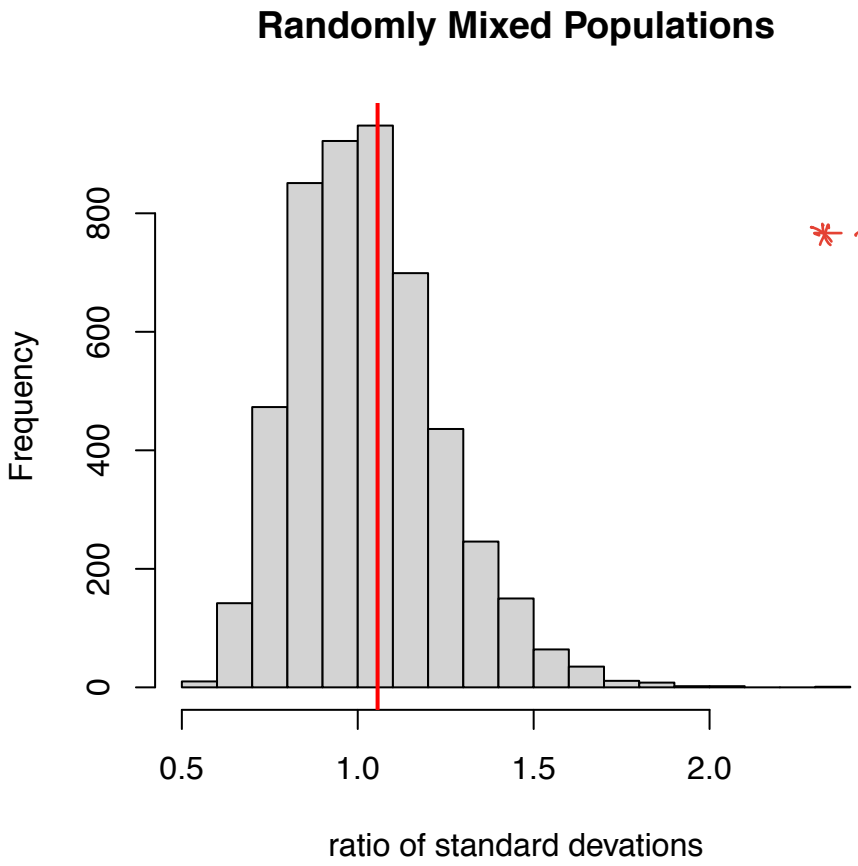
$$a(\mathcal{P}_{Australia}) - a(\mathcal{P}_{USA})$$

– Does $a(\mathcal{P}_{Australia}) - a(\mathcal{P}_{USA})$ seem to be *extreme* relative to the randomly mixed differences $a(\mathcal{P}_1^\star) - a(\mathcal{P}_2^\star)$? No, it seems fairly typical.

- Next for the **standard deviations**:

```r
set.seed(341)
ratioLengths <- sapply(1:5000,
                       FUN = function(...){ratioSDLengths(mixRandomly(pop))})

hist(ratioLengths, breaks=20,
     main = "Randomly Mixed Populations", xlab="ratio of standard devations",
     col="lightgrey")
abline(v=ratioSDLengths(pop), col = "red", lwd=2)
```



* these two findings suggest that $P_{Australia}$ and P_{USA} are not very different with respect to means and standard deviations.

- The red line represents the ratio of the standard deviation of shark lengths in each of the sub-populations

$$\frac{SD(\mathcal{P}_{Australia})}{SD(\mathcal{P}_{USA})}$$

  - Does $SD(\mathcal{P}_{Australia})/SD(\mathcal{P}_{USA})$ seem to be *extreme* relative to the randomly mixed ratios $SD(\mathcal{P}_1^\star)/SD(\mathcal{P}_2^\star)$? No, it seems pretty typical

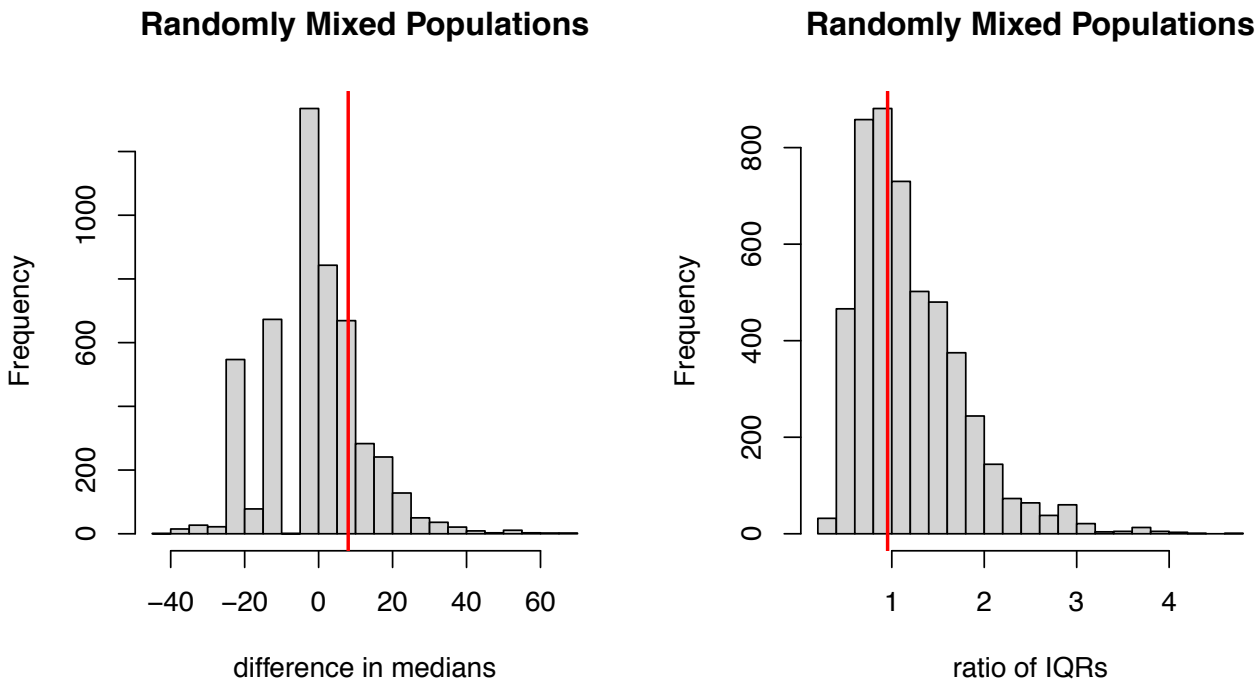
- We can also compare these sub-populations on the basis of their **medians** and **interquartile ranges**.
  - But first let's define some helpful functions like we did before:

```r
getMedianDiffsFn <- function(variate) {
  function(pop) {median(pop$pop1[, variate]) - median(pop$pop2[,variate])}
}

getIQRRatioFn <- function(variate) {
  function(pop) {IQR(pop$pop1[, variate])/IQR(pop$pop2[, variate])}
```

```
}
diffMedianLengths <- getMedianDiffsFn("Length")
ratioIQRLengths <- getIQRRatioFn("Length")
```

- Now, randomly mix the populations and summarize the results with a histogram.



- In both cases the red line represents the discrepancy measure calculated on the observed sub-populations $\{\mathcal{P}_{Australia}, \mathcal{P}_{USA}\}$
- What have we learned from these plots?

