Inductive Inference

Contents

4.2 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 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.
  - i.e. swapping units would not change the features of the two sub-populations.
- Suppose we have the population, $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$,
  - 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

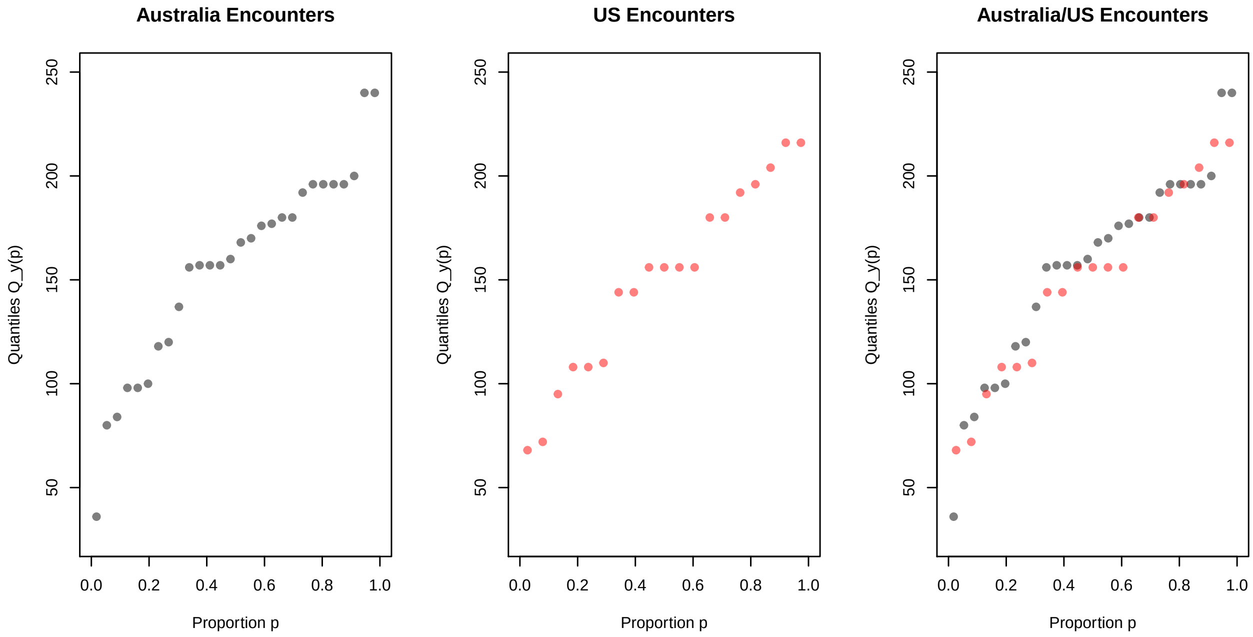
- After loading the shark data, we can compare the sharks lengths from the two populations numerically

```
directory = "../Data"
dirsep= "/"
sharkfile <- paste(directory, "Sharks", "sharks.csv", sep=dirsep)
sharks <- read.csv(sharkfile)

### Units in the large population of all encounters
popSharks <- rownames(sharks)
### get the sub-population that is just those encounters in Australian waters
popSharksAustralia <- popSharks[sharks$Australia == 1]
### the units in the sub-population are

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
```

- or graphically 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 = "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 = "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 = "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) )
```



- To summarize the differences in the two populations we typically use a single measure such as the difference in averages of the two sub-populations, i.e.

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

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

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

Randomly Mixing Population

- 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.
  - i.e. swapping units would not change the features of the two sub-populations.

- Here we shuffle the two sub-populations together $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$ then
  - randomly draw two new sub-populations $\mathcal{P}_1^\star$ & $\mathcal{P}_2^\star$
  - Should we maintain the population sizes?

```
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 $\{\mathcal{P}_1, \mathcal{P}_2\}$ and $\{\mathcal{P}_1^\star, \mathcal{P}_2^\star\}$, i.e. compare
  - $a(\mathcal{P}_1)$ to $a(\mathcal{P}_1^\star)$, $a(\mathcal{P}_2)$ to $a(\mathcal{P}_2^\star)$, or
  - $a(\mathcal{P}_1) - a(\mathcal{P}_2)$ to $a(\mathcal{P}_1^\star) - a(\mathcal{P}_2^\star)$, or
  - $a(\mathcal{P}_1)/a(\mathcal{P}_2)$ to $a(\mathcal{P}_1^\star)/a(\mathcal{P}_2^\star)$, or
  - some other measure of difference among the sub-populations.

Example (Australian vs. US shark encounters)

- Considering the variable in shark length, to measure or summarize differences between the Australia and US sub-populations attributes, we might calculate

$$\overline{y}_1 - \overline{y}_2 \qquad \text{and/or} \qquad \frac{SD\left(\mathcal{P}_1\right)}{SD\left(\mathcal{P}_2\right)}$$

```
round( c( mean(pop$pop1[,"Length"]) - mean(pop$pop2[,"Length"]),
sd(pop$pop1[,"Length"])/sd(pop$pop2[,"Length"]) ), 3)
```

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

- Then we can shuffle or mix the two populations ($N_1 = 28$ and $N_2 = 19$) and then compare the same measures. i.e. calculate

$$\overline{y}_1^{\star} - \overline{y}_2^{\star} \qquad \text{and/or} \qquad \frac{SD\left(\mathcal{P}_1^{\star}\right)}{SD\left(\mathcal{P}_2^{\star}\right)}$$

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

round( c( mean(mixedPop$pop1[,"Length"]) - mean(mixedPop$pop2[,"Length"]),
sd(mixedPop$pop1[,"Length"])/sd(mixedPop$pop2[,"Length"]) ), 3)
```

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

- Is this difference unusual?
- 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)
  - We will discuss such tests from a computational prospective.

Aside: 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 can be obtained using the functions below:

```
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 (comparing Australian vs. US shark encounters)

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

- For US and Australia populations.

