Section 3: Permutation Inference

Yotam Shem-Tov

Fall 2015

Introduction

- Throughout this slides we will focus only on randomized experiments, i.e the treatment is assigned at random.
- We will follow the notation of Paul Rosenbaum and the book *Observational Studies*, which is highly recommended.

Fisher's exact inference

- Fisher introduced the idea of exact inference in the book, The Design of Experiments, in 1935.
- Exact inference uses only the random assignment of treatment to test the sharp null of no treatment effect,

$$H_0 : \tau_i = 0 \quad \forall\, i$$

- The test of the sharp null hypothesis is **distribution** and **model** free!
- The key elements in Fishers argument are:
  1. For a valid test of no treatment effect on the units included in an experiment, it is sufficient to require that treatment be allocated at random to experimental units - these units may be both heterogeneous in their responses and not a sample from a population.
  2. Probability enters the experiment only through the random assignment of treatments, a process controlled by the experimenter.

Introduction

As with any statistical hypotheses test, we need the following elements:

1. Data
2. Null hypothesis
3. Test statistic
4. The distribution of the test statistic under the null hypothesis

Definitions: basic setup

- Using Rosenbaums notation: there are N units divided into S strata or blocks, which are formed on the basis of pre-treatment characteristics
- There are n_s units in stratum s for $s = 1, ..., S$ so $N = \sum_{s=1}^{S} n_s$
- Define Z_{si} as an indicator variable whether the ith unit in stratum s receives treatment or control.

  If unit i in stratum s receives treatment, $Z_{si} = 1$ and if the unit receives control, $Z_{si} = 0$
- Define m_s as the number of treated units in stratum s, so $m_s = \sum_{i=1}^{n_s} Z_{si}$, and $0 \leq m_s \leq n_s$

Definitions: Unit

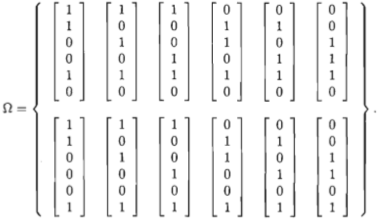
- We will simplify the notation and focus on the case in which there is only *one* strata, i.e $S = 1$ and $N = n_s$. The number of treated units is $m = \sum_{i=1}^{N} Z_{si}$
- What is a unit?
  Answer: *A unit is an opportunity to apply or withhold the treatment*
- A unit may be a person who will receive either the treatment or the control.
- A group of people may form a single unit: all children in a particular classroom or school .
- A single person may present several opportunities to apply different treatments, in which case each opportunity is a unit.

Notation

- Let $\mathbf{r} = (r_1, \ldots, r_N)$ be the vector of observed responses
- Let Ω be the set containing all possible treatment assignments
- Let $\mathbf{z} = (z_1, \ldots, z_N)$ be a treatment assignment, $z \in \Omega$, $z_i \in \{0, 1\}$

Treatment assignment

- The set Ω contains $K = \binom{N}{m}$, possible treatment assignments $\mathbf{z}$.
- In the most common experiments, each possible treatment assignments is given the same probability, $Pr(\mathbf{Z} = \mathbf{z}) = 1/K$ for all $\mathbf{z}$ in Ω.
- For example consider a randomized experiment with 2 strata, $S = 2$, four units in the first stratum, $n_1 = 4$, and two units in the second stratum, $n_2 = 2$. Half of the units in each stratum received treatment, $m_1 = 2$ and $m_2 = 1$. What is the set of all possible treatment assignment? $\Omega = \binom{4}{2} \cdot \binom{2}{1} = 12$



$$\Omega = \left\{ \begin{matrix} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \\ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \end{matrix} \right\}.$$

Treatment assignment: example

- Example 1: $N = 30$, given that the number of treated units is $m = 15$, how large is Ω? Is the allocation of treatment I.I.D?
  Answer: $cov(T_i, T_j) \neq 0$, the treatment assignment is not I.I.D

  ```
  > choose(30,15)
  [1] 155117520
  ```

- Example 2: $N = 30$, the treatment is assignment mechanism is, $T_i \sim Bernulli(p = 1/2)$. How large is Ω? Is the allocation of treatment I.I.D?
  Answer: The treatment assignment is I.I.D

  ```
  [1] 155117520
  > 2^30
  ```

- In example 2, Ω is 6.92 times larger than in example 1. We will usually consider the situation in which m is given