It appears as though shark lengths in Australia vs. USA are not that different. But, this conclusion should be taken tentatively because we've only explored 4 attributes.

Do shark lengths differ in fatal vs. non-fatal encounters?

- The two sub-populations are $\mathcal{P}_1$ = Fatal shark encounters and $\mathcal{P}_2$ = Non-Fatal shark encounters
  - And the variate of interest is again shark length
  - The code below constructs the two sub-populations:

```
Fatpop <- list(pop1 = sharks[sharks[,"Fatality"] ==1, ],
               pop2 = sharks[sharks[,"Fatality"] ==0, ])
```

- We can numerically compare the sharks lengths from the two populations:

```
summary(Fatpop[[1]]$Length)
```

Fatal

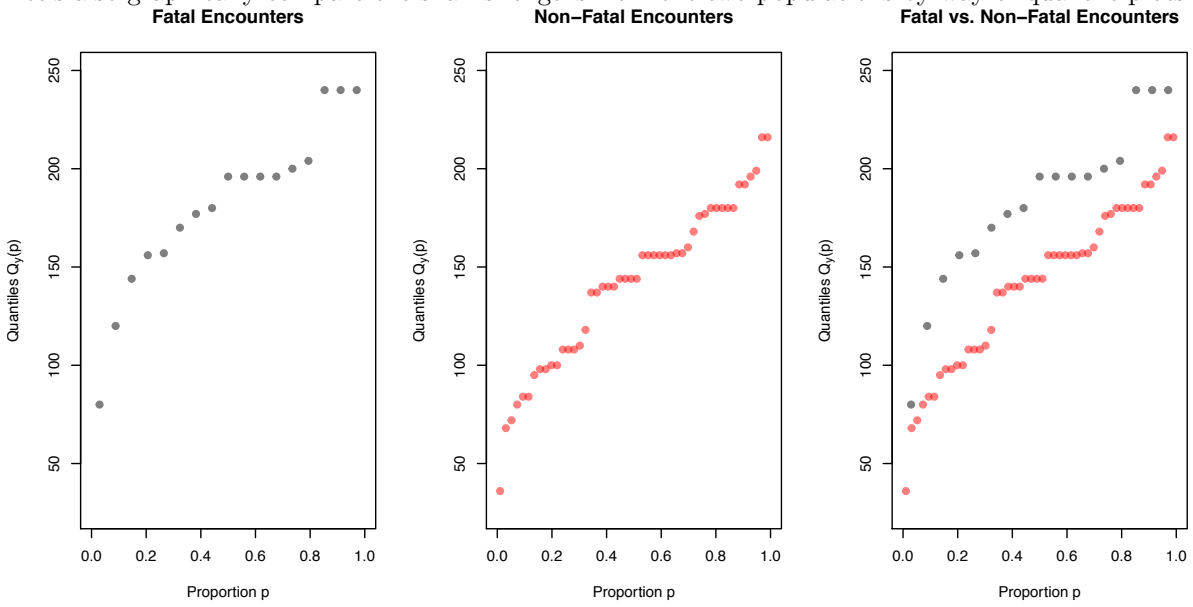
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    80.0   157.0   196.0   181.9   200.0   240.0
```

```
summary(Fatpop[[2]]$Length)
```

Non-Fatal

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    36.0   108.0   144.0   141.2   176.2   216.0
```

- Let's also graphically compare the sharks lengths from the two populations by way of quantile plots:



- It seems that fatal encounters involve bigger sharks compared to non-fatal encounters. (Why?)

- For shark encounters involving fatal and non-fatal encounters we quantify the difference in the average and standard deviation of the shark lengths from the two sub-populations by randomly mixing the sub-populations.

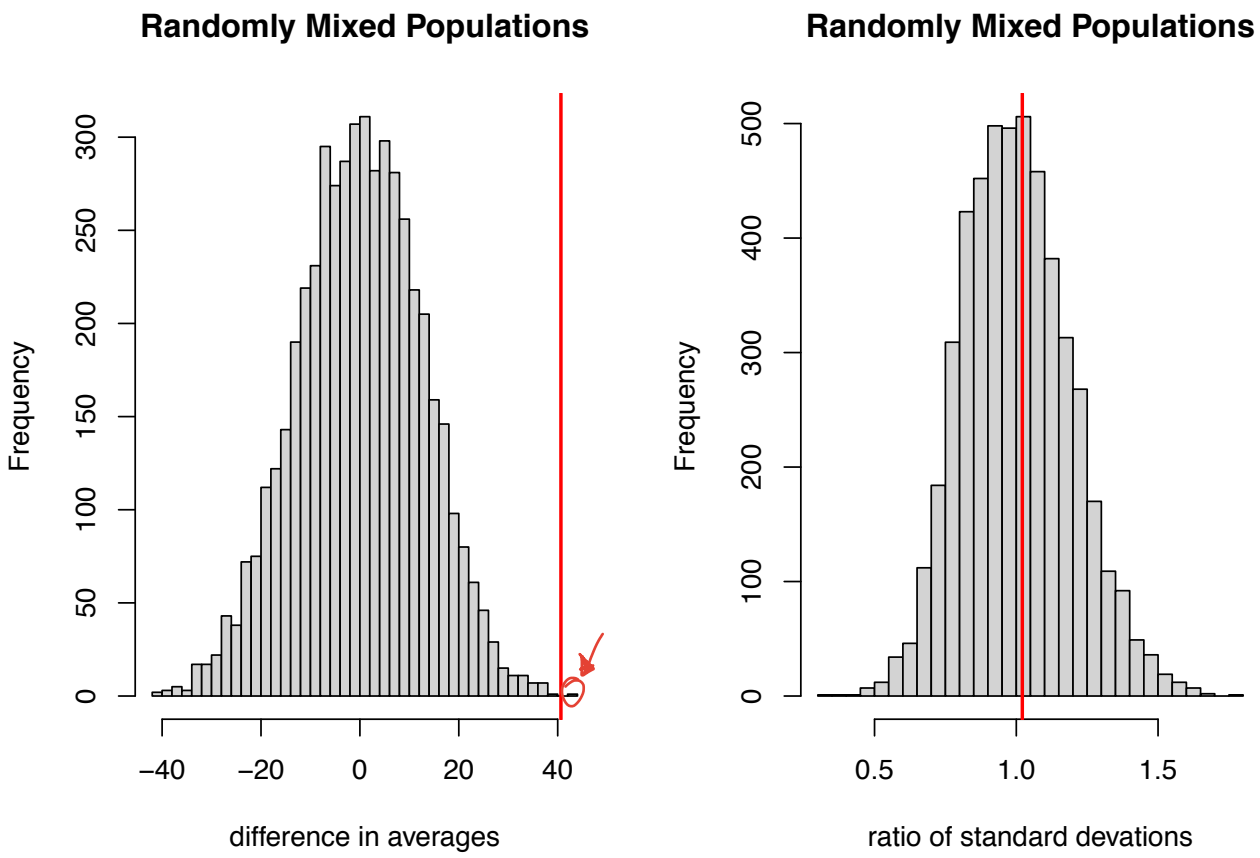
```
Fatpop <- list(pop1 = sharks[sharks[,"Fatality"] ==1, ],
               pop2 = sharks[sharks[,"Fatality"] ==0, ])

par(mfrow=c(1,2),oma=c(0,0,2,0))

set.seed(341)
fatpair <- sapply(1:5000,
   FUN = function(...){
     tmixpop = mixRandomly(Fatpop)
     c( diffAveLengths(tmixpop), ratioSDLengths(tmixpop))  })

hist(fatpair[1,], breaks="FD",
     main = "Randomly Mixed Populations", xlab="difference in averages",
     col="lightgrey")
abline(v=diffAveLengths(Fatpop), col = "red", lwd=2)

hist(fatpair[2,], breaks="FD",
     main = "Randomly Mixed Populations", xlab="ratio of standard devations",
     col="lightgrey")
abline(v=ratioSDLengths(Fatpop), col = "red", lwd=2)
```



- What do these plots say about the lengths of sharks involved in fatal vs. non-fatal encounters?