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

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

- For shuffled populations.

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

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

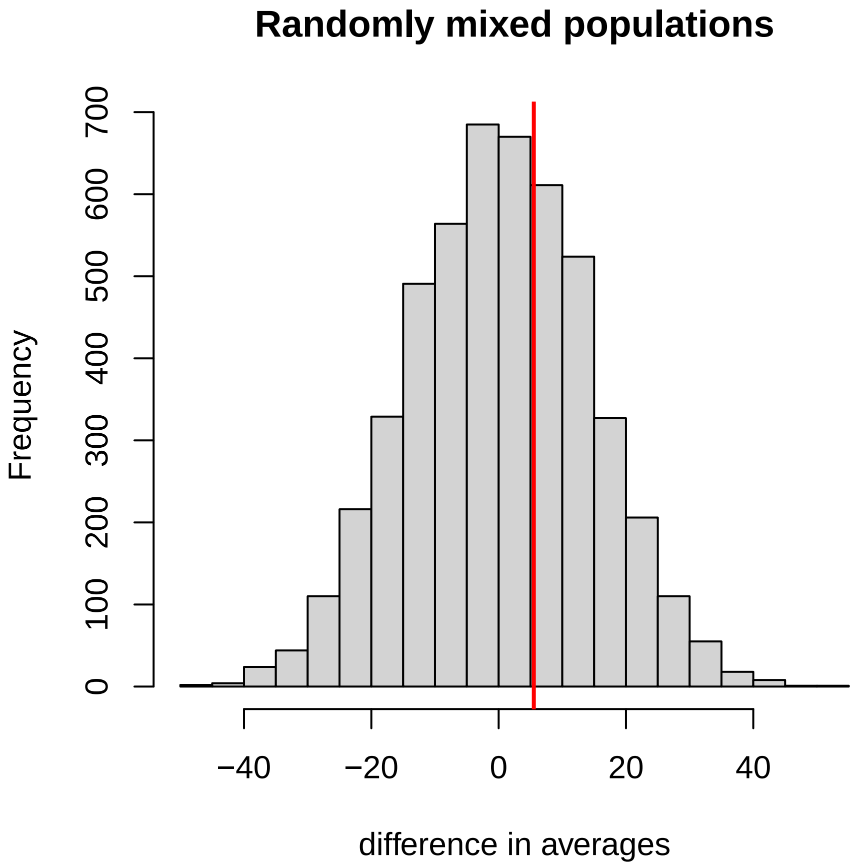
This matches the computations we did above, confirming the code is doing what it is expected to.

Shuffing the Populations (Australia vs. USA)

- To see how unusual the given pair of sub-populations are to any randomly shuffled pair, we can perform a simulation study.
- Ideally, we could look at all possible shufflings.
  - This requires about $\binom{N_1+N_2}{N_1} = \binom{28+19}{19} = 6.97 \times 10^{12}$ shuffles for the shark length data.
  - We use 5,000 shuffles instead.

- For each shuffled population we calculate $\overline{y}_1 - \overline{y}_2$

```
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)
```

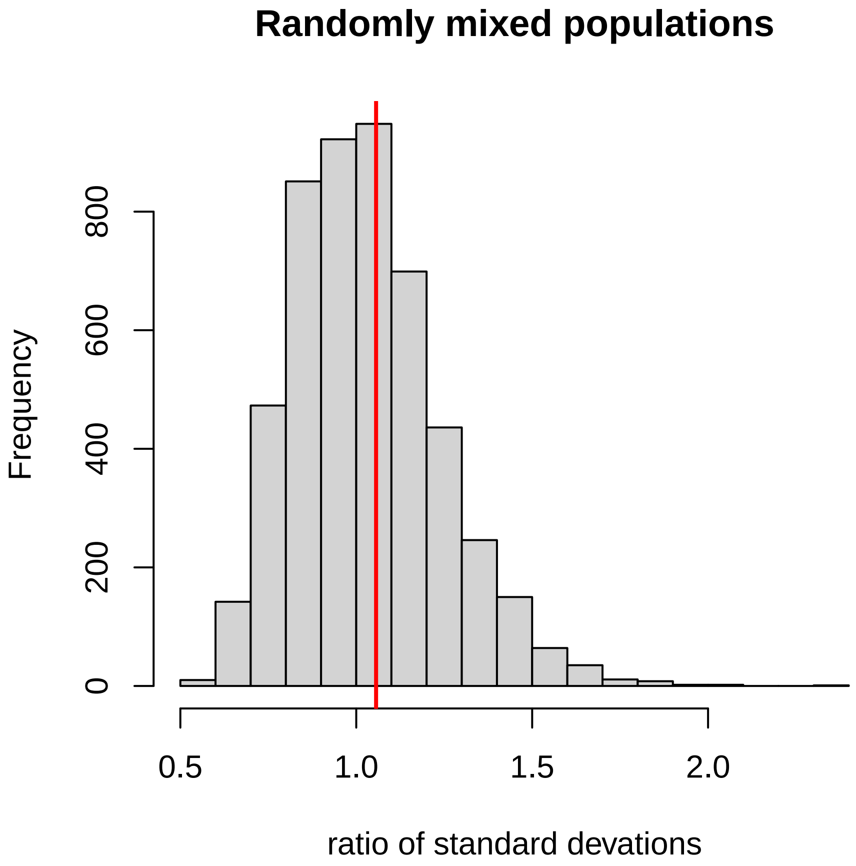


- The red line represent the difference between the shark length mean of the two populations, i.e. $a(\mathcal{P}_{Australia}) - a(\mathcal{P}_{USA})$.
- Is $a(\mathcal{P}_{Australia}) - a(\mathcal{P}_{USA})$ any different from the randomly mixed differences $a(\mathcal{P}_1^\star) - a(\mathcal{P}_2^\star)$?

