During today’s practice we will investigate how to perform **A/B Testing**. A/B testing is used in numerous ways to test different versions of web pages, UX, surveys and questionnaires, changes in policies, different marketing campaigns, emails and so on.

Broadly speaking, A/B tests are run mostly on two types of data:

- Continuous or discrete numbers, for example **average number of clicks, time spent on the page**
- Proportions or percentages, for example, **conversion rates**

For the first data type, the **t-test** is most frequently used. For the second type of data, the **Pearson’s Chi-squared test** is the obvious choice. Let’s take a look at both cases.

Let us take a look at the following data:

```r
library(data.table)
library(dplyr)
library(ggplot2)
setwd("path of your file")
dt <- fread("AB_clicks.csv")
dt$Version <- as.factor(dt$Version)
head(dt)
```

<table>
<thead>
<tr>
<th>Element_ID</th>
<th>Tag_name</th>
<th>Name</th>
<th>No_clicks</th>
<th>Visible</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>area</td>
<td>Montana State University - Home</td>
<td>1291</td>
<td>FALSE</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>FIND</td>
<td>842</td>
<td>TRUE</td>
</tr>
<tr>
<td>3</td>
<td>input</td>
<td>s.q</td>
<td>508</td>
<td>TRUE</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>lib.montana.edu/find/</td>
<td>166</td>
<td>TRUE</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>REQUEST</td>
<td>151</td>
<td>TRUE</td>
</tr>
<tr>
<td>6</td>
<td>a</td>
<td>Hours</td>
<td>102</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Interact</td>
</tr>
<tr>
<td>2: Interact</td>
</tr>
<tr>
<td>3: Interact</td>
</tr>
<tr>
<td>4: Interact</td>
</tr>
<tr>
<td>5: Interact</td>
</tr>
<tr>
<td>6: Interact</td>
</tr>
</tbody>
</table>
This is the cleaned version of the data from [https://scholarworks.montana.edu/xmlui/handle/1/3507](https://scholarworks.montana.edu/xmlui/handle/1/3507). University of Montana explored that the button *Interact* on their page is heavily underused. They surveyed the problem by conducting questionnaires and realized that the name might be one of the reasons being too intimidating. They came up with several other versions:

**t-test**

Let’s try to perform t-test like many people do (A is default and B the *Connect* version):

```r
dt_cleaned <- filter(dt, Tag_name!='area')
dt_interact_connect <- filter(dt_cleaned, Version %in% c("Interact", "Connect"))
t.test(No_clicks ~ Version, data=dt_interact_connect) # where No_clicks is numeric and Version is a binary factor
```


```r
##
## Welch Two Sample t-test
##
## data:  No_clicks by Version
## t = -0.50683, df = 118.96, p-value = 0.6132
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -45.37063  26.87786
## sample estimates:
## mean in group Connect mean in group Interact
##     26.38596     35.63235
```

**Task** (1 pt): What can we conclude?

**Something to ponder:** Did we perform it correctly?
# Is it gaussian distribution?

```r
ggplot(dt_cleaned, aes(x=No_clicks, fill=Version)) + geom_density(alpha=0.3) + theme_bw()
```
# May be this one better?

ggplot(dt_cleaned, aes(x=No_clicks, fill=Version)) + geom_density(alpha=0.3) + theme_bw() + scale_x_log10()
# common ways to test for normality

# qq plot

```{r}
library(nortest)
qqnorm(filter(dt_cleaned, Version=='Interact')$No_clicks, cex=0.5)
qqline(filter(dt_cleaned, Version=='Interact')$No_clicks, col = 2)
```

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. See reference [3]
Statistical tests for normality can be easily formulated in the framework of hypothesis testing. The null hypothesis is that the data is normally distributed. Here is the catch: we assume it by default, we are not trying to prove this. Basically, we can only reject (or accept) the hypothesis that data is not statistically distributed (if it is not significant, it is either normally distributed or we do not have enough data to reject it).

```r
ad.test(filter(dt_cleaned, Version=='Interact')$No_clicks)
```

```r
## Anderson-Darling normality test
## data:  filter(dt_cleaned, Version == "Interact")$No_clicks
## A = 17.76, p-value < 2.2e-16
```

Anderson-Darling Test: The test rejects the hypothesis of normality when the p-value is less than or equal to 0.05. See reference [4]

```r
shapiro.test(filter(dt_cleaned, Version=='Interact')$No_clicks)
```

```r
## Shapiro-Wilk normality test
## data:  filter(dt_cleaned, Version == "Interact")$No_clicks
## W = 0.29608, p-value < 2.2e-16
```

Low value of W means the test has failed and so thus p value. See reference [5].