Sharp null hypothesis

- The most common hypothesis associated with randomization inference is the sharp null of no effect for all units
- A unit labeled as treated will have the exact same outcome as a unit labeled as control
- Let r_z be the vector of potential responses for randomization assignment z
- The sharp null hypothesis is that r_z is the same for all $\mathbf{z}$, $\forall_z \; r_z = r$
- *Under the null, the units responses are fixed and the only random element is the meaningless rotation of labels (between control and treatment)*

Test statistic

- A test statistic $t(Z, r)$ is a quantity computed from the treatment assignment Z and the response r.
- Consider the following test statistic: the difference in sample means for the treated and control groups:

$$t(Z, r) = \sum_{i=1}^{N} \left\{ \frac{Z_i r_i}{m} - \frac{(1 - Z_i) r_i}{N - m} \right\} = \frac{\sum_{i=1}^{N} Z_i r_i}{m} - \frac{\sum_{i=1}^{N} (1 - Z_i) r_i}{N - m}$$

and in matrix notation,

$$t(Z, r) = \frac{Z^T r}{Z^T 1} - \frac{(1 - Z)^T r}{(1 - Z)^T 1}$$

- Why is Z in a capital letter and r not? To indicate that under the null Z is a random variable and r is fixed

Hypothesis testing

- The hypothesis test of the sharp null,

$$H_0 : \ r_z = r$$

$$H_1 : \ r_z \neq r$$

- We seek the probability of a value of the test statistic as extreme or more extreme than observed, under the null hypothesis
- In order to calculate the $P - value$, we need to know (or approximate) the distribution of the test statistic
- The treatment assignment Z follows a known randomization mechanism which we can simulate or exhaustively list

Calculating significant level ($P - value$)

- Let T be the observed value of this test statistic. Suppose we would like to reject the null for large values of T . The p-value is,

$$Pr_{H_0}(t(Z, r) \geq T) = \sum_{z \in \Omega} I\left[t(z, r) \geq T\right] Pr_{H_0}(Z = z)$$

  where $I\left[t(z, r) \geq T\right]$ is an indicator whether the value of the test statistic under the treatment assignment z is higher than the observed test statistic, T

- Under the null, H_0, the treatment has no effect and hence r is fixed regardless of the assignment $\mathbf{z}$

Calculating significant level ($P - value$)

In the case that all treatment assignments are equally likely, $Pr_{H_0}(Z = z) = \frac{1}{|\Omega|}$ and,

$$Pr_{H_0}(t(Z,r) \geq T) = \frac{\sum_{z \in \Omega} I\left[t(z,r) \geq T\right]}{|\Omega|}$$

$$= \frac{|\{z \in \Omega : \ t(z,r) \geq T\}|}{|\Omega|}$$

The indicator variable $I\left[t(z,r) \geq T\right]$ is a random variable which is distributed, $B(n = 1, \ prob = P_{H_0}(I\left[t(z,r) \geq T\right]))$ (Bernoulli distribution)

$$\frac{|\{z \in \Omega : \ t(z,r) \geq T\}|}{|\Omega|} = \frac{1}{|\Omega|} \sum_{Z \in \Omega} I\left[t(Z,r) \geq T\right] = \mathbb{E}\left(I\left[t(z,r) \geq T\right]\right)$$

Calculating significant level (*P-value*)

When Ω is small we can exhaustively go over all the elements in Ω and calculate, $|\{z \in \Omega : \; t(z, r) \geq T\}|$ - as in the Lady tasting tea example

How can we calculate the *P-value* when Ω is to large to enumerate all possible treatment assignments?

1. Use a Monte-Carlo approximation
2. Use an asymptotic approximation for the distribution of the test statistic

Monte-Carlo approximation: step-by-step

1. Draw a SRS (simple random sample) of size m from the data and call it X (the treatment group), and call the rest of the data Y (the control group)
2. Compute the test statistic, $t(Z, r)$, as you would if X and Y would have been the originals data, denote this test statistic by $t^b(Z, r)$
3. Repeat this procedure B times (many times), saving the results, so you have:
$$t^1(Z,r), t^2(Z,r), t^3(Z,r), \ldots, t^B(Z,r)$$
4. The distribution of $t^b(Z, r)$ approximates the true distribution of $t(Z, r)$ under the null (the sharp null)
In particular, a p-value can be computed by using,
$$\frac{1}{B} \times \#\{b: \ t^b(z,r) \geq T\}$$