We see that there appears to be a significant difference in average shark lengths, but variability in shark lengths is fairly similar.

4.2.1 Anatomy of a Significance Test

We would like to quantify, numerically, how unusual the difference between $a(\mathcal{P}_1)$ and $a(\mathcal{P}_2)$ is relative to randomly mixed sub-populations.

- If the two sub-populations are actually similar, we want to provide numerical evidence in favour of the notion that the two sub-populations are similar to a randomly mixed sub-population.

- If the two sub-populations are actually different, we want to provide numerical evidence against the notion that the two sub-populations are similar to a randomly mixed sub-population.

The following steps are used to gather such evidence:

1. We suppose the sub-populations were randomly drawn from the same population. This is known as the **null hypothesis**.

2. We construct a **discrepency measure** that quantifies how inconsistent our data is with the null hypothesis

- where large values indicate evidence against the null hypothesis

3. We obtain the **observed discrepancy** by calculating
   - the discrepency measure on the two observed (i.e., unshuffled) sub-populations

4. Finally, we obtain the **observed p-value** by calculating
   - the probability that a randomly shuffled sub-population has a discrepency measure at least as large as the observed discrepancy
   - where small values indicate evidence against the null hypothesis

Let's elaborate on each of these steps.

* This approach to hypothesis testing is referred to as a "permutation test"

The Null Hypothesis

- Each of the following (equivalent) statements consitutes the null hypothesis we are testing.
  - H_0 : The sub-populations $\mathcal{P}_1$ and $\mathcal{P}_2$ were randomly drawn from the same population.
  - H_0 : $\mathcal{P}_1$ and $\mathcal{P}_2$ were created by randomly assigning units in the same population to one of the two sub-populations.
  - H_0 : $\mathcal{P}_1$ and $\mathcal{P}_2$ were generated by random mixing.

  H_0: P_1 and P_2 are indistinguishable
- Regardless of how the null hypothesis is stated, the **alternative hypothesis** H_A is the complement of H_0.

  H_A: P_1 and P_2 are distinguishable
- In the context of the shark data, we have
  - H_0 : $\mathcal{P}_{Australia}$ and $\mathcal{P}_{USA}$ are drawn from the same population of shark encounters.
  - H_A : $\mathcal{P}_{Australia}$ and $\mathcal{P}_{USA}$ are **not** drawn from the same population of shark encounters.

* Note that we do not state the null hypothesis in terms of the equivalence of attribute values, i.e. $a(\mathcal{P}_1) = a(\mathcal{P}_2)$.
  - Although such a statement is true if H_0 holds, it is *weaker* and so we avoid using it.

The Discrepency Measure

- A discrepency measure $D(\mathcal{P}_1, \mathcal{P}_2)$ quantifies how inconsistent our data is with the null hypothesis, and is defined so that large values indicate evidence against the null hypothesis.

- As a point of interest, the discrepancy measure is technically an attribute for the popuation $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$ and so we could consider properties such as equivariance and invariance.
- In other statistical texts the discrepancy measure is often referred to as a **test statistic**

- The form of $D(\mathcal{P}_1, \mathcal{P}_2)$ depends on *how* we want to compare $\mathcal{P}_1$ and $\mathcal{P}_2$

- If we want to compare measures of location the discrepancy measure is typically based on *differences*: $a(\mathcal{P}_1) - a(\mathcal{P}_2)$

- If we want to compare measures of spread the discrepancy measure is typically based on *ratios*: $\frac{a(\mathcal{P}_1)}{a(\mathcal{P}_2)}$

- For example:
  - if we hypothesized that the averages from the two sub-populations were the same, then a discrepency measure for this might be
$$D(\mathcal{P}_1, \mathcal{P}_2) = |\overline{y}_1 - \overline{y}_2|$$
  - if we hypothesized that the standard deviation from the two sub-populations were the same, then a discrepency measure for this might be
$$D(\mathcal{P}_1, \mathcal{P}_2) = \left| \frac{SD(\mathcal{P}_1)}{SD(\mathcal{P}_2)} - 1 \right|$$
  - if we hypothesized that the average from the first population was smaller than the average from the second population, then a discrepency measure for this might be
$$D(\mathcal{P}_1, \mathcal{P}_2) = \overline{y}_1 - \overline{y}_2$$
  - if we hypothesized that the average from the first population was larger than the average from the second population, then a discrepency measure for this might be
$$D(\mathcal{P}_1, \mathcal{P}_2) = \overline{y}_2 - \overline{y}_1$$

The Observed Discrepancy

- The observed discrepancy, d_{obs}, is the value of discrepancy measure D calculated on the two observed (i.e., unshuffled) sub-populations:
$$d_{obs} = D(\mathcal{P}_1, \mathcal{P}_2)$$

- It's important to recognize that the discrepancy measure quantifies only one type of discrepancy between the populations
  - e.g., discrepancy in averages
  - e.g., discrepancy in standard deviations

— or something else.

- All other differences are completely ignored.

For example P_1 and P_2 could have very similar means and standard deviations, but very different skewness coefficients.

The Observed p-value → The probability of observing a result at least as extreme as what we observed, if H_0 is true.

- The observed p-value is the probability that a randomly shuffled sub-population has a discrepency measure at least as large as the observed discrepancy

$$p\text{-value} = Pr\left(D \geq d_{obs} \,\middle|\, H_0 \text{ is true}\right)$$

- If the p-value is very small then either
  - the null hypothesis is true and we have observed a very unusual value of d_{obs}
  - OR the null hypothesis is false. → this is what we tend to believe.

- The smaller the p-value, the greater the evidence against the null hypothesis.
  - p-value < 0.001 means that there is **very strong evidence** against H_0
  - $0.001 < p$-value < 0.01 means that there is **strong evidence** against H_0
  - $0.01 < p$-value < 0.05 means that there is **evidence** against H_0
  - $0.05 < p$-value < 0.1 means that there is **weak evidence** against H_0
  - p-value > 0.1 means that there is **no evidence** against H_0
  - In the extreme case where p-value $= 0$, then we have observed something impossible and the hypothesis must therefore be false – this would be a proof by contradiction.

