Comparing sub-populations

Comparing sub-populations

- Oftentimes, interest lies in two or more sub-populations.
    - e.g., the encounters that occurred in Australian and US waters (two sub-populations).
    - If the encounters are essentially the same, then the sub-populations observed should not look too different if we were to mix them up with one another.

- Suppose the population, $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$, made from two sub-populations.
    - Then we might compare the differences of the two attributes based on the two sub-populations, e.g. for averages

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

    - or the ratio of the attributes, e.g. standard deviations

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

    - or the compare the populations graphically such as a histogram or quantile plot.

Comparing Shark Encounters

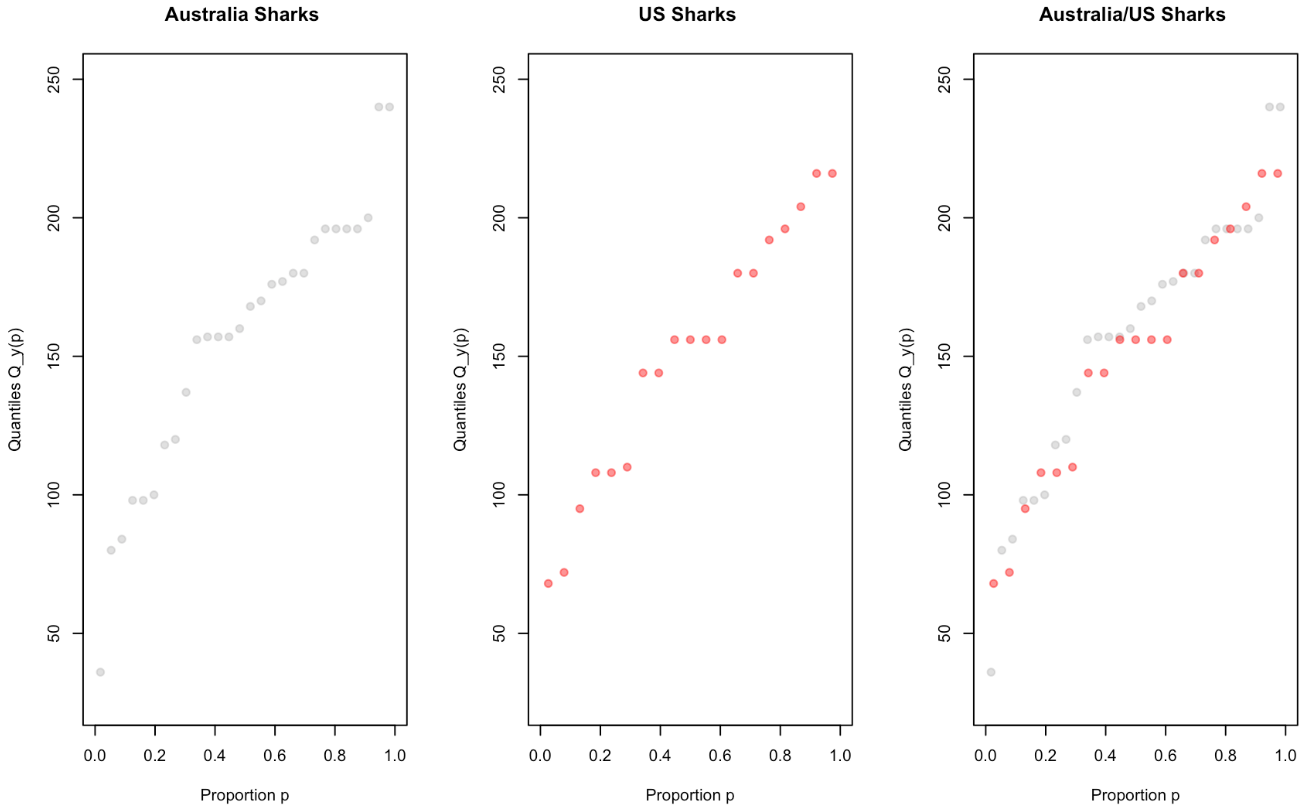
- We can compare the sharks lengths from the two populations

```
pop <- list(pop1 = sharks[sharks[,"Australia"] ==1, ],
            pop2 = sharks[sharks[,"USA"] ==1, ])

Map( function(popi) { summary(popi$Length) }, pop)
```

```
## $pop1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    36.0   119.5   164.0   155.9   193.0   240.0
##
## $pop2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    68.0   109.0   156.0   150.4   186.0   216.0
```

- A quantile plot of sharks from the two populations



shuffle 57 encounters and divide them into 2 groups
always pop2 -> 29
pop1 -> 28

Randomly Mixing Population

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

```
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)
}
```

- Note that the mixing of the two sub-populations maintains the population sizes.

Example

- We can shuffle or mix the two populations and then compare the attributes values.

```
set.seed(341)
mixedPop <- mixRandomly(pop)

c( mean(mixedPop$pop1[,"Length"]) - mean(mixedPop$pop2[,"Length"]),
sd(mixedPop$pop1[,"Length"])/sd(mixedPop$pop2[,"Length"]) )
```

```
## [1] -18.1522556   0.9283716
```

- Then we might compare the randomly shuffled populations attributes to the Australia and US attributes.

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

```
## [1] 5.524436 1.056418
```

Some convenient functions

- It will be convenient to write functions that return functions which in turn calculate these attributes *for any of the variates* in the population.
    - The difference in the averages and the ratio of the 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 lengths

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

- For US and Australia populations.

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

```
## [1] 5.524436 1.056418
```

- For shuffled populations.

```
c(diffAveLengths(mixedPop), ratioSDLengths(mixedPop))
```

```
## [1] -18.1522556   0.9283716
```

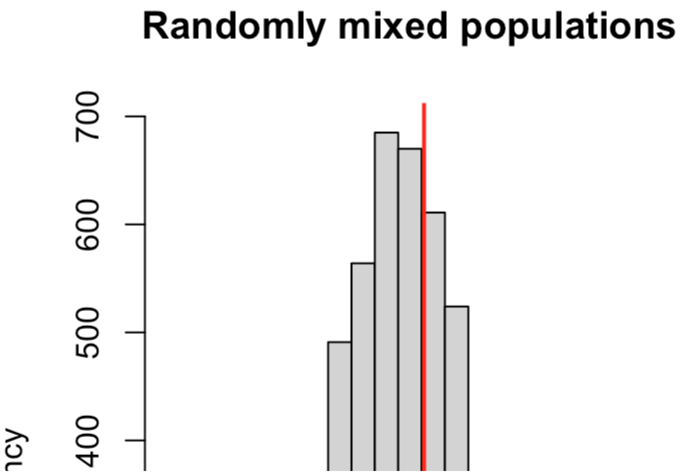
- It seems that the standard deviation does not change much under shuffling, but the mean does change.
    - To make this claim formal (statistically sound) we need to do more statistical analysis (e.g. perform a test of hypothesis)