Standard Deiviation

- To see how unusual the given pair of sub-populations are to any randomly shuffled pair, we perform a simulation study.
- This is similar to what we did with the above, except we looked at the difference between averages, but we will look at the ratio between the standard deviations here.

```
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)
```



- The red line represent the ratio of shark length standard deviation of the two populations, i.e. $SD(\mathcal{P}_{Australia})/SD(\mathcal{P}_{USA})$.
- What do you learn from this graph?

Difference in Surfing

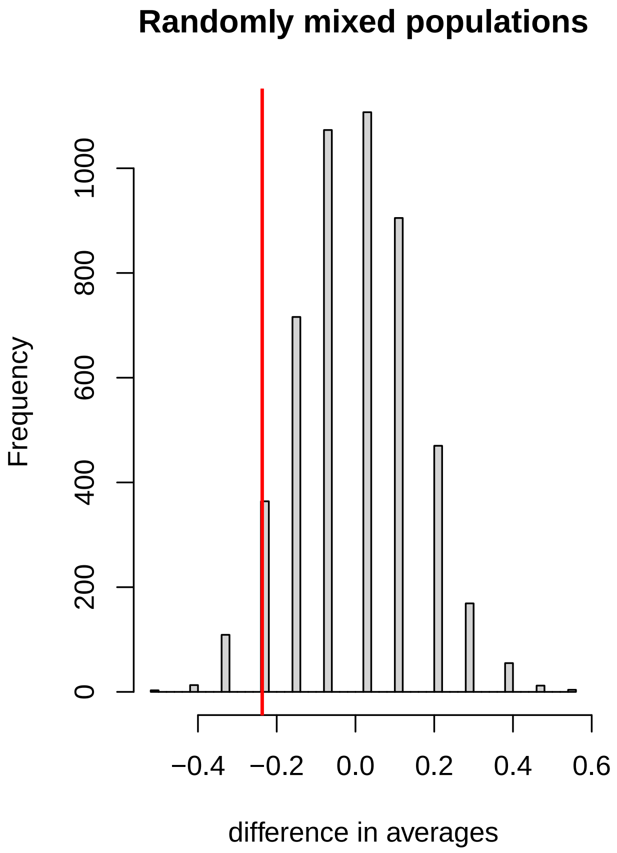
- Comparing the shark encounters involving Surfing from Australia and the USA.
    - The functions below performs the comparisons:

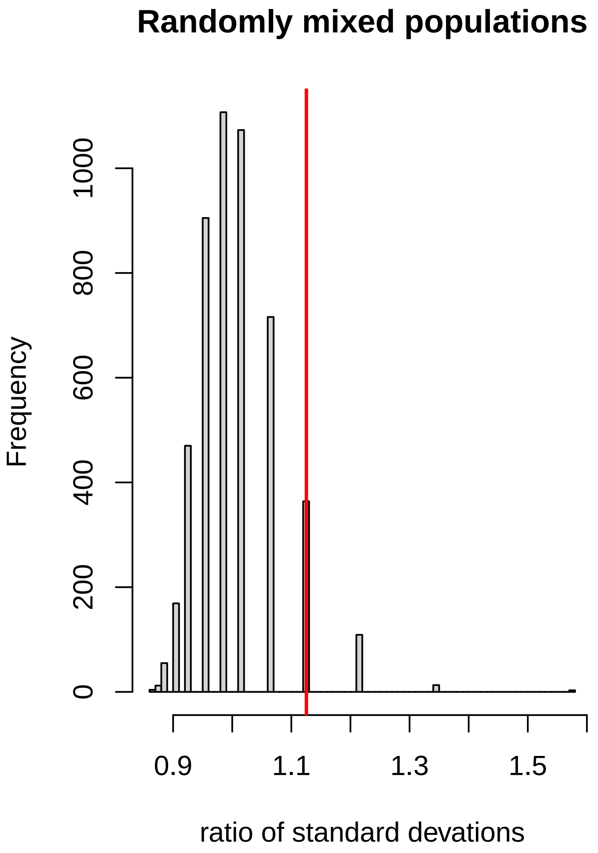
```
diffAveSurf <- getAveDiffsFn("Surfing")
ratioSDSurf <- getSDRatioFn("Surfing")
```

Then perform the mixing and summarize with a plot.

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

set.seed(341)
```

```r
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="lightgrey")
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)
```





What do you learn from the plots above?

Differences in Shark Length with Fatality

- The two sub-populations are

  - Fatal shark encounters and
  - Non-Fatal shark encounters

The codes below construct the two sub-populations:

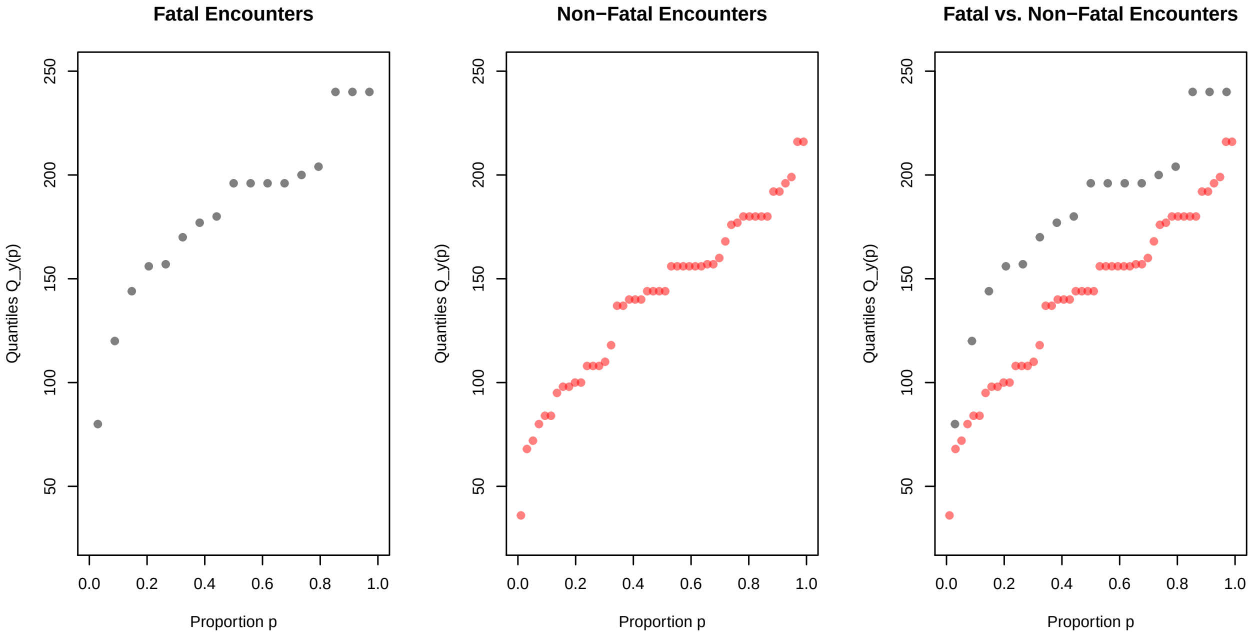
```
Fatpop <- list(pop1 = sharks[sharks[,"Fatality"] ==1, ],
               pop2 = sharks[sharks[,"Fatality"] ==0, ])
```

- We can compare the sharks lengths from the two populations.

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