- In order to calculate the p-value *exactly* one must consider all $\binom{N_1+N_2}{N_1} = \binom{N_1+N_2}{N_2}$ possible permutations of the observed data → this would require knowing the null distribution exactly
  - The exact p-value is the fraction of $D(\mathcal{P}_1^\star, \mathcal{P}_2^\star)$ values greater than or equal to d_{obs}

- Because $\binom{N_1+N_2}{N_1} = \binom{N_1+N_2}{N_2}$ is in practice too many permutations to consider, we typically just use M (a large number) of them
  - In particular we generate M shuffled pairs:

$$(\mathcal{P}_{1,1}^\star, \mathcal{P}_{2,1}^\star), (\mathcal{P}_{1,2}^\star, \mathcal{P}_{2,2}^\star), \ldots, (\mathcal{P}_{1,M}^\star, \mathcal{P}_{2,M}^\star)$$

  - The p-value is then *approximated* as

$$\hat{p}\text{-value} = \frac{1}{M}\sum_{i=1}^{M} I_{[d_{obs},\infty)}\left(D(\mathcal{P}_{1,i}^\star, \mathcal{P}_{2,i}^\star)\right)$$

is $D\left(P_{1,i}^*, P_{2,i}^*\right) \in [d_{obs}, \infty)$? AKA is $D(P_{1,i}^*, P_{2,i}^*) \geq d_{obs}$?

  - If $M = \binom{N_1+N_2}{N_1} = \binom{N_1+N_2}{N_2}$ and we've considered all possible shuffles the equation above would yield the *exact* p-value

Putting it All Together

Test of Significance Algorithm

1. State the **null hypothesis**: H_0 :$\mathcal{P}_1$ and $\mathcal{P}_2$ are drawn from the same population.
2. Construct a measure of **discrepancy** $D = D(\mathcal{P}_1, \mathcal{P}_2)$ where large values indicate **evidence against the null hypothesis**
3. Calculate the **observed discrepancy** $d_{obs} = D(\mathcal{P}_1, \mathcal{P}_2)$.
4. Shuffle the sub-populations M times and calculate the **observed p-value**:

$$p\text{-value} = Pr\left(D \geq d_{obs} \,\middle|\, H_0 \text{ is true}\right) \approx \frac{1}{M}\sum_{i=1}^{M} I_{[d_{obs},\infty)}\left(D(\mathcal{P}^{\star}_{1,i}, \mathcal{P}^{\star}_{2,i})\right)$$

Shark Example

- In R, since we already have a sample of $M = 5000$ shuffled pairs, we can calculate the approximate p-value with `sum(abs(diffLengths) >= abs(diffAveLengths(pop))) / length(diffLengths)`
  - Note that this assumes a discrepancy measure of $D(\mathcal{P}_1, \mathcal{P}_2) = |\overline{y}_1 - \overline{y}_2|$

- For the null hypothesis

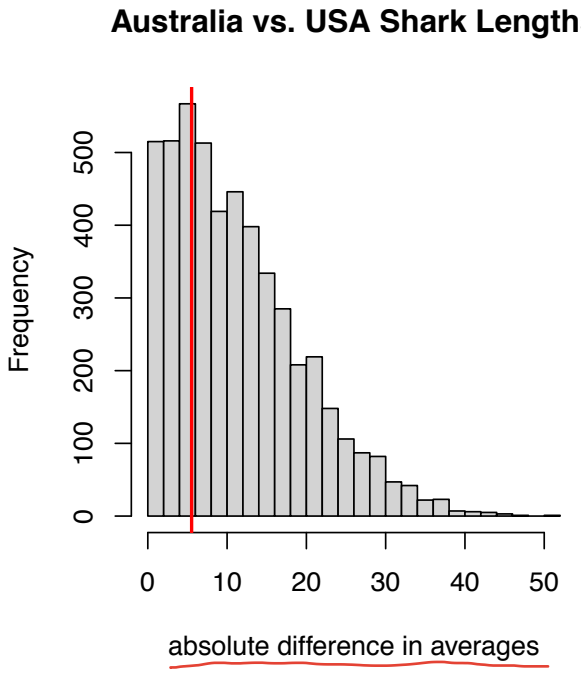
$H_0 : \mathcal{P}_{Australia}$ and $\mathcal{P}_{USA}$ are drawn from the same population of shark lengths

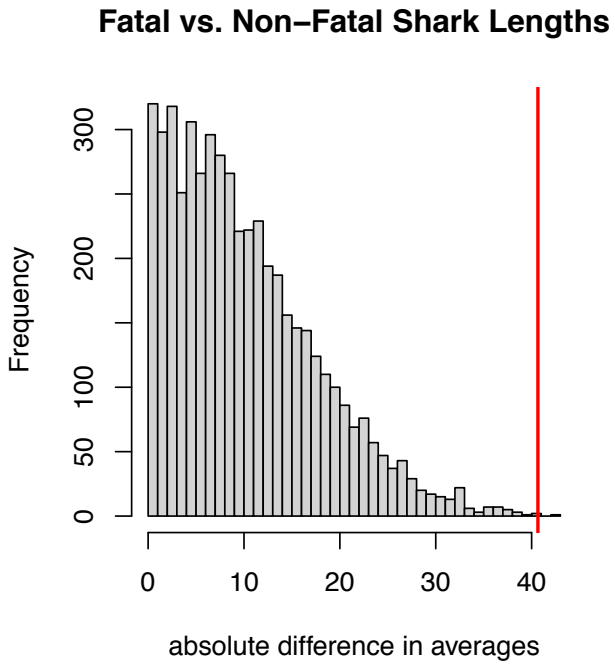
we find p-value ≈ 0.704

- For the null hypothesis

$H_0 : \mathcal{P}_{Fatal}$ and $\mathcal{P}_{Non-Fatal}$ are drawn from the same population of shark lengths

we find p-value ≈ 0.0002





- Suppose that the pair $(\mathcal{P}_{Australia}, \mathcal{P}_{USA})$ is a random draw:
  - then the probability of seeing at least as large a difference as we observed in $(\mathcal{P}_{Australia}, \mathcal{P}_{USA})$ is approximately 0.704
  - thus, there is **no evidence against the null hypothesis** that the pair $(\mathcal{P}_{Australia}, \mathcal{P}_{USA})$ was randomly drawn
  - in other words, we have no evidence against the hypothesis that the two populations $\mathcal{P}_{Australia}$ and $\mathcal{P}_{USA}$ are indistinguishable

  * as far as the average is concerned.
- Suppose that the pair $(\mathcal{P}_{Fatal}, \mathcal{P}_{Non-Fatal})$ is a random draw:
  - then the probability of seeing at least as large a difference as we observed in $(\mathcal{P}_{Fatal}, \mathcal{P}_{Non-Fatal})$ is approximately 0.0002
  - thus, there is **very strong evidence against the null hypothesis** that the pair $(\mathcal{P}_{Fatal}, \mathcal{P}_{Non-Fatal})$ was randomly drawn
  - in other words, we have very strong evidence against the hypothesis that the two populations $\mathcal{P}_{Fatal}$ and $\mathcal{P}_{Non-Fatal}$ are indistinguishable

  * as far as averages are concerned.