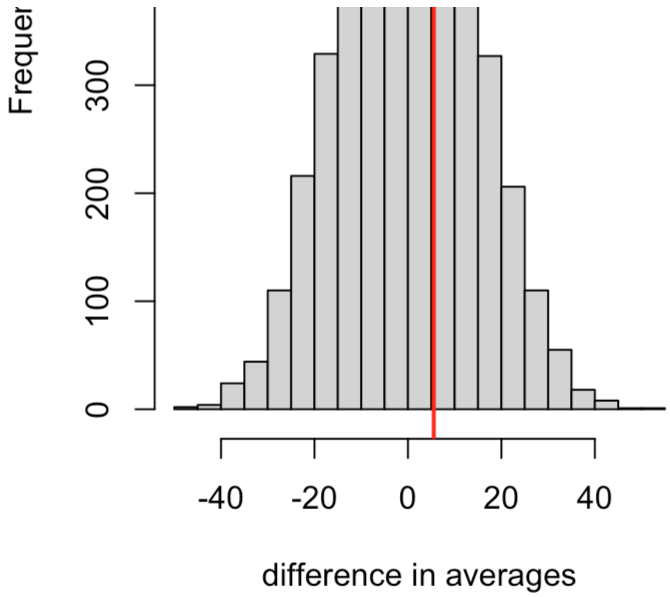
Shuffing the Populations

- To see how unusual the given pair of sub-populations are to any randomly shuffled pair.
- Ideally, we could look at all possible shufflings.
    - This is the same as all possible permutations of the numbers 1 to N where $N = N_1 + N_2$ is the sum of the two sub-population sizes.
    - This requires about 2.6×10^{59} shuffles in the shark length data.
    - We use 5,000 shuffles instead.

```
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)
```



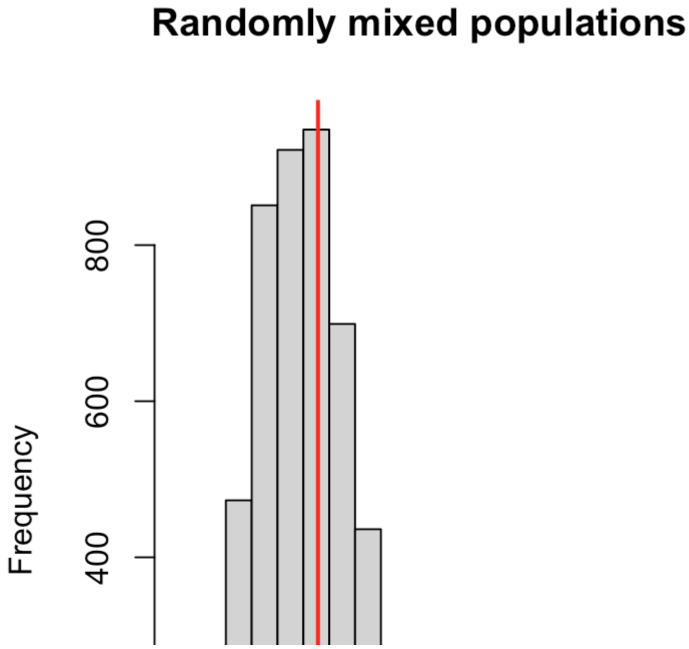


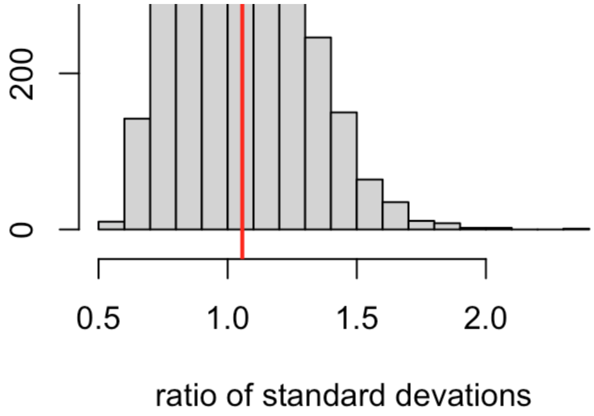
Standard Deiviation

- To see how unusual the given pair of sub-populations are to any randomly shuffled pair.

```
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)
```





Difference in Surfing

- Comparing the shark encounters involving Surfing from Australia and the USA.

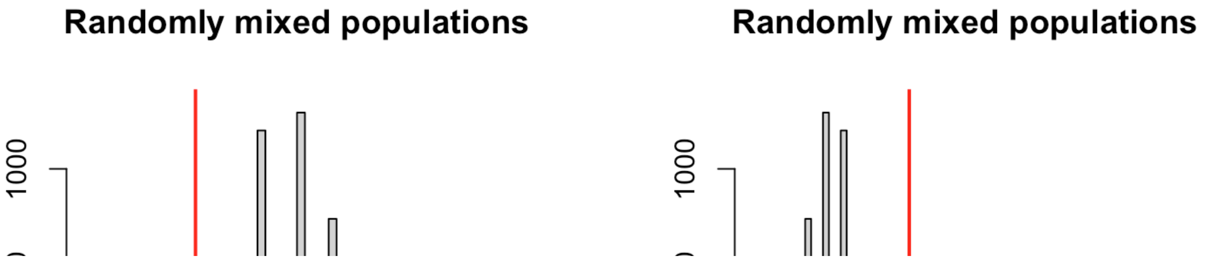
```
diffAveSurf <- getAveDiffsFn("Surfing")
ratioSDSurf <- getSDRatioFn("Surfing")
```

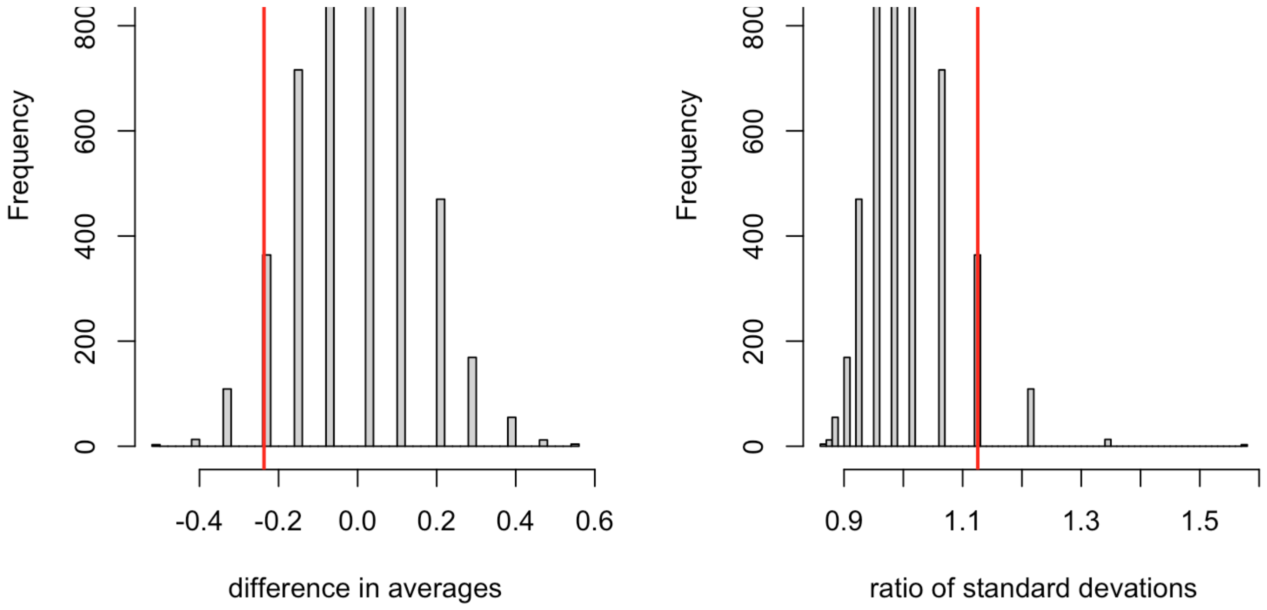
```
par(mfrow=c(1,2),oma=c(0,0,2,0))

set.seed(341)
pair <- sapply(1:5000,
   FUN = function(...){
     tmixpop = mixRandomly(pop)
     c( diffAveSurf(tmixpop), ratioSDSurf(tmixpop))  })

hist(pair[1,], breaks="FD",
     main = "Randomly mixed populations", xlab="difference in averages", col="ligh
tgrey")
abline(v=diffAveSurf(pop), col = "red", lwd=2)


hist(pair[2,], breaks="FD",
     main = "Randomly mixed populations", xlab="ratio of standard devations", col=
"lightgrey")
abline(v=ratioSDSurf(pop), col = "red", lwd=2)
```