```
## $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
```

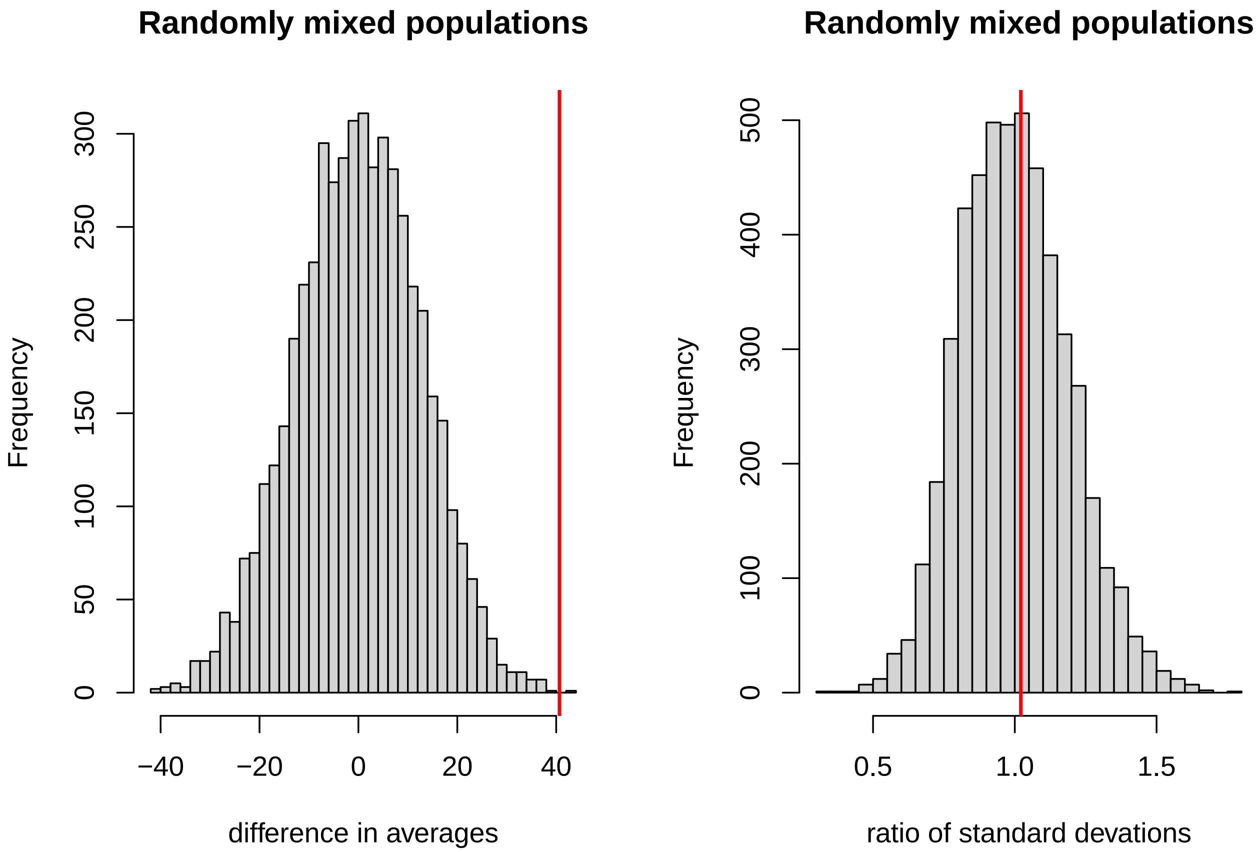
- A quantile plot of sharks from the two populations:



- It seems that fatal encounters involve bigger sharks compared to non-fatal encounters. (how did we come to this conclusion?)

- 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)
```