Significance Testing Errors

Courtroom Analogy

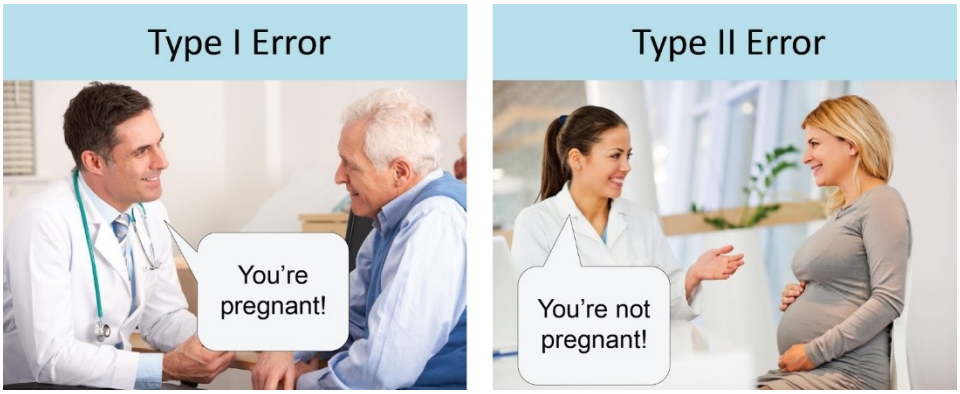
Decision	defendent is innocent	the defendent is guilty
Convicted	Error (Type I Error)	Correct
Acquitted	Correct	Error (Type II Error)

In Hypothesis Testing

Decision	H_0 is true	H_0 is false
Reject H_0	Error (Type I Error)	Correct
Do Not Reject H_0	Correct	Error (Type II Error)

Another Example

H_0: person is not pregnant vs. H_A: person is pregnant



A note on the language:

- We will try to avoid using terms such as "reject", "fail to reject", "accept", etc.
- Instead, we use the significance test simply to measure the evidence against H_0

Important Remarks

- The observed p-value provides a common (probabilistic) scale on which to measure the **evidence against the null hypothesis**

- The observed p-value does **not** measure evidence **in favour** of the null hypothesis.
  - in science, we try to falsify hypotheses and entertain only those which remain standing;

* Absence of evidence ≠ evidence of absence

- A test of significance therefore **neither accepts nor rejects a null hypothesis**; it simply provides a measure of the evidence against it
  - the decision taken in light of this evidence is the choice of the researcher

Equivalence Testing

- There is **no magic level** for a p-value such as 0.05 or 0.01,
  - there is no practical or scientific difference between $p = 0.048$ and $p = 0.051$, for example

- The fact that the evidence against the null hypothesis is **statistically significant** based on some discrepancy measure **does not imply that the discrepancy is practically significant**
  - the p-value measures how unusual a discrepancy of that size might be when the null hypothesis holds,
  - it says nothing about whether a discrepancy of that size matters for any practical or scientific purpose
  - e.g., for the shark lengths data, the average shark length in fatal vs. non-fatal encounters differed by 3.25 feet. Is that difference of practical importance?

- Every test of significance is based on some measure of discrepancy and **different discrepancy measures can detect different departures** from the null hypothesis, so one needs to understand the nature of the departure from the hypothesis that the discrepancy is trying to measure.

4.2.2 A t-like Discrepancy Measure

- When comparing two sub-populations on the basis of a measure of location, one particularly useful discrepancy measure is

$$D(\mathcal{P}_1, \mathcal{P}_2) = \frac{a(\mathcal{P}_1) - a(\mathcal{P}_2)}{SD\big[a(\mathcal{P}_1) - a(\mathcal{P}_2)\big]}$$

- This discrepancy measure is "physically dimensionless"
  - Whatever scale the numerator is measured in (e.g. inches as in the shark lengths), the scale of the denominator will match, leaving the ratio free of any measurement scale.
  - This naturally makes this discrepancy measure scale-invariant.

- The challenge is determining the denominator of the discrepancy measure.
  - In rare cases, the denominator might be known and then this discrepancy measure is a rescaling of $a(\mathcal{P}_1) - a(\mathcal{P}_2)$ and would not yield different results.
  - However, more commonly, we will estimate the denominator using information from $\mathcal{P}_1$ and $\mathcal{P}_2$.

- Suppose that the populations $\mathcal{P}_1$ and $\mathcal{P}_2$ are **independently** drawn from the **same** larger population. Then the denominator would instead be

$$\widetilde{SD}\left[a(\mathcal{P}_1) \;-\; a(\mathcal{P}_2)\right] = \sqrt{\widetilde{Var}[a(\mathcal{P}_1)] \;+\; \widetilde{Var}[a(\mathcal{P}_2)]}$$

  where
  - $\widetilde{SD}[\cdots]$ denotes an estimator of the standard deviation of its argument and
  - $\widetilde{Var}[\cdots]$ denotes an estimator of the variance of its argument.

- But determining the form of $\widetilde{SD}\left[a(\mathcal{P}_1) \;-\; a(\mathcal{P}_2)\right]$ can also be difficult
  - Except in the common special case when $a(\mathcal{P})$ is an average

Differences in Averages

Suppose we were interested in differences in averages:

- In this case $a(\mathcal{P}_i) = \overline{Y}_i$ and $\mathcal{P}_i$ has size N_i, $i = 1, 2$
- And the discrepancy measure becomes:

$$D(\mathcal{P}_1, \mathcal{P}_2) \quad = \quad \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{\widetilde{\sigma}^2}{N_1} + \frac{\widetilde{\sigma}^2}{N_2}}}$$

  where $\widetilde{\sigma}$ is an estimator of the standard deviation of the Y values in the population $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$.
  - If $\widetilde{\sigma}_1$ and $\widetilde{\sigma}_2$ denote the estimators of the standard deviations from each of $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively, then the pooled estimator of σ would be

$$\widetilde{\sigma} = \left(\frac{(N_1 - 1)\widetilde{\sigma}_1^2 + (N_2 - 1)\widetilde{\sigma}_2^2}{(N_1 - 1) + (N_2 - 1)}\right)^{\frac{1}{2}}$$

  - **Note:** if it were inappropriate to assume the variability in the two sub-populations was equivalent we could instead use the denominator

$$\sqrt{\frac{\widetilde{\sigma}_1^2}{N_1} \;+\; \frac{\widetilde{\sigma}_2^2}{N_2}}$$

This just looks like the t-test

- This *is* the "two-sample" Student t statistic used to test the equality of the means of two normal distributions with common (but unknown) standard deviation σ.

  - If the Y values were in fact normally distributed, the discrepancy measure would follow a Student t distribution with $n_1 + n_2 - 2$ degrees of freedom under the null hypothesis that the means were identical.

- Note, however, in our procedure of randomly mixing the populations we make **no such normality assumption**.
  - We simply proceed with *this* discrepancy measure just as we did with the earlier measures. The only difference is that now we need to first calculate the denomonator (the standard error).