Difference in Shark Length with Fatality

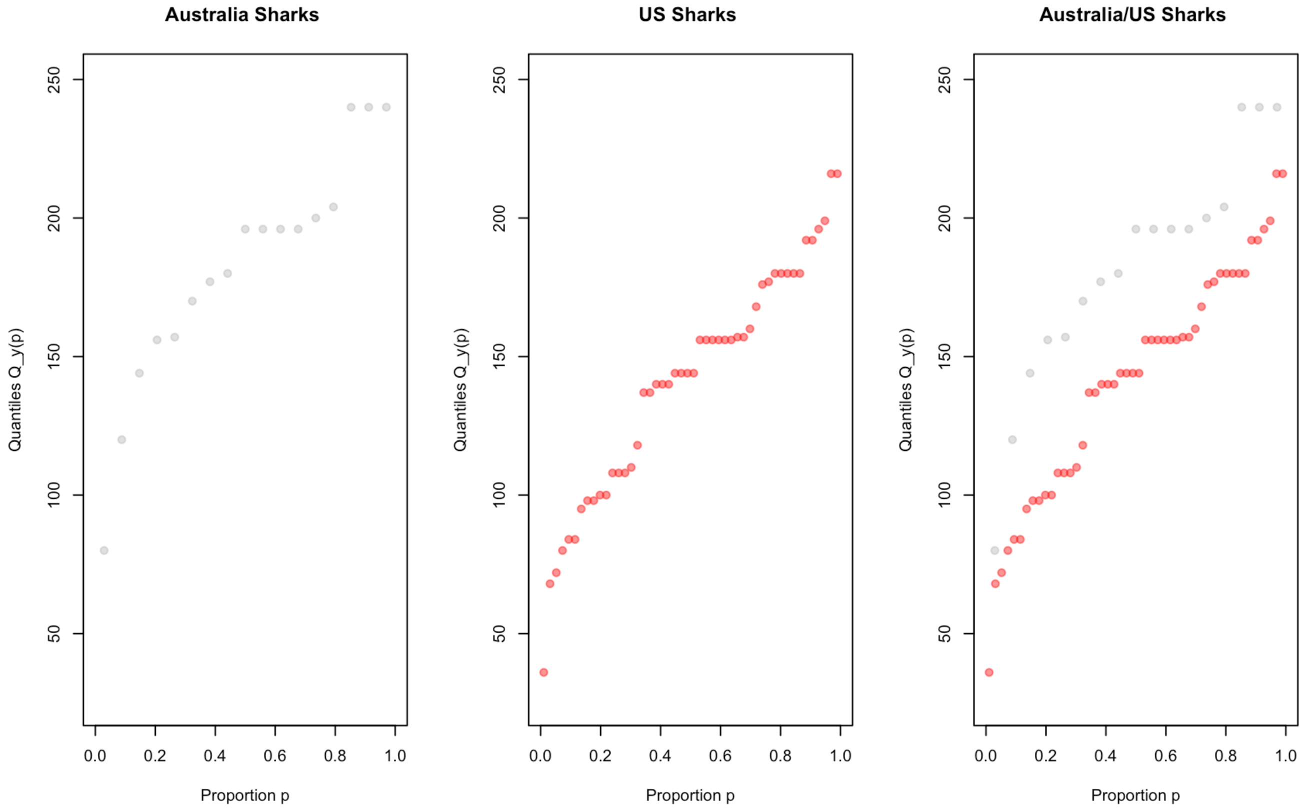
- Two other sub-populations;
  - Fatal shark encounters and
  - Non-Fatal shark encounters

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

- We can compare the sharks lengths from the two populations

```
## $pop1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    80.0   157.0   196.0   181.9   200.0   240.0
##
## $pop2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    36.0   108.0   144.0   141.2   176.2   216.0
```

- A quantile plot of sharks from the two populations



Comparing Shark Encounters

- Then we quantify the difference in the average and standard deviation of the shark lengths from the two populations by
  - randomly mixing the sub-populations.