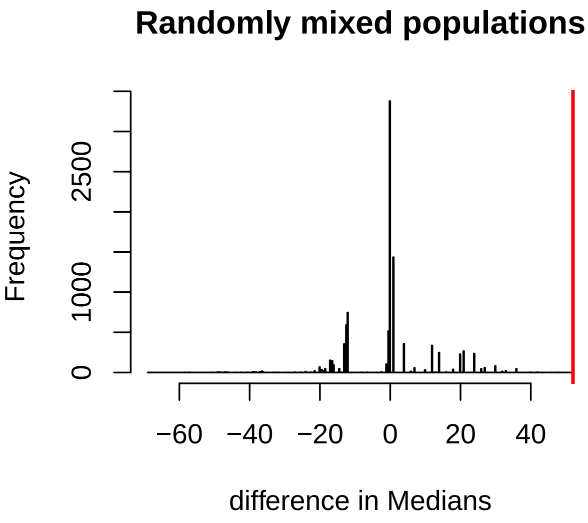


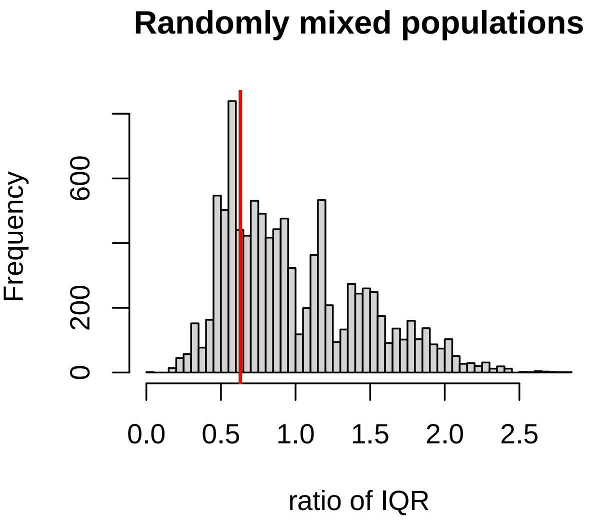
- What do these plots say about shark encounters involving fatal and non-fatal encounters?

- We can also compare other attributes, e.g.
  - We can quantify the difference in the median and IQR of the shark lengths from the two populations by randomly mixing the sub-populations.
  - But we need to construct these functions:

```
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.





What do you learn from these two plots?

4.3.1 Anatomy of a test of significance

- We would like to quantify, numerically, how unusual is the difference between the population averages relative to randomly mixed sub-populations.

  - If the two sub-populations are actually different, we want to provide numerical evidence against the idea that the two sub-populations are similar to randomly shuffled sub-populations.

- We use the following steps to gather this 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** to quantify how much the data is inconsistent 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 given (or unshuffled) sub-populations
4. Finally, we obtain the **observed significance level**
   - by finding probability that a randomly mixed (or shuffled) sub-population has a larger discrepency measure than the observed discrepancy.
   - We interpret this level as evidence against the null.

The null hypothesis

- The null hypothesis can be stated in several ways.
  - The sub-populations, $\mathcal{P}_1$ & $\mathcal{P}_2$, were randomly drawn from the same population.
  - $\mathcal{P}_1$ & $\mathcal{P}_2$ were generated by random mixing.
  - $\mathcal{P}_1$ & $\mathcal{P}_2$ were created by randomly assigning units in the same population to one or other of the sub-populations.

- 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.

- Note we cannot say that the two populations $\mathcal{P}_{Australia}$ and $\mathcal{P}_{USA}$ are equal in terms of their attribute values, i.e. $a(\mathcal{P}_{Australia}) = a(\mathcal{P}_{USA})$.
- The null hypothesis assumes the two 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

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

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

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

Discrepency Measure

- A *discrepency measure* is used to quantify how much the data is inconsistent with the null hypothesis, where large values indicate evidence against the null hypothesis.
  - It is an attribute for the popuation $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$.
  - We might consider properties such as equivariance and invariance.

- If we hypothesized that the averages from the two sub-populations were similar,
  - then a discrepency measure for this might be

$$D(\mathcal{P}_1, \mathcal{P}_2) = |\overline{y}_1 - \overline{y}_2|$$

- If we wanted evidence that the average from the first population was larger than the average from the second population,
  - then a discrepency measure for this might be

$$D(\mathcal{P}_1, \mathcal{P}_2) = \overline{y}_1 - \overline{y}_2$$

- If we hypothesized that the standard deviation from the two sub-populations were the same,
  - then a discrepency measure for this might be

$$D(\mathcal{P}_1, \mathcal{P}_2) = \left| \frac{SD(\mathcal{P}_1)}{SD(\mathcal{P}_2)} - 1 \right|$$

The Observed Discrepancy

- The observed discrepancy, d_{obs}, is the discrepancy measure, D, on the two given (or unshuffled) sub-populations

$$d_{obs} = D(\mathcal{P}_1, \mathcal{P}_2)$$

- The discrepancy measure quantifies only one type of discrepancy between the populations
  - e.g. shark length.
  - Any other differences are completely ignored.

The observed significance level

- The **observed significance level**, denoted by SL, is the probability that a randomly mixed (or shuffled) sub-population has a larger discrepancy measure than the observed discrepancy.

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

- If SL is very small then either
  - the hypothesis is true and we have observed a very unusual value of d_{obs},
  - or, the hypothesis is false.

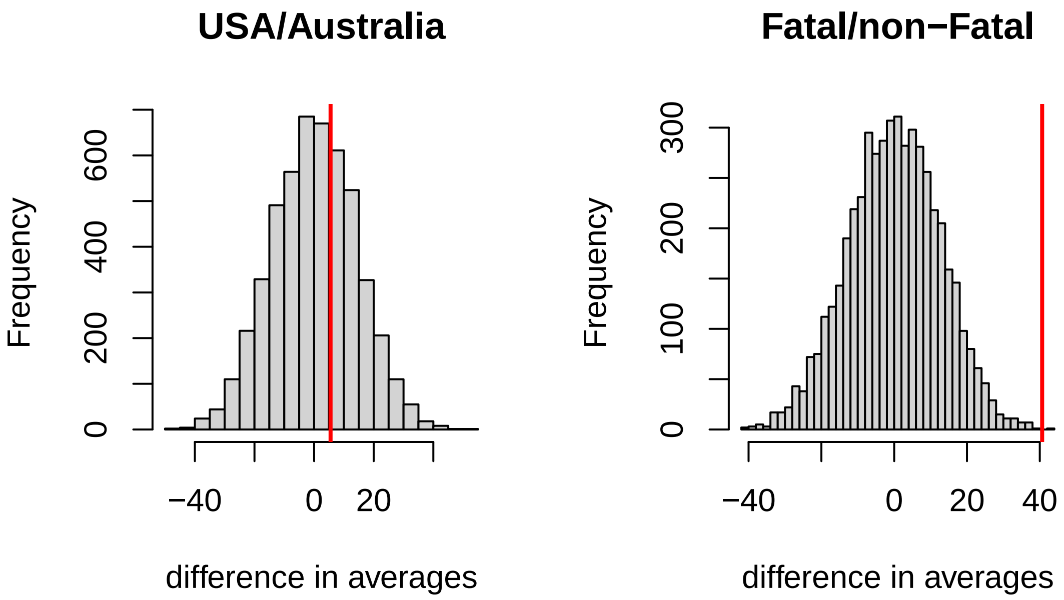
- The smaller is SL, the greater the evidence against the hypothesis.
  - In the extreme case where $SL = 0$, then we have observed something impossible and the hypothesis must therefore be false – this would be a proof by contradiction.

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

Calculating SL

- We may not be able to enumerate all possible permutations, so may not have the exact value of SL.
- Instead, we approximate SL by
  - generating a sample of shuffled pairs,
  - determining the discrepancy measure on each shuffled pair
  - calculating the proportion of the calculated discrepancy measures which are bigger than the observed discrepancy measure.

- In R speak, since we already have a sample of shuffled pairs, we can calculate this approximation with `sum(abs(diffLengths) >= abs(diffAveLengths(pop))) / length(diffLengths)`
  - For US and Australia Shark Lengths gives $SL \approx \widehat{SL} = 0.704$.
  - For Fatal and non-Fatal shark Lengths gives $SL \approx \widehat{SL} = 0.0002$



Interpretation

- 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.

- A large SL $(\mathcal{P}_{Australia}, \mathcal{P}_{USA})$ indicates
  - there is **no evidence against the null 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 Algorithm

1. State the **null hypothesis** that $\mathcal{P}_1$ 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**,
   - e.g. $D(\mathcal{P}_1, \mathcal{P}_2) = |a(\mathcal{P}_1) - a(\mathcal{P}_2)|$
3. Calculate the **observed discrepancy** $d_{obs} = D(\mathcal{P}_1, \mathcal{P}_2)$.
4. Find the **observed significance level**, SL,

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

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

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

Some More Important Things

- 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**
  - 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
  - e.g., for sharks lengths data, the difference between mean sharks length in fatal and non-fatal encounters was 3 and 1/4 ft.
- 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.

Errors

- In Court

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

A note on the language:

- we will try to avoid using terms such as "reject," "fail to reject," "accepts," etc. Instead, we quantify the risk in taking the action of "rejecting" H_0.

4.3.2 A t-like discrepancy measure

- 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"
  - 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.

- 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 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 denominator

$$\widetilde{SD}\,(a(\mathcal{P}_1) \ - \ a(\mathcal{P}_2)) = \sqrt{\widetilde{Var}(a(\mathcal{P}_1)) \ + \ \widetilde{Var}(a(\mathcal{P}_2))}$$

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.

Differences in Averages

Suppose we were interested in differences in averages, i.e. we have $a(\mathcal{P}_i) = \overline{Y}_i$ and $\mathcal{P}_i$ has size n_i, $i = 1, 2$ then we might use

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

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 from each of $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively,
    - then the pooled estimator of σ would be

$$\widetilde{\sigma} = \left(\frac{(n_1-1)\widetilde{\sigma}_1^2 + (n_2-1)\widetilde{\sigma}_2^2}{(n_1-1)+(n_2-1)}\right)^{\frac{1}{2}}.$$

    - When is this estimate of the standard deviation acceptable, and what other estimate is available?

Gausssian Assumption?

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

- 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.
    - Note that if the standard deviation of the two populations $\mathcal{P}_1$ and $\mathcal{P}_2$ are not assumed the same, then the other discrepency measure $D(\mathcal{P}_1, \mathcal{P}_2) = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{\widetilde{\sigma}_1^2}{n_1} + \frac{\widetilde{\sigma}_2^2}{n_2}}}$ may be used.

Calculating the Obseraved Values

- Below is a function that will return this discrepancy measure (assuming the variance of sub-populations are equal) 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
  }
}
```