Monte-Carlo approximation: Theory

- Recall $P_{H_0}(I\left[t(z,r) \geq T\right]) = \mathbb{E}\left(I\left[t(z,r) \geq T\right]\right)$
- The intuitive estimator for the *P-value* is the proportion of times the indicator variable receives a value of 1 in the Monta-Carlo simulation:

$$\widehat{P - value} = \widehat{\mathbb{E}}\left(I\left[t(z,r) \geq T\right]\right) = \frac{1}{B}\sum_{b=1}^{B} I\left[t^{b}(z,r) \geq T\right]$$

where B is the number of samples

Example of permutation inference

- We want to compare x_1 and x_2, denote x_2 as the treatment group and x_1 as the control group. The treatment was allocated randomly

  ```
  > set.seed(13)
  > x1 = rexp(1000,rate=0.6)
  > x2 = rexp(1000,rate=0.5)
  ```

  The observed difference in means is,

  ```
  > mean(x2)-mean(x1)
  [1] 0.4204367
  ```

- In order to calculate a significant level we need to know (or approximate) the distribution of $t(Z, r)$ under the null

Example continue

- In this case the size of Ω is large and we cannot go over all the elements of Ω

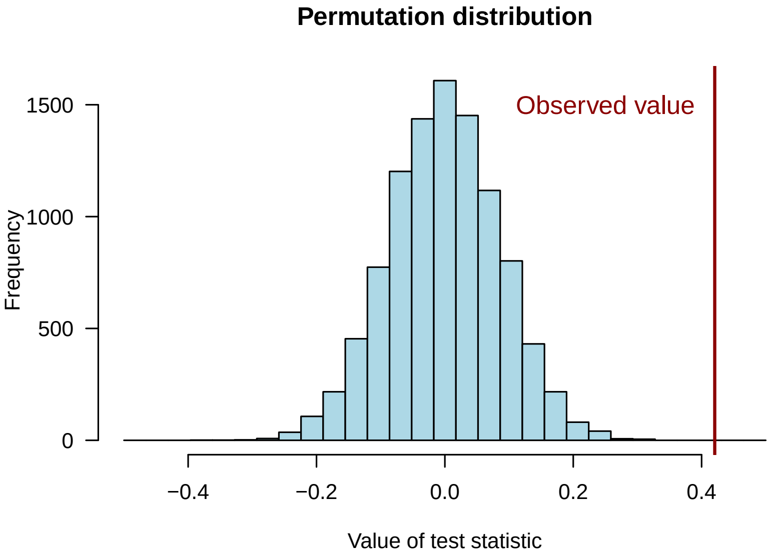
```
> choose(200,100)
[1] 9.054851e+58
```

- We will use Monta-Carlo simulations. The R code is bellow,

```
f.permute = function(){
  id = sample(c(1:length(x)),length(x2))
  t0= rep(0,length(x))
  t0[id]=1
  statistic0 = mean(x[t0==1])-mean(x[t0==0])
  return(statistic0)
}

stat.permutation = replicate(10000,f.permute())
```

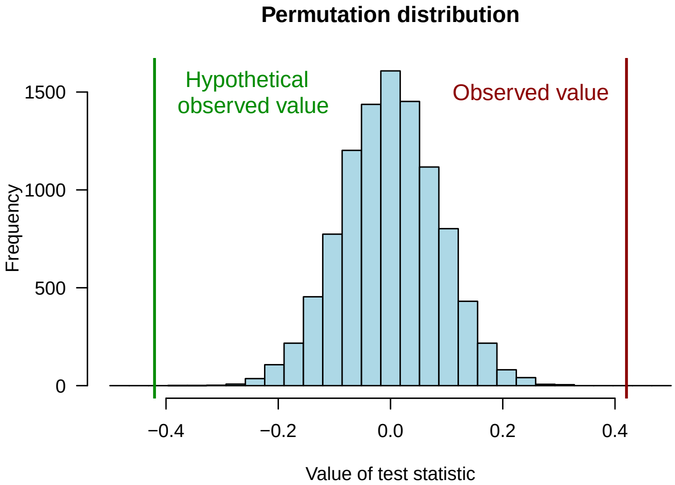
What is B in the code? $B = 10000$

Example continue



Example continue

Should we reject the null if we would have observed the green line?



Example continue

- What is the *P-value*?
- A one sided *P-value* needs to specify if we are looking for extreme values from the right or the left, $Pr_{H_0}(t(Z,r) \geq T)$ or $Pr_{H_0}(t(Z,r) \leq T)$

Solutions:

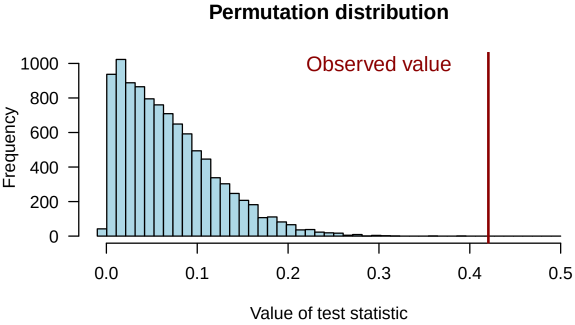
1. Choose the lowest option, and double the *P-value* by 2 (Bonforoni correction):

$$P - value = min\left[Pr_{H_0}(t(Z,r) \geq T), Pr_{H_0}(t(Z,r) \leq T)\right] \times 2$$