- Below is a factory function that will return a function that calculates this discrepancy measure for any two sub-populations for a given variate `var` (assuming the variance of the sub-populations are equal)

```
### The t statistic

getDiscrepancyFn <- function(var) {
  function(pop) {
    ## First sub-population
    pop1 <- pop$pop1
    n1 <- nrow(pop1)
    m1 <- mean(pop1[, var])
    v1 <- var(pop1[, var])

    ## Second sub-population
    pop2 <- pop$pop2
    n2 <- nrow(pop2)
    m2 <- mean(pop2[, var])
    v2 <- var(pop2[, var])

    ## Pool the variances
    v <- ((n1 - 1) * v1 + (n2 - 1) * v2)/(n1 + n2 - 2)

    ## Determine the t-statistic
    t <- (m1 - m2) / sqrt(v * ( (1/n1) + (1/n2) ) )

    ## Return the t-value
    t
  }
}
```

Comparing Shark Lengths in Australia vs. USA

- Get the this t-like discrepancy measure for the variate `Length`

```
tStatLengths <- getDiscrepancyFn("Length")
```

- The t-like discrepancy measure calculated on the Australia vs. USA sub-populations is

```
tStatLengths(pop)
```

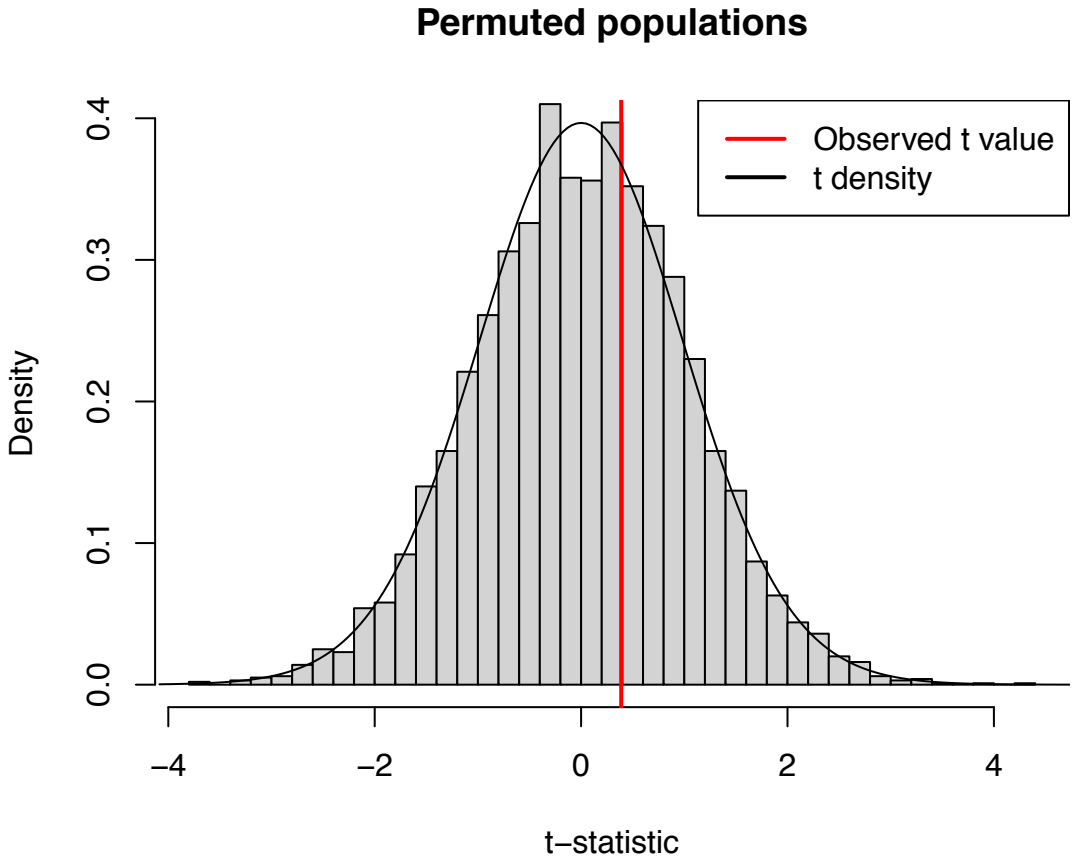
```
## [1] 0.3886752
```

- To gauge the size of the these discrepancy measures we
  - mix / shuffle / permute the sub-populations $M = 5000$ times and plot the histogram as before
  - for comparison we also overlay the Student t density (with $n_1 + n_2 - 2$ degrees of freedom) which is what we would use to calculate p-values if we were willing to make an assumption of normality

```
set.seed(341)
tVals <- sapply(1:5000, FUN = function(...){tStatLengths(mixRandomly(pop))})
xvals <- extendrange(tVals)
xvals <- seq(from = min(xvals), to = max(xvals), length.out = 200)

### We will overlay the histogram with the theoretical t-density
n1 <- nrow(pop$pop1)
n2 <- nrow(pop$pop2)
densityVals <-dt(xvals, df = (n1 +  n2 - 2))
histHeights <- hist(tVals, breaks=20, plot = FALSE)$density
heightRange <- c(0, max(densityVals, histHeights))

### Plot the histogram
hist(tVals, breaks=50, probability = TRUE,
     ylim = heightRange,
     main = "Permuted populations", xlab="t-statistic",
     col="lightgrey")
abline(v=tStatLengths(pop), col = "red", lwd=2)
### Add the density to the plot
lines(xvals, densityVals, col = "black")
legend("topright",
       legend=c("Observed t value", "t density"),
       lwd = c(2, 2), col = c("red", "black"))
```



- Remarkably, the Student t density closely approximates the histogram!
  - In many instances, even when no normality distribution is assumed, the Student t distribution will roughly approximate the histogram that arises from randomly mixing the sub-populations.
  - This in fact was one of the early justifications (by R.A. Fisher) for using the t distribution broadly in application; namely that it approximated the random mixing procedure.

- The p-value can now be estimated two ways:

1) Using the sub-populations generated via random mixing:

```
tobs = tStatLengths(list(pop1 = sharks[sharks[,"Australia"] ==1, ], pop2 = sharks[sharks[,"USA"] ==1, ])

mean(abs(tVals) >= abs(tobs) )
```

```
## [1] 0.704
```

2) Using the t-distribution with $n_1 + n_2 - 2$ degrees of freedom:

```
2*pt( abs(tobs), df = (n1 +  n2 - 2), lower.tail=FALSE)
```

```
## [1] 0.6993494
```

- The similarity of these two numbers shows that, in this particular example, the t-distribution could be a good approximation (though we are not using it!)
- The p-value is so large that the observed discrepancy measure is not at all unusual when the hypothesis is true.

- This test provides **no evidence against the null hypothesis**.
- From a practical point of view, this means that the two sub-populations Australia and USA are not different from perspective of average shark length.

Comparing Shark Lengths in Fatal vs. Non-Fatal Encounters

- The t-like discrepancy measure calculated on the Fatal vs. Non-Fatal sub-populations is

```
tStatLengths(Fatpop)
```