- Get the this t-like discrepancy measure 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))
```

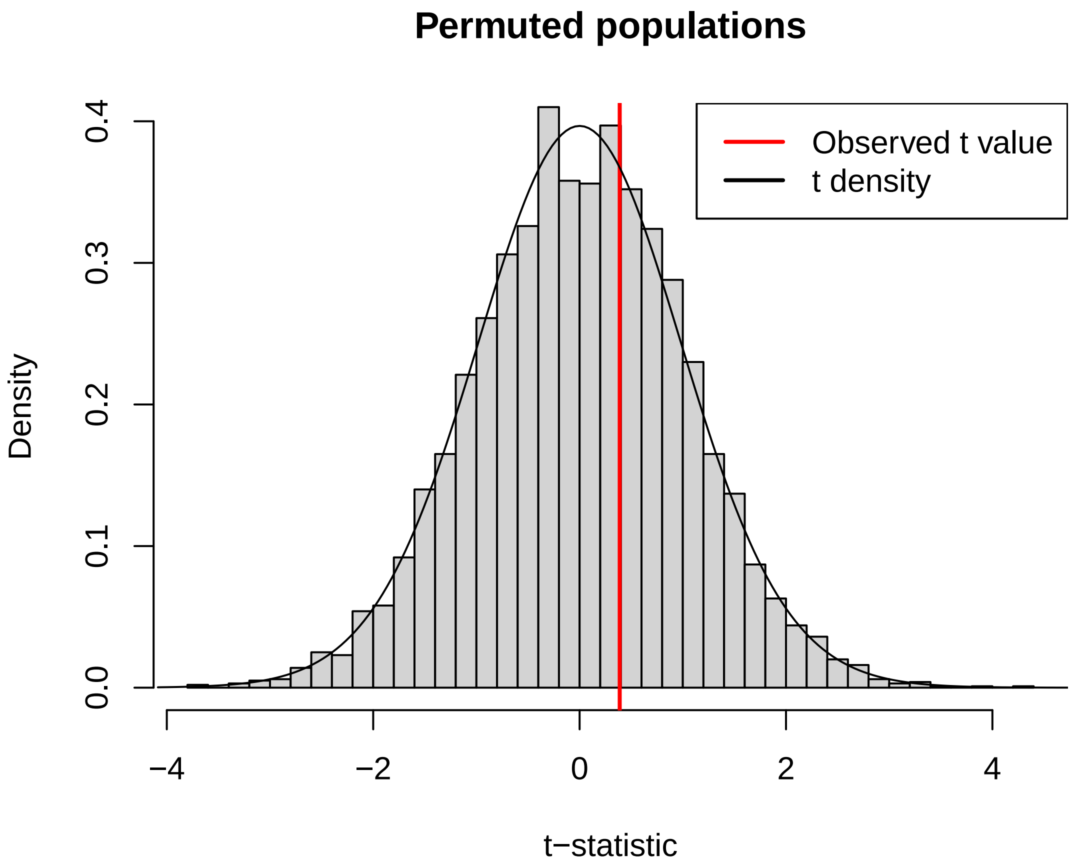
```
## [1] 0.3886752 3.4454919
```

- 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.

US and Australia sub-populations

- The Histogram of discrepancy measures of the
  - shark lengths of the US and Australia sub-populations with
  - t-density with $n_1 + n_2 - 2$ degrees of freedom.