Example continue

Solutions:

2. Adjust the test statistic in away which will make us want to reject the null only for extreme values from one side (this is not always possible). In our case we re-define $t(Z, r)$ as, $\left|\frac{Z^T r}{Z^T 1} - \frac{(1-Z)^T r}{(1-Z)^T 1}\right|$, i.e the absolute difference in means.



Wilcoxon rank sum test (WRST)

- The Wilcoxon rank sum test is one of the most commonly used non-parametric tests
- Let $q = (q_1, q_2, \ldots, q_N)$ be the ranks of the responses **r**
- The test statistic is the sum of the ranks of the treated units,

$$t(Z, r) = W = Z^T q = \sum_{i=1}^{N} Z_i q_i$$

  where $q_i = rank(r_i)$
- WRST is usually used to test for differences in medians (and means) between two distributions. It has a lower power detecting differences in the variance (when the medians are similar).

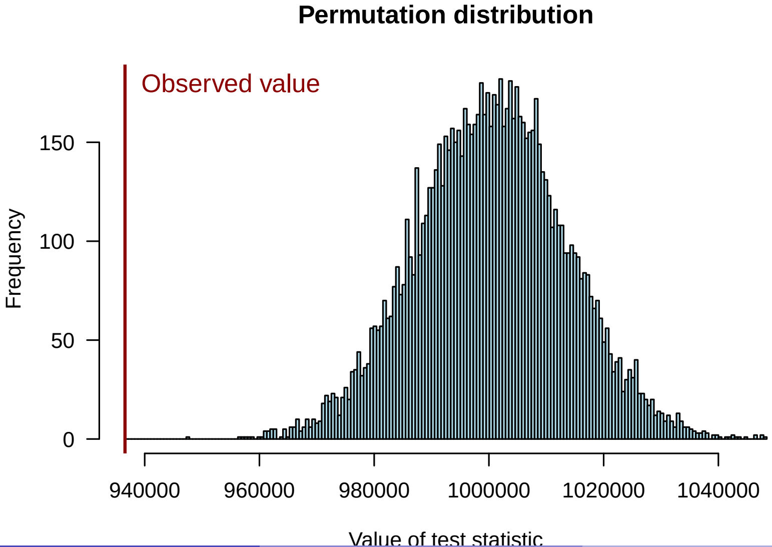
WRST: example

Consider the same data as in the previous example. The code bellow calculates the permutation distribution, and observed statistic for the WRST,

```
q = rank(x)
f.permute = function(){
  id = sample(c(1:length(x)),length(x2))
  t0= rep(0,length(x))
  t0[id]=1
  statistic0 = sum(q[t0==1])
  return(statistic0)
}
stat.permutation = replicate(10000,f.permute())
statistic.obs = sum(q[t==1])
```

WRST: example

The permutation distribution is,



WRST: example

The *P-value* (two-sided hypothesis test) is,

$$P - value = min\left[Pr_{H_0}(t(Z,r) \geq W), Pr_{H_0}(t(Z,r) \leq W)\right] \times 2$$

$$= \frac{1}{10000+1} \times 2 = 0.00019998$$

The implementation in *R*:

```
wilcox.test()
```

Kolmogorov-Smirnov (KS) test

- The KS test is used to detect differences between two distributions, it can detect differences in other moments except the expectations, such as the variance or quantiles
- The hypothesis test to have in mind is,

$$H_0;\ F_x = F_y$$

$$H_0;\ F_x \neq F_y$$

- The KS test statistic is the largest difference between the CDF's of group x and group y,

$$D = max_w \left|F_x(w) - F_y(w)\right|$$

Kolmogorov-Smirnov (KS) test

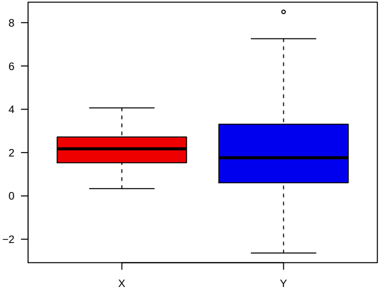
- The CDF is not a known function and needs to be approximated. We use the empirical CDF which is defined as,

$$\widehat{F_x}(w) = \frac{\#\{x \leq w\}}{n_x}$$

- Consider the following example:

```
set.seed(16)
x=rnorm(50,mean=2,sd=1)
y=rnorm(100,mean=2,sd=2)

### The emperical CDF:
Fx = function(w,x){
  return(sum(x<=w)/length(x))
}
```