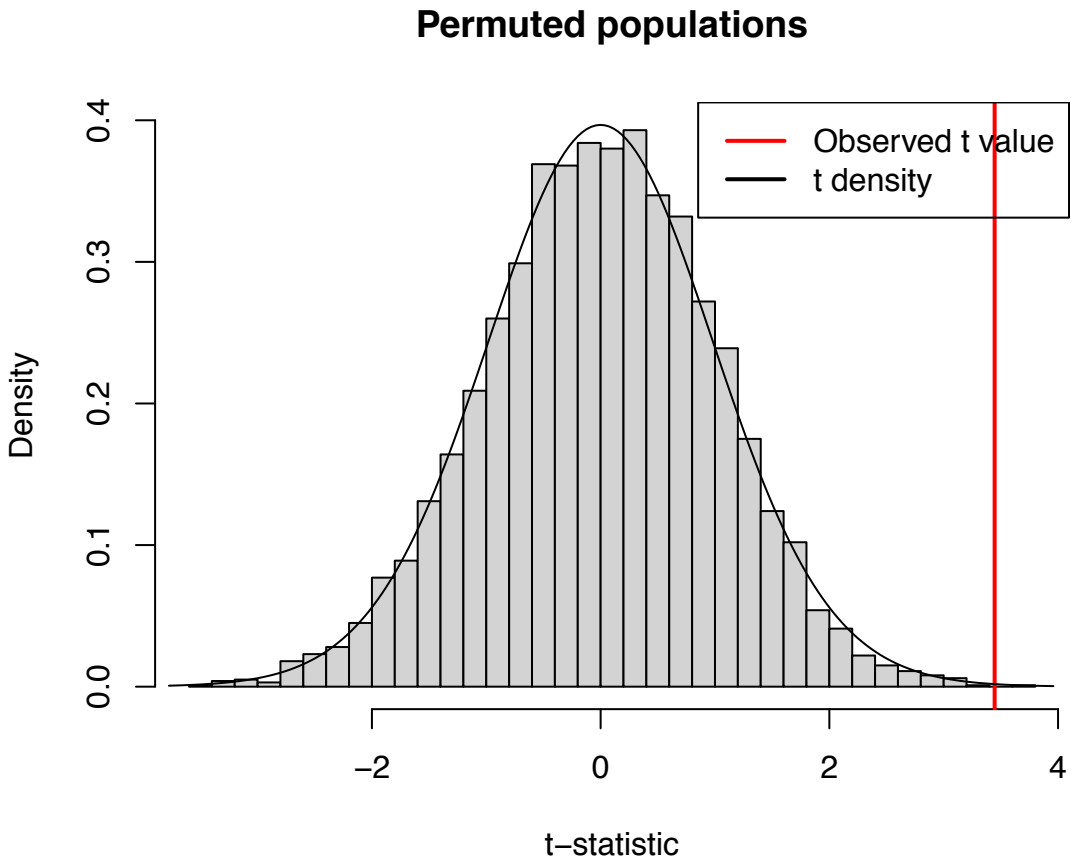
```
## [1] 3.445492
```

- To gauge the size of the these discrepancy measures we
  - mix / shuffle / permute the sub-populations $M = 5000$ times and plot the histogram as before
  - for comparison we also overlay the Student t density (with $n_1 + n_2 - 2$ degrees of freedom) which is what we would use to calculate p-values if we were willing to make an assumption of normality

```
set.seed(341)
tVals <- sapply(1:5000, FUN = function(...){tStatLengths(mixRandomly(Fatpop))})
xvals <- extendrange(tVals)
xvals <- seq(from = min(xvals), to = max(xvals), length.out = 200)

### We will overlay the histogram with the theoretical t-density
n1 <- nrow(pop$pop1)
n2 <- nrow(pop$pop2)
densityVals <-dt(xvals, df = (n1 +  n2 - 2))
histHeights <- hist(tVals, breaks=20, plot = FALSE)$density
heightRange <- c(0, max(densityVals, histHeights))

### Plot the histogram
hist(tVals, breaks=50, probability = TRUE,
     ylim = heightRange,
     main = "Permuted populations", xlab="t-statistic",
     col="lightgrey")
abline(v=tStatLengths(Fatpop), col = "red", lwd=2)
### Add the density to the plot
lines(xvals, densityVals, col = "black")
legend("topright",
       legend=c("Observed t value", "t density"),
       lwd = c(2, 2), col = c("red", "black"))
```



- Again, the Student t density appears to closely approximate the histogram.

- As before, the p-value can now be estimated two ways:

1) Using the sub-populations generated via random mixing:

```
tobs = tStatLengths(list(pop1 = sharks[sharks[,"Fatality"] ==1, ], pop2 = sharks[sharks[,"Fatality"] ==
SL.hat = mean(abs(tVals) >= abs(tobs))
sprintf("%.4f", SL.hat )
```

```
## [1] "0.0002"
```

2) Using the t-distribution with $n_1 + n_2 - 2$ degrees of freedom:

```
SL.that = 2*pt( abs(tobs), df = (n1 +  n2 - 2), lower.tail=FALSE)
sprintf("%.5f", SL.that )
```

```
## [1] "0.00125"
```

- Despite what we see in the figure above, the dissimilarity between these two values suggests that the t approximation may not be as good in this case.
  - This is because the approximation does not work well on the tails, and the p-value is, afterall, a tail probability

- The observed significance level is so small that the observed discrepancy measure is unusual when the hypothesis is true.

- This test provides **evidence against the null hypothesis**.
- From a practical point of view, this means that the two sub-populations (Fatal vs. Non-Fatal encounters) seem to be different from perspective of average shark length.

4.2.3 Multiple Testing

The null hypothesis we have been testing is the following:

H_0 : The sub-populations $\mathcal{P}_1$ and $\mathcal{P}_2$ were randomly drawn from the same population

Or, in other words:

H_0 : The sub-populations $\mathcal{P}_1$ and $\mathcal{P}_2$ are indistinguishable

- We have tested this hypothesis using a variety of discrepancy measures:
  - Comparisons of means
  - Comparisons of medians
  - Comparisons of standard deviations
  - Comparisons of interquartile ranges
  - ⋮
- In general, we could use any number of discrepancy measures $D_1, D_2, \ldots, D_K$ to compare sub-populations
  - Each with an associated p-value: $pv_1, pv_2, \ldots, pv_K$.
- However, when a *family* of statistical inferences is considered simultaneously one encounters the **multiple testing problem** (also known as the **multiple comparison problem**).
  - The more inferences made, the more likely an error is going to occur
  - e.g., Even if H_0 is true, the more discrepancy measures (and hence tests) we consider, the more likely it becomes that one of them will erronesouly suggest that null hypothesis should be rejected.
- If a single test has probability $\alpha \in (0,1)$ of yielding a Type I Error, then this probability becomes inflated when considering K simultaneous tests
- Such an inflation is commonly quantified by two metrics:
  - The **family-wise error rate (FWER)** is the probability of making a Type I Error on *any* of the K tests.
  - The **false discocery rate (FDR)** is the expected number of Type I Errors in K tests. $E[\#\text{ of T1 errors}]$

$$P(\text{At least 1 T1 error}) = 1 - P(\text{No T1 errors})$$

$$= 1 - P(\text{No T1 error on test 1 and} \cdots \text{and no T1 error on test k})$$

Assuming the K tests are independent →

$$= 1 - \prod_{i=1}^{K} P(\text{No T1 error on test } i)$$

$$= 1 - \prod_{i=1}^{K} (1-\alpha) = 1 - (1-\alpha)^K$$

Let's Assume that the K tests are dependent in some way.

$P(\text{At least 1 T1 Error}) = P(\text{T1 error on test 1 or } \cdots \text{ or T1 error on test } K)$

$= P\left(\bigcup_{i=1}^{K} \text{T1 error on test } i\right)$ Bode's Inequality

$\leq \sum_{i=1}^{K} P(\text{T1 error on test } i) = K\alpha \leq 1$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- When multiple testing cannot be avoided, many statistical methods have been developed to control FWER and FDR at acceptable values:
  - Bonferroni Correction
  - Sidak Correction
  - Holm-Bonferroni Method
  - Benjamini-Hochberg Procedure
  - ...

  Outside the scope of the course

  These protect you if you're performing K different tests of K different null hypotheses.
- However, we have been considering a special case in which we are testing the same hypothesis using different discrepancy values.
  - And so we can employ a more tailored solution which simply *combines* the information gained by each of the K discrepancy measures rather than considering them in isolation.