```r
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 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 observed significance level can now be estimated two ways:

1) using randomly generated sub-populations we obtain

```
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 we obtain

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

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

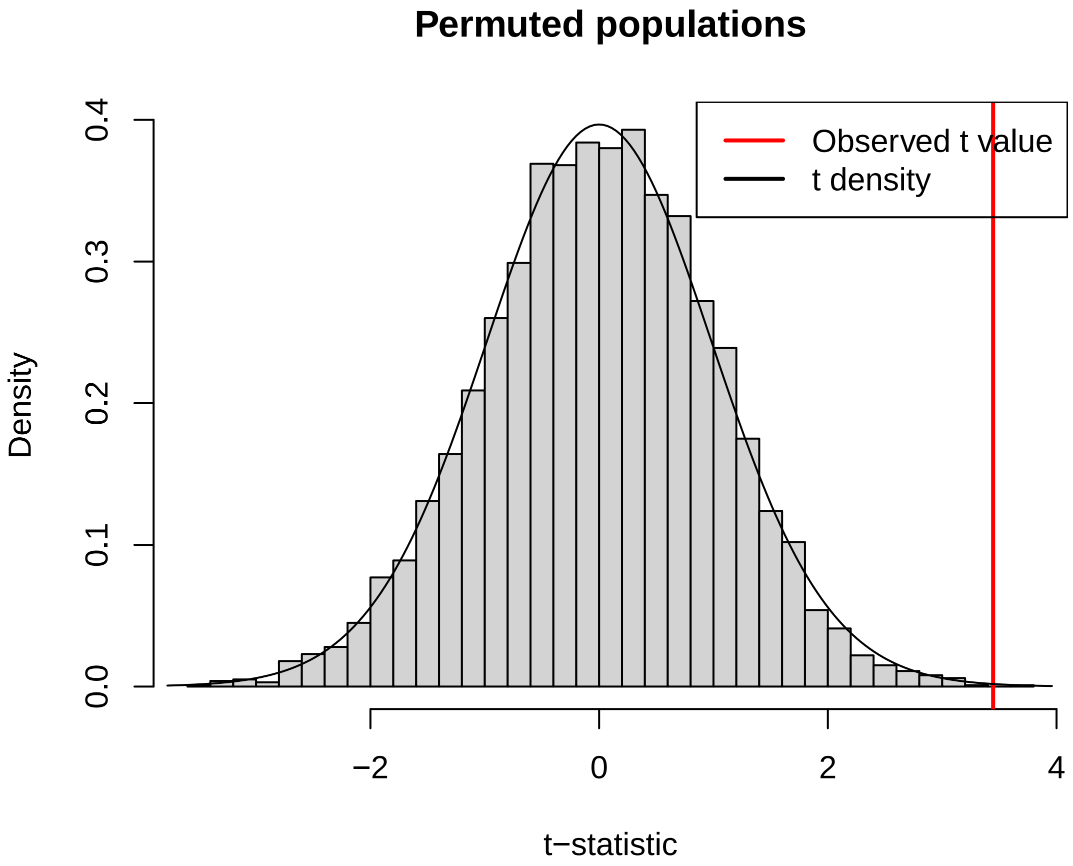
- The closeness of these two numbers shows that, in this particular example, t distribution could be a good approximation (though we are not using it!)
- The observed significance level is so large that the observed discrepancy measure is not at all unusual when the hypothesis is true.
  - This test provides **no evidence against the null hypothesis**.

- From a practical point of view, this means that the two sub-populations US and AUstralia are not different from the length of sharks point of view.

Fatal and non-fatal sub-populations

- The Histogram of discrepancy measures of the
  - sharks lengths from the fatal and non-fatal sub-populations with t-density with $n_1 + n_2 - 2$ degrees of freedom.

```
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"))
```



The observed significance level can now be estimated two ways:

1) using randomly generated sub-populations we obtain

```
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 we obtain

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

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

- Notice 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 SL is, in fact, the tail of the distribution.
- 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 US and Australia seem to be different from fatality of the enncounters point of view.

4.3.3 Multiple Testing with Random Noise

- We might consider any number of discrepancy measures, $D_1, D_2, \ldots, D_K$ to compare two sub-populations
    - each with an associated observed significance level say $SL_1, SL_2, \ldots, SL_K$.

- Each SL value is a measure of evidence against the null hypothesis, but
    - we cannot consider these individually because we are aware of type I and type II error.

- We have to consider them collectively... but first an example.