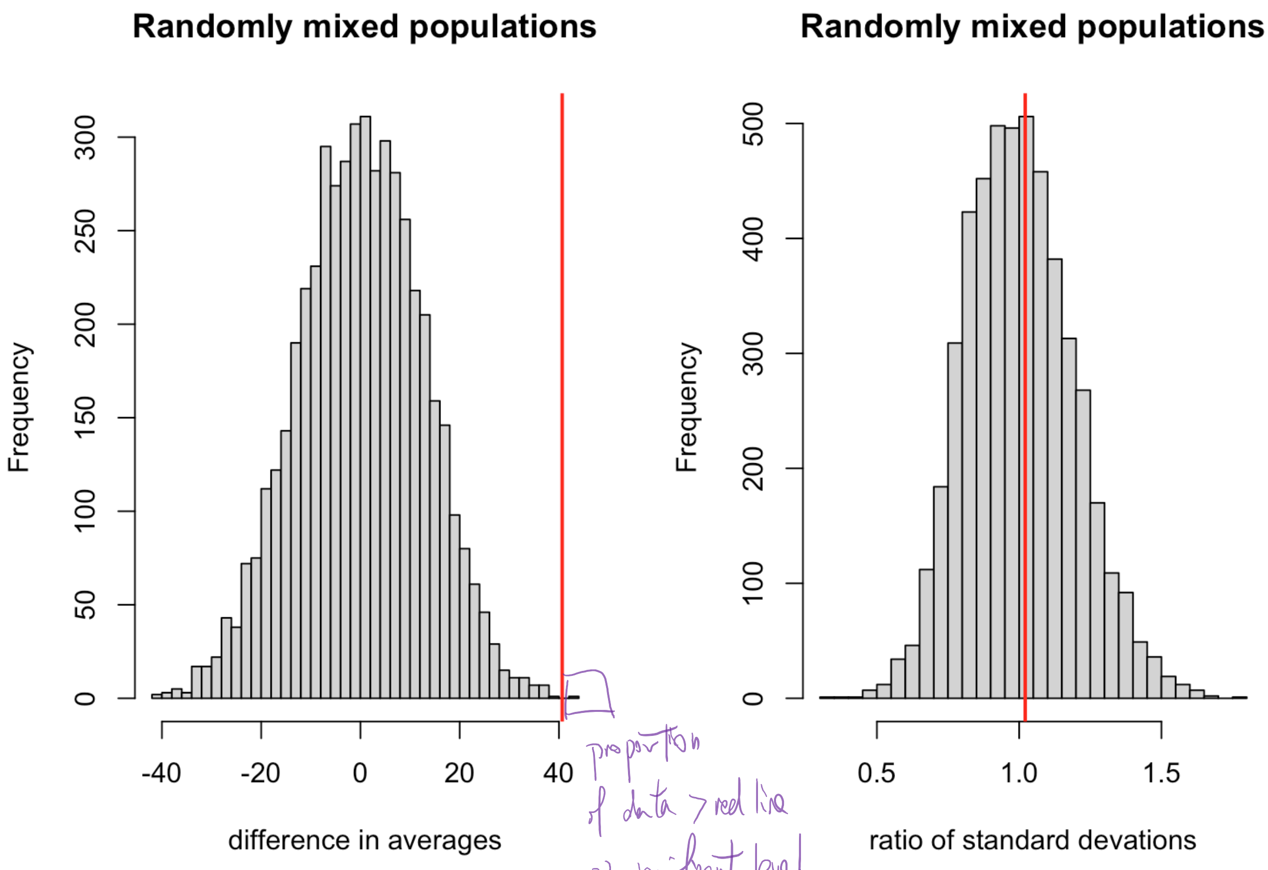
```
Fatpop <- list(pop1 = sharks[sharks[,"Fatality"] ==1, ],
            pop2 = sharks[sharks[,"Fatality"] ==0, ])

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)
```



proportion of data > red line => significant level

Difference in Shark Length with Fatality

- We can quantify the difference in the median and IQR of the shark lengths from the two populations by
  - randomly mixing the sub-populations.

```
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")
```

Anatomy of a test of significance

- We would like to quantify ``how unusual is the difference between the population averages". e.g.
  - the average sharks length from Australia and US
  - the average sharks length involving a survival and fatality.
- We start by a null hypothesis about the parameter of interest, e.g. $H_0 : \theta = \theta_0$
- We then need a *discrepency measure* to quantify how much the data is inconsistent with H_0.
- The last step is to interpret the value of the discrepancy measure in terms of the amount of evidence against H_0.

$\text{data} \longrightarrow \boxed{\text{discrepency measure}} \longrightarrow \text{a number}$

a measure of how inconsistent is data with H_0

Anatomy of a test of significance

- One measure to check how unusual the difference between the two sub-populations is, is called the **observed significance level**

the larger the discrepency measure, the more evidence saying that the 2 populations are different

$$SL = \Pr\left(\overset{\text{discrepency measure}}{\underline{\mid a(\mathcal{P}_1) - a(\mathcal{P}_2) \mid}} \geq \underset{\text{red line}}{\underline{\mid a(\mathcal{P}_{Australia}) - a(\mathcal{P}_{USA}) \mid}} \right)$$

where the populations $\mathcal{P}_1$ and $\mathcal{P}_2$ are randomly drawn (with equal probability) from the set of all pairs $(\mathcal{P}_1, \mathcal{P}_2)$ where

$$\mathcal{P}_1 \cup \mathcal{P}_2 = \mathcal{P}_{Australia} \cup \mathcal{P}_{USA},$$

$$\mathcal{P}_1 \cap \mathcal{P}_2 = \emptyset,$$

$$size(\mathcal{P}_1) = size(\mathcal{P}_{Australia}), \quad \text{and} \quad size(\mathcal{P}_2) = size(\mathcal{P}_{USA}).$$

The smaller is the value, SL, of this probability the more different are the pair of populations $(\mathcal{P}_{Australia}, \mathcal{P}_{USA})$ in terms of the attribute $a(\cdot)$.

Anatomy of a test of significance

- If we cannot enumerate all possible permutations, we do not have the exact value of SL.
  - It can, of course, be well approximated by using the sample of 5,000 pairs $(\mathcal{P}_1, \mathcal{P}_2)$ that we generated according to this probability mechanism.
- Calculating this approximation as

```
sum(abs(diffLengths) >= abs(diffAveLengths(pop))) / length(diffLengths)
```

  - For US and Australia Shark Lengths gives $SL \approx \widehat{SL} = 0.704$.
  - For Fatality and non-Fatality shark Lengths gives $SL \approx \widehat{SL} = 0.0002$

SL large -> no diff
SL small -> diff

Interpretation

if red line is likely wrt shuffled population, then SL is large is saying that there is no evidence against H0: 2 populations are the same; otherwise, if SL is small, then there is evidence against H0: 2 populations are the same

- Suppose that the pair $(\mathcal{P}_{Australia}, \mathcal{P}_{USA})$ is a random draw from the above set of pairs $(\mathcal{P}_1, \mathcal{P}_2)$,
  - 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.
- A large SL $(\mathcal{P}_{Australia}, \mathcal{P}_{USA})$ being as large as that indicates
  - there is **no evidence against the hypothesis** that the pair $(\mathcal{P}_{Australia}, \mathcal{P}_{USA})$ was randomly drawn.
  - We have no evidence against the hypothesis that the two populations $\mathcal{P}_{Australia}$ and $\mathcal{P}_{USA}$ are indistinguishable.

Test of Significance

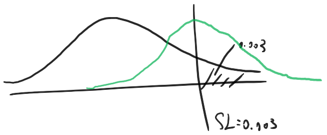
1. A **hypothesis**, here expressed equivalently as
   - $\mathcal{P}_{Australia}$ and $\mathcal{P}_{USA}$ are drawn from the same population of shark encounters, or
   - the pair of sub-populations $(\mathcal{P}_{Australia}, \mathcal{P}_{USA})$ were created by randomly assigning units in the same population to one or other of the sub-populations, or
   - The two populations $\mathcal{P}_{Australia}$ and $\mathcal{P}_{USA}$ are equal in terms of their attribute values, i.e. $a(\mathcal{P}_{Australia}) = a(\mathcal{P}_{USA})$.
2. A measure of **discrepancy** $D = D(\mathcal{P}_1, \mathcal{P}_2)$ where large values indicate **evidence against the hypothesis**,
   - e.g. $D(\mathcal{P}_1, \mathcal{P}_2) = |a(\mathcal{P}_1) - a(\mathcal{P}_2)|$
3. The **observed discrepancy** $\underline{d = D(\mathcal{P}_{Australia}, \mathcal{P}_{USA})}$, and red line
4. The probability of $D \geq d$ **when the hypothesis is true.**
5. The **observed significance level**, SL, is then

$$SL = Pr\left(D \geq d \,\middle|\, \text{the hypothesis is true}\right).$$ proportion

Significance level

- The **observed significance level**, SL, is then

$$SL = Pr\left(D \geq d \,\middle|\, \text{the hypothesis is true}\right).$$



- If SL is very small then either
  - the hypothesis is true and we have observed a very unusual value of d,

=> Either shuffling is allowed, but what we observe is rare
OR shuffling is not allowed

  - or, the hypothesis is false.

