Permutation tests

Peter Ralph

13 October – Advanced Biological Statistics

Permutation tests

```
##
##  Welch Two Sample t-test
##
## data:  airbnb$price[airbnb$instant_bookable] and airbnb$price[!airbnb$instant_bookab
## t = 3.6482, df = 5039.8, p-value = 0.0002667
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   4.475555 14.872518
## sample estimates:
## mean of x mean of y
##  124.6409  114.9668
```

But, the t test relies on *Normality*. Is the distribution of AirBnB prices too “weird”? How can we be sure?

```
##
##  Welch Two Sample t-test
##
## data:  airbnb$price[airbnb$instant_bookable] and airbnb$price[!airbnb$instant_bookab
## t = 3.6482, df = 5039.8, p-value = 0.0002667
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   4.475555 14.872518
## sample estimates:
## mean of x mean of y
##  124.6409  114.9668
```

But, the t test relies on *Normality*. Is the distribution of AirBnB prices too “weird”? How can we be sure?

Methods:

1. Remove the big values and try again.

2. Use a nonparametric test.

The permutation test

Observation: **If** there was no meaningful difference in prices between “instant bookable” and not, **then randomly shuffling that label** won’t change anything.

The permutation test

Observation: **If** there was no meaningful difference in prices between “instant bookable” and not, **then randomly shuffling that label** won’t change anything.

Strategy:

1. Shuffle the `instant_bookable` column.
2. Compute the difference in means.
3. Repeat, many times.
4. Compare: the p-value is the proportion of “shuffled” values more extreme than observed.

The permutation test

Observation: **If** there was no meaningful difference in prices between “instant bookable” and not, **then randomly shuffling that label** won’t change anything.

Strategy:

1. Shuffle the `instant_bookable` column.
2. Compute the difference in means.
3. Repeat, many times.
4. Compare: the p-value is the proportion of “shuffled” values more extreme than observed.

Why is this a p-value? For what hypothesis?

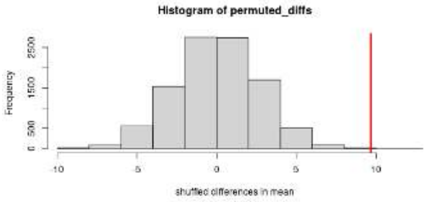
Shuffle once

```
fake_is_instant <- sample(airbnb$instant_bookable)
(mean(airbnb$price[fake_is_instant], na.rm=TRUE) -
                mean(airbnb$price[!fake_is_instant], na.rm=TRUE))
```

```
## [1] 2.837541
```

Many times

```
real_diff <- (mean(airbnb$price[airbnb$instant_bookable], na.rm=TRUE)
               - mean(airbnb$price[!airbnb$instant_bookable], na.rm=TRUE))
permuted_diffs <- replicate(10000, {
        fake_is_instant <- sample(airbnb$instant_bookable)
        (mean(airbnb$price[fake_is_instant], na.rm=TRUE)
             - mean(airbnb$price[!fake_is_instant], na.rm=TRUE))
   } )
hist(permuted_diffs, xlab="shuffled differences in mean", xlim=range(c(permuted_diffs,
        real_diff)))
abline(v=real_diff, col='red', lwd=3)
```



How surprising was the real value?

```
mean(abs(permuted_diffs) > abs(real_diff))
```

```
## [1] 3e-04
```

How surprising was the real value?

```
mean(abs(permuted_diffs) > abs(real_diff))
```

```
## [1] 3e-04
```

> *The difference in price between instant bookable and not instant bookable is highly statistically significant ($p \approx 0.001$, permutation test).*

Our turn

Let’s do the analogous thing for the ANOVA comparing price between neighbourhoods:

```
anova(lm(price ~ neighbourhood, data=airbnb))
```

```
## Analysis of Variance Table
##
## Response: price
##                 Df   Sum Sq Mean Sq F value    Pr(>F)
## neighbourhood   91  6015248   66102  7.6277 < 2.2e-16 ***
## Residuals     5510 47749952    8666
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
// reveal.js plugins
```

/