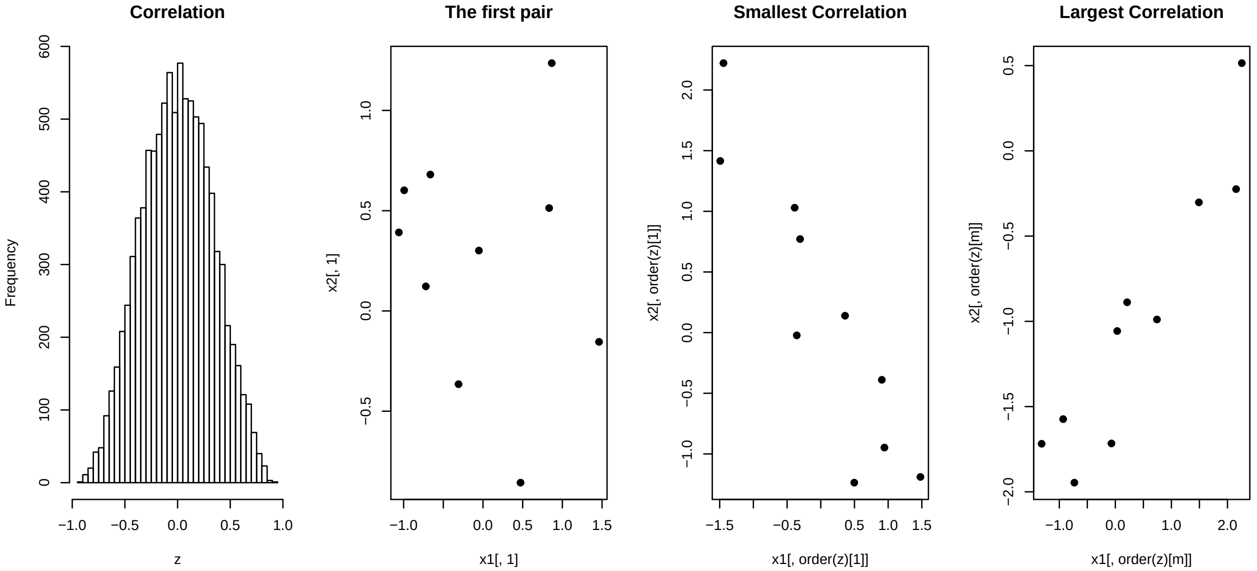
Example: Correlation Coefficient for Random Noise

- 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 (code below).
    - each $\mathbf{x}_{ij}$ is a vector, representing realizations of a variable j within sub-population i.

```r
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 $j = 1, \ldots, m$ and generate some plots.

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

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 independent, there are samples in the simulation with large sample correlation coefficients (both large positive and large negative correlations).
  - note: the first pair (second panle in the plots above) is equivalent to a randomly selected pair.

Multiple testing (restart)

- We might consider any number of discrepancy measures, $D_1, D_2, \ldots, D_K$
  - each with an associated observed significance level say $SL_1, SL_2, \ldots, SL_K$.
  - Fortunely, unlike the several discrepancy measures, these significance levels **are on a common and interpretable scale** (probability).

- To consider the SL collectively, and 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 null hypothesis, i.e.

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

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

- Note: 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} a discrepancy measure, we let

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

  - $D^{\star}$ is arranged 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 significance level that describes this combined evidence is denoted by

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

- $SL^{\star}$ will be larger than SL_{min} because
  - SL_{min} is the smallest significance level among $SL_1, SL_2, \ldots, SL_K$ and so
  - SL_{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 $SL^{\star}$ can be calculated.

Recall: The SL with a single discrepancy measure

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

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

  - where d_{obs} is the observed discrepancy measure based on the given sub-populations, $\mathcal{P}_1$ & $\mathcal{P}_2$.

- 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

- Suppose we only have **one** discrepancy measure.
- For $i = 1, \ldots, M$
  - randomly construct two sub-populations, $\mathcal{S}_{i1}$ & $\mathcal{S}_{i2}$, while maintaining the sub-population sizes.

  - this can be done by sampling without replacement from the population $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$ and calculate $d_i = D(\mathcal{S}_{i1}, \mathcal{S}_{i2})$

- then we estimate the SL with
$$\widehat{SL} = \frac{1}{M} \sum_{i=1}^{M} \mathrm{I}\left(d_i \geq d_{obs}\right)$$

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

Estimating $d^\star_{obs}$ in multiple testing

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

- For $i = 1, \ldots, M$
  - randomly construct two sub-populations, $\mathcal{S}_{i1}$ & $\mathcal{S}_{i2}$, from $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$
  - calculate $d_{ik} = D_k(\mathcal{S}_{i1}, \mathcal{S}_{i2})$
  - then 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 $SL^\star$

- Suppose we have K discrepancy measures $D_1, D_2, ..., D_K$.
  - Then the combined discrepancy measure is
$$d^\star = 1 - SL_{min} = 1 - \min_{k=1,\ldots,K} SL_k$$

- We estimate $\widehat{d}^\star_{obs}$ based on the given sub-populations, $\mathcal{P}_1$ & $\mathcal{P}_2$ (see previous subsection).
- To estimate $SL^\star$ we repeat the following steps $M^\star$ times
  - randomly construct two sub-populations, $\mathcal{S}_{i1}$ & $\mathcal{S}_{i2}$, from $\{\mathcal{P}_1, \mathcal{P}_2\}$ and
  - estimate $d^\star_i$ by the same procedure used to calculate $d^\star_{obs}$ (see the previous subsection) but now using $\mathcal{S}_{i1}$ & $\mathcal{S}_{i2}$ as the given sub-populations.

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

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

R code Example

- 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
}
```

- Let us compare the encounters happened in Australia versus the USA.
- We examine two discrepancies: average and standard deviation.
  - this is multiple testing, because we have two discrepancy measures involving in the test.

```
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 US and Australian encounters were randomly drawn from the same population based on the average and standard deviation of shark lengths.
  - increase the `B_outer` and `B_inner` values above to get a more accurate estimate of the significance level (computationally intensive though).

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