- The smaller is SL the greater the evidence against the hypothesis.

- In the extreme case where $SL = 0$, then we have observed something impossible and the hypothesis must therefore be false – this would be a proof by contradiction.

- Note that SL is also called the p-value by many writers.

Some Important Things

- the observed significance level provides a common (probabilistic) scale on which to measure the **evidence against the hypothesis** assumed;

- the observed significance level does **not** measure evidence **in favour** of the hypothesis
  - in science, we try to falsify hypotheses and entertain only those which remain standing;
- a test of significance therefore **neither accepts nor rejects a hypothesis** but simply provides a measure of the evidence against;

- there is **no magic level for SL** such as 0.05 or 0.01,
  - there being no practical or scientific difference between $SL = 0.048$ and $SL = 0.051$ for example;

Some More Important Things

- the fact that the evidence against the hypothesis is **statistically significant** based on some discrepancy measure **does not imply that the discrepancy is practically significant**
  - i.e. the SL measures how unusual a discrepancy of that size might be when the hypothesis holds,
  - it says nothing about whether a discrepancy of that size matters for any practical or scientific purpose
- every test of significance is based on some measure of discrepancy and **different discrepancy measures can detect different departures** from the hypothesis, so one needs to understand the nature of the departure from the hypothesis that the discrepancy is trying to measure.
  - For sharks lengths data, the difference between sharks length in fatal and non-fatal encounters was 3 and $1/4$ ft.
- The discrepancy measure quantifies only one type of discrepancy between the populations
  - e.g. shark length.
  - Any other differences are completely ignored.

Errors

- In Judgement

Decision	the person is guilty	the person is innocent
Convicted	Correct	Error (Type I Error)

Decision	the person is guilty	the person is innocent
Acquitted	Error (Type I Error)	Correct

- In Hypothesis Testing

Decision	the hypothesis is true	the hypothesis is false
Not Reject	Correct	Error (Type II Error)
Reject	Error (Type I Error)	Correct

we don't take the action of reject or not
we just provide risks/probability

A t-like discrepancy measure

- Another 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”
  - in that 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.
- **Question:** What conditions on $a(\ldots)$ would be required for the measure to also be location-invariant?

The denominator

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

- The challenge is determining the denominator
  - In rare cases, the denominator might be known and then this discrepancy measure is a rescaling of the difference.
  - More commonly, we will estimate the denominator using information from $\mathcal{P}_1$ and $\mathcal{P}_2$.

Independent Samples

- Suppose that the populations $\mathcal{P}_1$ and $\mathcal{P}_2$ are **independently** drawn
  - Then the discrepancy measure would become

$$\begin{aligned} D(\mathcal{P}_1, \mathcal{P}_2) &= \frac{a(\mathcal{P}_1) - a(\mathcal{P}_2)}{\widetilde{SD}(a(\mathcal{P}_1) - a(\mathcal{P}_2))} \\ &= \frac{a(\mathcal{P}_1) - a(\mathcal{P}_2)}{\left(\widetilde{SD}^2(a(\mathcal{P}_1)) + \widetilde{SD}^2(a(\mathcal{P}_2))\right)^{\frac{1}{2}}} \end{aligned}$$

where $\widetilde{SD}(\cdots)$ denotes an estimator of the standard deviation of its argument.

Differences in Averages

- Suppose that $a(\mathcal{P}_i) = \overline{Y}_i$ and $\mathcal{P}_i$ is size n_i, $i = 1, 2$

  $$D(\mathcal{P}_1, \mathcal{P}_2) = \frac{\overline{Y}_1 - \overline{Y}_2}{\widetilde{\sigma}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}}}$$

  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 of the Y values from each of $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively, and we know that the standard deviation of Y remains the same in population 1 and 2,
  - then the pooled estimator of σ would be

  $$\widetilde{\sigma} = \left(\frac{(n_1 - 1)\widetilde{\sigma}_1^2 + (n_2 - 1)\widetilde{\sigma}_2^2}{(n_1 - 1) + (n_2 - 1)}\right)^{\frac{1}{2}}.$$
- If the two populations $\mathcal{P}_1$ and $\mathcal{P}_2$ have different standard deviations σ_1 and σ_2, respectively, then

  $$D(\mathcal{P}_1, \mathcal{P}_2) = \frac{\overline{Y}_1 - \overline{Y}_2}{\left(\widetilde{\frac{\sigma_1}{n_1}} + \widetilde{\frac{\sigma_2}{n_2}}\right)^{\frac{1}{2}}}$$

Gausssian Assumption?

$$D(\mathcal{P}_1, \mathcal{P}_2) = \frac{\overline{Y}_1 - \overline{Y}_2}{\widetilde{\sigma}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}}} \qquad \text{where} \qquad \widetilde{\sigma} = \left(\frac{(n_1 - 1)\widetilde{\sigma}_1^2 + (n_2 - 1)\widetilde{\sigma}_2^2}{(n_1 - 1) + (n_2 - 1)}\right)^{\frac{1}{2}}.$$

- This is the “two-sample” Student t statistic used to test the equality of the means of two Gaussian (or “normal”) distributions with common (but unknown) standard deviation σ.
  - If the Y values were in fact Gaussian distributed, the discrepancy would follow a Student t distribution with $n_1 + n_2 - 2$ degrees of freedom under the hypothesis that the means were identical.
- Note, however, in our procedure of randomly mixing the populations we make **no such Gaussian assumption**.

  - We proceed with this discrepancy measure just as we did with the earlier measures
  - but now we need to calculate a standard error as well.

Rcode

A function that will return this discrepancy measure for any variate `var` is

```
### 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
  }
}
```

Calculating the Obseraved Values

- Get the t-function for “Length”

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

- The value for the two sets of sub-populations,
  - US and Australia encounters and
  - fatal and non-fatal encounters.

```
c(tStatLengths(pop), tStatLengths(Fatpop))
```

```
## [1] 0.3886752 3.4454919
```

- To gauge the size of the these discrepancy measures we
  - mix, shuffle, or permute the sub-populations 5,000 times and plot the histogram as before and
  - overlay the Student t density on $n_1 + n_2 - 2$ degrees of freedom which we would use if the Gaussian models applied.