Combining Information Across Tests

- To consider the p-values collectively we might consider the smallest of them as measuring the combined evidence against the null hypothesis, i.e.

$$p\text{-value}_{min} = \min_{k=1,\ldots,K} pv_k$$

  - Note that this is appropriate only because the p-values are on a common (interpretable) scale, i.e., they're probabilities
  - We could not, for instance, combine discrepancy measures in the same way

- The smaller the value of p-value$_{min}$ the greater is the evidence against the null hypothesis.

- Note: p-value$_{min}$ is **not** a p-value in the traditional sense
  - but it is a measure of the evidence against the hypothesis.

- Thus we can construct a discrepancy measure out of it:

  This single discrepancy measure combines information across the K tests done previously.

$$D^\star = 1 - p\text{-value}_{min}$$

  - $D^\star$ is defined so that large values, again, indicate evidence against the null hypothesis (unlike the significance level)
  - Therefore, $D^\star$ is a discrepency measure.

- If the observed value of $D^\star$ is $d^\star_{obs}$, then the p-value that describes this combined evidence is denoted by

$$p\text{-value}^\star = Pr\left(D^\star \geq d^\star_{obs} \mid H_0 \text{ is true}\right)$$

- p-value* will necessarily be larger than p-value$_{min}$ because
  - p-value$_{min}$ is the smallest significance level among $pv_1, pv_2, \ldots, pv_K$ and so
  - p-value$_{min}$ exaggerates the evidence against the hypothesis and is misleading as a significance level.

- Given the data, all probabilities are proportions, hence $D^\star$ and p-value* can be calculated.

Estimating $d^\star_{obs}$

- Suppose we have K discrepancy measures $D_1, D_2, ..., D_K$.
  - The combined discrepancy measure is $D^\star = 1 - p\text{-value}_{min}$.

- For $i = 1, \ldots, M_{inner}$ and each discrepancy $k = 1, \ldots, K$
  - randomly mix the two sub-populations $\mathcal{P}_1$ and $\mathcal{P}_2$ yielding $\mathcal{P}^\star_{1,i}$ and $\mathcal{P}^\star_{2,i}$
  - calculate $d_{k,i} = D_k\left(\mathcal{P}^\star_{1,i}, \mathcal{P}^\star_{2,i}\right)$
  - then we estimate each pv_k with

$$\widehat{pv}_k = \frac{1}{M_{inner}} \sum_{i=1}^{M_{inner}} I_{[d_{obs,k},\infty)}(d_{k,i})$$

- Finally, we estimate p-value$_{min}$ and hence $d^\star_{obs}$:

$$\hat{d}^\star_{obs} = 1 - \min_{k=1,\ldots,K} \widehat{pv}_k$$

Estimating p-value*

- In order to estimate p-value* we need an idea of how extreme $\hat{d}^\star_{obs}$ is H_0 were true

- Thus we need to generate a distribution for $D^\star$ so that p-value* can be estimated.

- This is achieved by repeating the the steps above over and over again.
  - Conceptually this requires nested looping: an *inner* loop to calculate $\hat{d}^\star_{obs}$, and an *outer* loop to generate a distribution of $\hat{d}^\star_{obs}$ values

- In particular, we repeat the following steps M_{outer} times:
  - randomly construct two sub-populations and
  - estimate $d_j^\star$ by the same procedure used to calculate $d_{obs}^\star$ (see above)

- then we estimate the p-value* with

$$\widehat{p\text{-value}}^\star = \frac{1}{M_{outer}} \sum_{j=1}^{M_{outer}} I_{[\widehat{d_{obs}^\star},\infty)}(d_j^\star)$$

R Code + Example

- Below is a function that will calculate p-value*, the p-value that accounts for multiple discrepancy measures in the context of multiple testing.
- Notice that throughout the code the functions `sapply`, `Map`, and `Reduce` have been used instead of nested loops for efficiency.

```
calculateSLmulti <- function(pop, discrepancies, M_outer = 1000, M_inner){
  #pop is a list whose two members are two sub-populations

  if (missing(M_inner)) M_inner <- M_outer
  ## Local function to calculate the significance levels
  ## over the discrepancies and return their minimum

  getSLmin <- function(basePop, discrepanies, M) {
    observedVals <- sapply(discrepancies,
                           FUN = function(discrepancy) {discrepancy(basePop)})

    K <- length(discrepancies)

    total <- Reduce(function(counts, i){
      #mixRandomly mixes the two populations randomly, so the new sub-populations are indistinguishable
      NewPop <- mixRandomly(basePop)

      ## calculate the discrepancy and counts
      Map(function(k) {
        Dk <- discrepancies[[k]](NewPop)
        if (Dk >= observedVals[k]) counts[k] <<- counts[k] +1 },
        1:K)
      counts
    },
    1:M, init = numeric(length=K))

    SLs <- total/M
    min(SLs)
  }

  SLmin <- getSLmin(pop, discrepancies, M_inner)

  total <- Reduce(function(count, m){
    basePop <- mixRandomly(pop)
    if (getSLmin(basePop, discrepancies, M_inner) <= SLmin) count + 1 else count
```

```
  },   1:M_outer, init = 0)

  SLstar <- total/M_outer
  SLstar
}
```

- As an example, let's compare the lengths of sharks in Australian versus American encounters.
  - We examine two discrepancies: one that compares averages and another that compares standard deviations.
  - Strictly speaking, this is a multiple testing scenario because we are testing the same null hypothesis with multiple (two) discrepancy measures.

```
getAbsAveDiffsFn <- function(variate) {
  function(pop) {abs(mean(pop$pop1[, variate]) - mean(pop$pop2[,variate]))}
}

discrepancies <- list(getAbsAveDiffsFn("Length"), getSDRatioFn("Length"))

### The following takes a long time (about 20 minutes)
### for M_outer = M_inner = 1,000 say
### So for illustration much smaller values than would be sensible are
### used here
set.seed(341)
SLstar=calculateSLmulti(pop, discrepancies, M_outer = 100, M_inner=100)
SLstar
```

```
## [1] 0.68
```

- Since the p-value is large (0.68), there is no evidence against the hypothesis that the US and Australian encounters were randomly drawn from the same population based on the average and standard deviation of shark lengths.

Exercise: increase the `M_outer` and `M_inner` values above to get a more accurate estimate of the p-value, but note that this is computationally intensive.

4.2.4 An Important Variation on Comparisons

- Consider the population of northeast (`NE`) US counties from the agricultural census.
  - Suppose interest lies in how the number of acres devoted to farms compares between 1982 and 1992.

```
head(agpop[agpop$region == "NE", c("county", "acres82", "acres92")])
```

```
##                county acres82 acres92
## 284  FAIRFIELD COUNTY   17845    9975
## 285   HARTFORD COUNTY   67606   56510
```

```
## 286 LITCHFIELD COUNTY  103942   86581
## 287  MIDDLESEX COUNTY   23191   19830
## 288  NEW HAVEN COUNTY   30024   25882
## 289 NEW LONDON COUNTY   82709   65987
```

- While the counties now constitute a *single* sub-population there still seems to be two sub-populations in play, namely the first being the *counties in 1982* and the second the *counties in 1992*.

- How can we randomly mix the population while accounting for the link between `acres82` and `acres92`?
  - **Randomly swap the variate values** of a county in 1982 and with those of the **same** county in 1992.
    - The randomization thus requires **pairing**, like the paired t-test discussed in introductory statistics courses.

* paired permutation tests are a thing - you just need to think carefully about how to do the random shuffling.