**Task** (1 pt) Repeat these tests for the modified feature for Number of clicks by converting it into log. It can help to transform feature to normally distributed one.

What to do when the data is not gaussian (not normally distributed)? There are two large groups of tests: **parametric** - they have assumptions about distributions, and **non-parametric**, where you can forget about the distribution assumptions. However, it comes with a catch - often their **power** is lower. We can use **wilcoxon test for normality** instead of t-test (as the Normality assumption is not fulfilled):

```r
wilcoxon.test(No_clicks ~ Version, data=dt_interact_connect)
```

```r
## Wilcoxon rank sum test with continuity correction
## data:  No_clicks by Version
## W = 1600, p-value = 0.09027
## alternative hypothesis: true location shift is not equal to 0
```
wilcox.test(log_clicks ~ Version, data=dt_interact_connect) # same result

##
## Wilcoxon rank sum test with continuity correction
##
## data: log_clicks by Version
## W = 1600, p-value = 0.09027
## alternative hypothesis: true location shift is not equal to 0

Task (2 pt): Let’s try to change the hypothesis. Let’s check only those clicks on the objects, when the Visibility=TRUE. What are the results?

Currently, we do not have the possibility to plan the experiment and make decisions about the sample sizes. In real-life, you first need to calculate how many samples you need with respect to significance, power and effect size:

library(pwr)
pwr.t.test(d = 0.2, sig.level = 0.05, power = 0.8)

##
## Two-sample t test power calculation
##
## n = 393.4057
## d = 0.2
## sig.level = 0.05
## power = 0.8
## alternative = two.sided
##
## NOTE: n is number in *each* group

Chi-squared test

The hypothesis that in general one version of the page is more clicked than another one is quite optimistic. Let’s narrow down our hypothesis. What if we want to check whether the number of times clicked on this component (*Connect*) out of all clicks to this page is significantly better (worse) than the same proportion of clicks in the default version (Interact)? Here we test proportions. See reference [6] and [7] for more on this.
total_clicks <- group_by(dt, Version) %>%
  summarise(total = sum(No_clicks))  # total number of clicks
dt_button <- filter(dt, Name %in% c("SERVICES", "HELP", "LEARN","CONNECT","INTERACT")) %>%
  # choose only areas we focus
  left_join(total_clicks, by = "Version") %>%
  mutate(proportions = No_clicks/total)# combine
dt_button

```r
## Element_ID Tag_name Name No_clicks Visible Version total
## 1 87 a INTERACT INTERACT 42 TRUE 3714
## 2 92 a CONNECT CONNECT 53 TRUE 1587
## 3 87 a LEARN LEARN 21 TRUE 1652
## 4 92 a HELP HELP 38 TRUE 1717
## 5 87 a SERVICES SERVICES 45 TRUE 1348

## proportions
## 1 0.01130856
## 2 0.03339635
## 3 0.01271186
## 4 0.02213162
## 5 0.03338279
```
The assumption is that the two groups are **mutually exclusive**. In other words, a user can only see one of the versions. Null hypothesis is that the proportions are equal in both groups.

```r
prop.test(x=dt_button$No_clicks[c(1,2)], n=dt_button$total[c(1,2)])
```

```r
##
## 2-sample test for equality of proportions with continuity correction
##
## data:  dt_button$No_clicks[c(1, 2)] out of dt_button$total[c(1, 2)]
## X-squared = 29.579, df = 1, p-value = 5.368e-08
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.03200864 -0.01216692
## sample estimates:
## prop 1  prop 2
## 0.01130856 0.03339635
```

If the value is less or equal to than (df = 1, and for 0.05) value than null hypothesis cannot be rejected. See reference [7]

**Task (1 pt): So, if the null hypothesis can be rejected or not?**

There are many tests, but all of them fit into the general framework!
References:

2) https://www.statmethods.net/stats/ttest.html: Look for different types of t-test
3) http://data.library.virginia.edu/understanding-q-q-plots/ q-q plot
4) http://www.variation.com/da/help/hs140.htm anderson darling test
5) http://www.statisticshowto.com/shapiro-wilk-test/ shapiro test
6) https://www.youtube.com/watch?v=VskmMgXmkMQ Chi Squared Test
7) https://www.youtube.com/watch?v=qYOMO83Z1WU Chi Squared Test