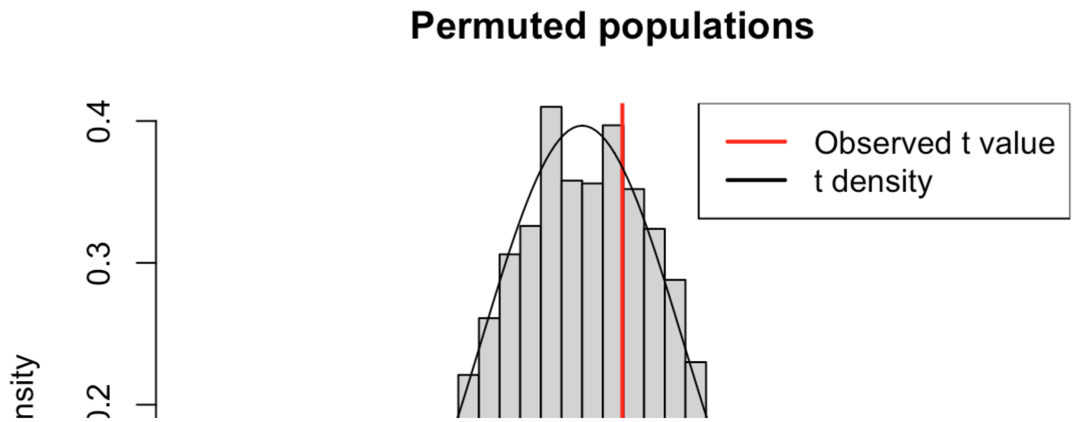
Histogram of discrepancy measures

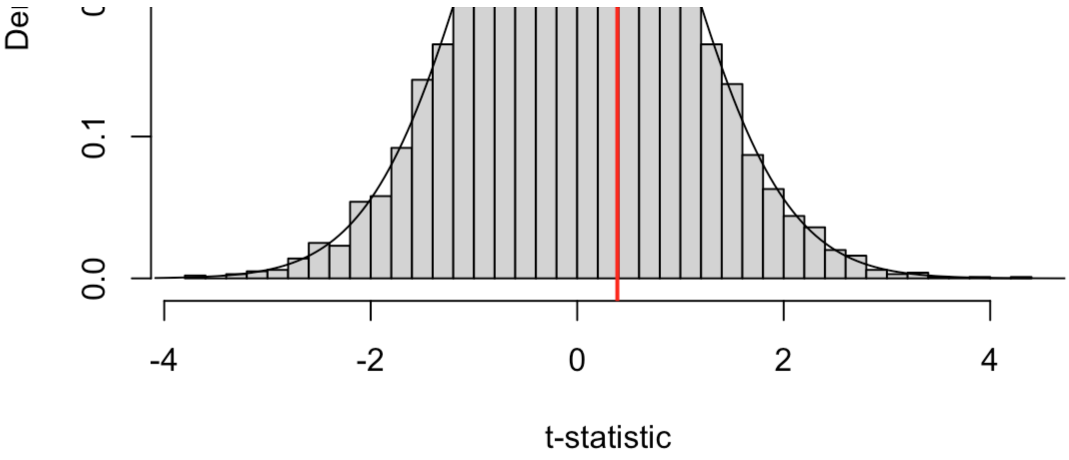
- The discrepancy measure on length of the US and Australia sub-populations.

```
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"))
```





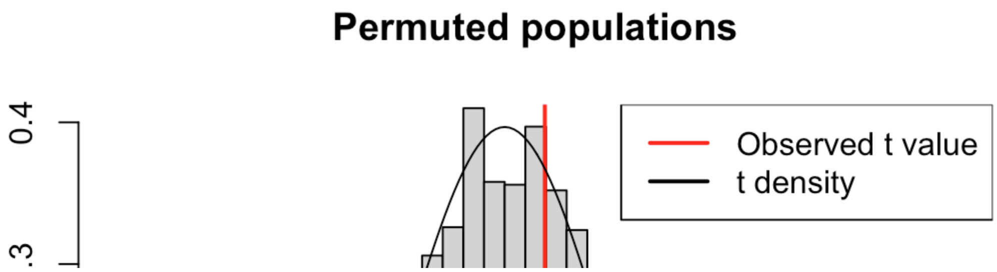
Comments on Histogram

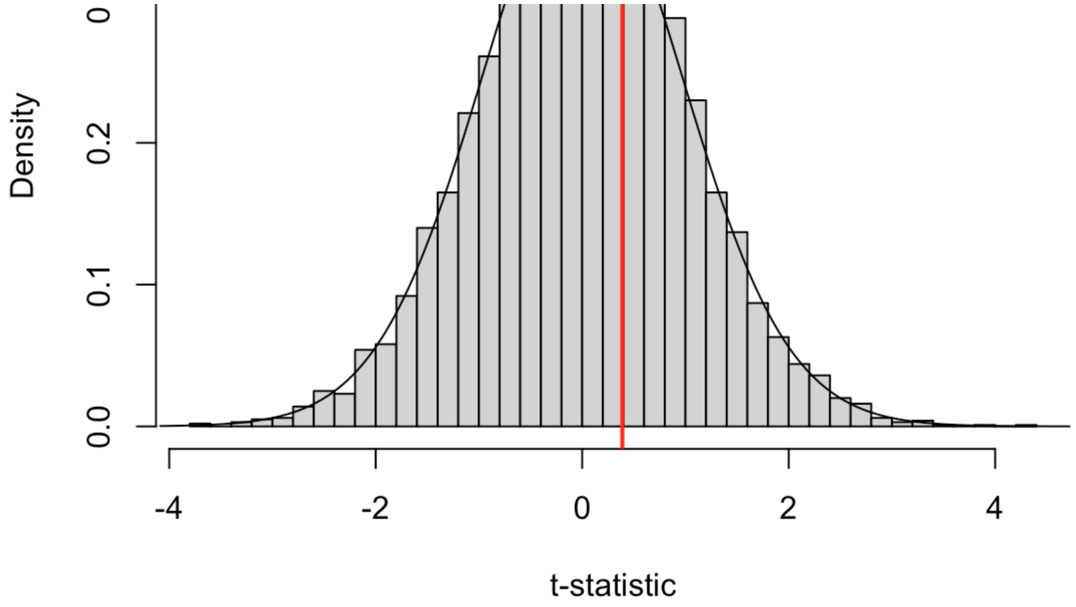
- The discrepancy measure on length of the US and Australia sub-populations.

```
set.seed(341)
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"))
```





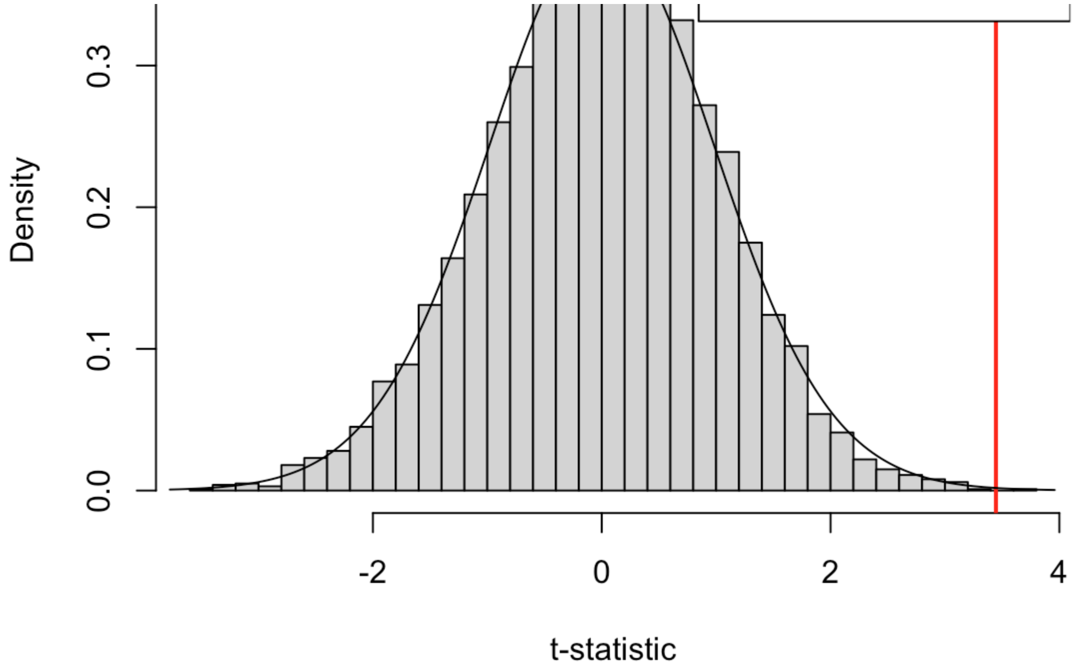
```
# The significance Level is:
SL = sum(abs(tVals) >= abs(tStatLengths(list(pop1 = sharks[sharks[,"Australia"] ==
1, ], pop2 = sharks[sharks[,"USA"] ==1, ])))) / length(tVals)
SL
```

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

- Remarkably, the Student t density closely approximates the histogram!
  - In many instances, even when no Gaussian 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 randomly mixed distribution.
- The significance level SL observed for this discrepancy measure in this example is $P(|D(\mathcal{P}_1, \mathcal{P}_2)| \geq red\ line)$= 0.704.
  - This is so large that the observed discrepancy measure is not at all unusual when the hypothesis of $H_0 : a(\mathcal{P}_1) = a(\mathcal{P}_2)$ is true.
  - This test provides **no evidence against the hypothesis**.

Comments on Histogram

- The discrepancy measure on length of the fatal and non-fatal sub-populations.



- The significance level observed for this discrepancy measure in this example is 210^{-4}.
    - This value is so small that the observed discrepancy measure is unusual when the hypothesis of $H_0 : a(\mathcal{P}_1) = a(\mathcal{P}_2)$ is true.
    - This test provides **evidence against the null hypothesis**.

Guideline to Interpret the Significance Level

- $SL < 0.001$ means that there is **very strong evidence** against H_0
- $0.001 < SL < 0.01$ means that there is **strong evidence** against H_0
- $0.01 < SL < 0.05$ means that there is **evidence** against H_0
- $0.05 < SL < 0.1$ means that there is **weak or some evidence** against H_0
- $SL > 0.1$ means that there is **no evidence** against H_0

Multiple Testing with Random Noise

- Let's see what happens if we simulate data from independent populations and try to study the correlation coefficient.
- Create two populations, $\mathcal{P}_1 = \{\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1m}\}$ and $\mathcal{P}_2 = \{\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2m}\}$, with independent random noise.

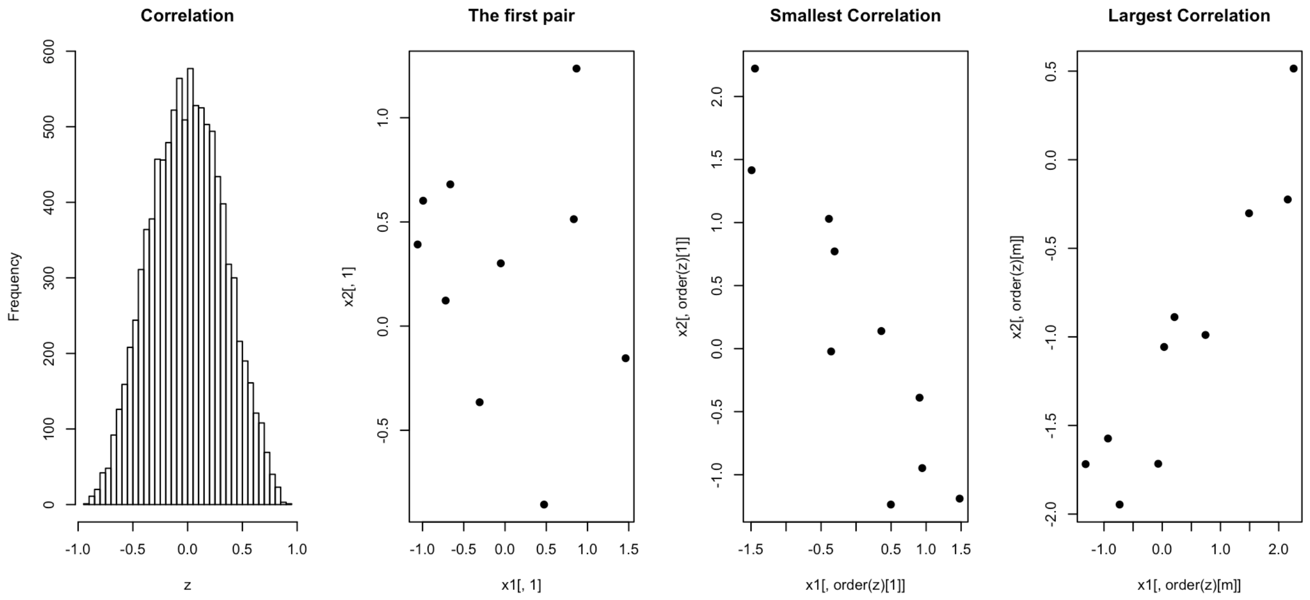
```
n = 10;  m = 10^4;
set.seed(341)
x1 = matrix(rnorm(n*m), nrow=n, ncol=m)
x2 = matrix(rnorm(n*m), nrow=n, ncol=m)
```

- Calculate the sample correlation between $\mathbf{x}_{1j}$ and $\mathbf{x}_{2j}$, where $j = 1, \ldots, m$.

```
z = numeric(m)
for (j in 1:m) z[j] = cor(x1[,j], x2[,j])
```

- Plot some values

```
par(mfrow=c(1,4),oma=c(0,0,2,0))
hist( z, main="Correlation", breaks="FD")
plot( x1[, 1], x2[,1], main="The first pair", pch=19)
plot( x1[, order(z)[1]], x2[,order(z)[1]], main="Smallest Correlation", pch=19)
plot( x1[, order(z)[m]], x2[,order(z)[m]], main="Largest Correlation", pch=19)
```



- Although the data is randomly generated from X_1 and X_2, which are clearly independent, there are samples in the simulation with large sample correlation coefficients.

Multiple testing

- We might consider any number of discrepancy measures, $D_1, D_2, \ldots, D_K$
  - each with an associated observed significance level say $SL_1, SL_2, \ldots, SL_K$.
  - Unlike the several discrepancy measures, these significance levels **are on a common and interpretable scale** (probability).
- Because the significance levels are on a common and interpretable scale, we might consider the smallest of these as measuring the combined evidence against the hypothesis. i.e.

$$SL_{min} = \min_{k=1,\ldots,K} SL_k.$$

The smaller is SL_{min} the greater is the evidence against the hypothesis.

- SL_{min} is **not** a significance level
  - but it is a measure of the evidence against the hypothesis.

A discrepancy measure

- To make SL_{min} into a discrepancy measure, we let

$$D^\star = 1 - SL_{min}$$

  - $D^\star$ is arranged so that large values again indicate evidence against the hypothesis.
  - Therefore, $D^\star$ is a discrepency measure.
- If the observed value of $D^\star$ is $d^\star$, then the significance that describes this combined evidence is denoted by

$$SL^\star = Pr\left(D^\star \geq d^\star \quad | \text{ Hypothesis is true}\right),$$

which will be larger than SL_{min}
  - i.e. SL_{min} exaggerates the evidence against the hypothesis and so is misleading as a significance level.
- Given data, the values $d^\star$ and SL^* must be estimated.

Estimating the SL Value (no multiple testing)

- Suppose we only have **one** discrepancy measure. Then we
  - Recall that the **observed significance level**, SL, is

$$SL = Pr\left(D \geq d_{obs} \,\middle|\, \text{the hypothesis is true}\right).$$

  - calculate d_{obs} which is the observed discrepancy measure based on the given sub-populations, $\mathcal{P}_1$ & $\mathcal{P}_2$.
  - e.g. sharks encounters from Australia or the USA.
- If we could construct all possible samples from sub-populations $\mathcal{P}_1$ & $\mathcal{P}_2$
  - we could calculate the SL exactly
  - but there are too many possible samples, so we estimate SL.

Estimating SL:

- For $i = 1, \ldots, M$
  - randomly construct two sub-populations, $\mathcal{S}_{i1}$ & $\mathcal{S}_{i2}$, (while maintaining the sub-population sizes) by sampling without replacement from the population $\{\mathcal{P}_1, \mathcal{P}_2\}$
  - calculate $d_i = D(\mathcal{S}_{i1}, \mathcal{S}_{i2})$

- then we estimate the SL with

$$\widehat{SL} = \frac{1}{M} \sum_{i=1}^{M} \mathrm{I}\,(d_i \geq d_{obs})$$

- Interpret $\widehat{SL}$ according to the guidelines provided before.

Estimating $d^{\star}_{obs}$ (multiple testing)

- Recall that
    - the discrepancy measure is $D^{\star} = 1 - SL_{min}$, and
    - $SL^{\star} = Pr\left(D^{\star} \geq d^{\star} \;\mid\; \text{Hypothesis is true}\right)$ in which $d^{\star}$ is the observed value of the discrepancy measure $D^{\star}$
- Suppose we have K discrepancy measures $D_1, D_2, \ldots, D_K$. Then we
    - calculate $d_{k,obs} = D_k(\mathcal{P}_1, \mathcal{P}_2)$ based on the given sub-populations, $\mathcal{P}_1$ & $\mathcal{P}_2$.
    - e.g. sharks encounters from Australia or the USA.
- For $i = 1, \ldots, M$
    - randomly construct two sub-populations, $\mathcal{S}_{i1}$ & $\mathcal{S}_{i2}$, from $\{\mathcal{P}_1, \mathcal{P}_2\}$
    - calculate $d_{ik} = D_k(\mathcal{S}_{i1}, \mathcal{S}_{i2})$
- We estimate each SL_k with

$$\widehat{SL}_k = \frac{1}{M} \sum_{i=1}^{M} \mathrm{I}\left(d_{ik} \geq d_{k,obs}\right)$$

    - Finally we estimate SL_{min} and $d^{\star}_{obs}$ with

$$\widehat{d}^{\star}_{obs} = 1 - \widehat{SL}_{min,obs} = 1 - \min_{k=1,\ldots,K} \widehat{SL}_k$$

Estimating $SL^{\star}$ (multiple testing)

- The discrepancy measure is

$$d^{\star} = 1 - SL_{min} = 1 - \min_{k=1,\ldots,K} SL_k$$

    - we estimate this quantity with $\widehat{d}^{\star}_{obs}$ based on the given sub-populations, $\mathcal{P}_1$ & $\mathcal{P}_2$.
    - e.g. sharks encounters from Australia or the USA.
- Repeat $M^{\star}$ times
    - randomly construct two sub-populations, $\mathcal{S}_{i1}$ & $\mathcal{S}_{i2}$, from $\{\mathcal{P}_1, \mathcal{P}_2\}$
    - we estimate this quantity with $d^{\star}_i$
    - by the same procedure used to calculate $d^{\star}_{obs}$ (see the previous slide) but now using $\mathcal{S}_{i1}$ & $\mathcal{S}_{i2}$ as the given sub-populations.

- then we estimate the $SL^{\star}$ with

$$\widehat{SL}^{\star} = \frac{1}{M^{\star}} \sum_{i=1}^{M^{\star}} \mathrm{I}\left(d_i^{\star} \geq \widehat{d}_{obs}^{\star}\right)$$

- Interpret $\widehat{SL}$ according to the guidelines provided before.
- **Computational Time:** note that there are nested loops here, one a loop over $i = 1, \ldots, M^{\star}$ to calculate $\widehat{S}L^{\star}$, each iteration of which involves another loop over $k = 1, \ldots, K$ to estimate $d_{obs}^{\star}$.
    - we will use Map-Reduce approach to increase efficiency of the the code.

Rcode

- Below is the R code to calculate the significance level $SL^{\star}$ for multiple testing.
- Notice that throughout the code the functions `sapply` , `Map` , and `Reduce` have been used instead of nested loops.

```
#pop is a list whose two members are two sub-populations
calculateSLmulti <- function(pop, discrepancies, B_outer = 1000, B_inner){
  if (missing(B_inner)) B_inner <- B_outer
  ## Local function to calculate the significance levels
  ## over the discrepancies and return their minimum

  getSLmin <- function(basePop, discrepanies, B) {
  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:B, init = numeric(length=K))

    SLs <- total/B
    min(SLs)
  }

  SLmin <- getSLmin(pop, discrepancies, B_inner)

  total <- Reduce(function(count, b){
    basePop <- mixRandomly(pop)
    if (getSLmin(basePop, discrepancies, B_inner) <= SLmin) count + 1 else count
  },   1:B_outer, init = 0)

  SLstar <- total/B_outer
  SLstar
}
```

Rcode example

- Let us compare the encounters happened in Australia versus those happened in the USA.
- We would like to see if there is a difference in mean shark length between the two sub-populations

above (Australia vs USA).

- Recall from R codes in this slide set that `pop` is a list where `pop$pop1` and `pop$pop2` are the Australian and US sub-populations, respectively.

```
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 B_outer = B_inner = 1,000 say
### So for illustration much smaller values than would be sensible are
### used here
set.seed(341)
SLstar=calculateSLmulti(pop, discrepancies, B_outer = 100, B_inner=100)
SLstar
```

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

- Since the significance level is large (0.68), there is no evidence against the hypothesis that the mean shark length of US encounters is equal to that of the Australian encounters.
    - Note that this calculated significance level is based on a very small simulation (read the comments in the code above).
    - increase the `B_outer` and `B_inner` values above to get a more accurate estimate of the significance level (computationally intensive though).

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 to 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 would require **pairing**, like paired t-test discussed in introductory stats course.

Exercise: Implement this kind of significance test for any population attribute. You can start by comparing the mean attribute in `acres82` and `acres92` using `agpop` dataset.