Kolmogorov-Smirnov: Example

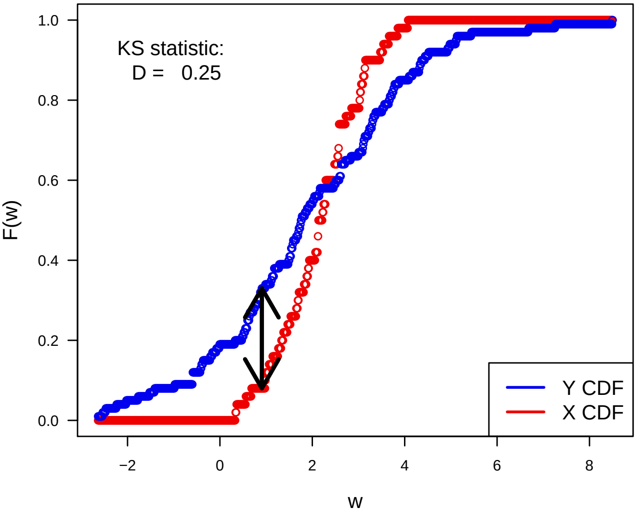


What do you think will be the results using WRST? Will a T-test detect the difference in distributions?
What is the null in a T-test? Are the assumption in a T-test satisfied?

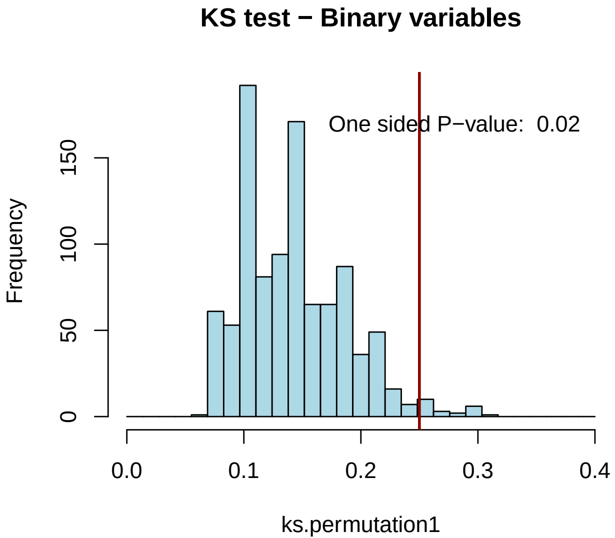
Kolmogorov-Smirnov: Example

```
Welch Two Sample t-test

data:  x and y
t = 0.9403, df = 144.938, p-value = 0.3486
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -0.2573415  0.7244316
sample estimates:
mean of x mean of y
 2.157051  1.923506

Wilcoxon rank sum test with continuity correction

data:  x and y
W = 2706, p-value = 0.4126
alternative hypothesis: true location shift is not equal to 0
```

Kolmogorov-Smirnov: Example

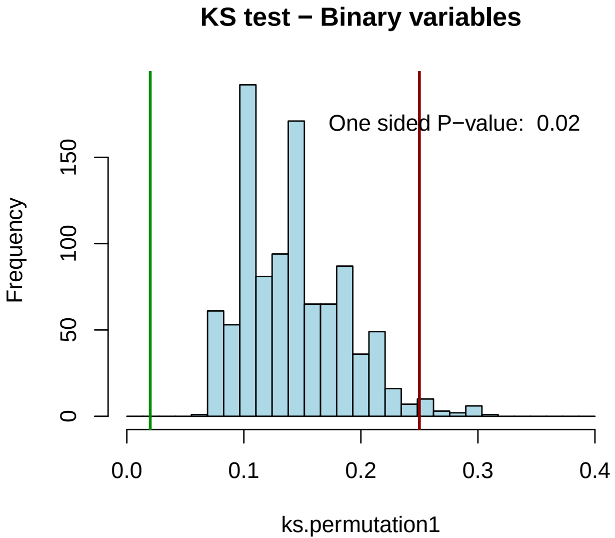


Kolmogorov-Smirnov: Example



Kolmogorov-Smirnov: Example

Should we reject the null if we observe the green line?



Kolmogorov-Smirnov: Binary variables

Mr. Sceptical argued that the KS test has low power when considering binary distributions (Bernoulli), and suggested using the difference in means instead (difference in proportions)
What do you think?

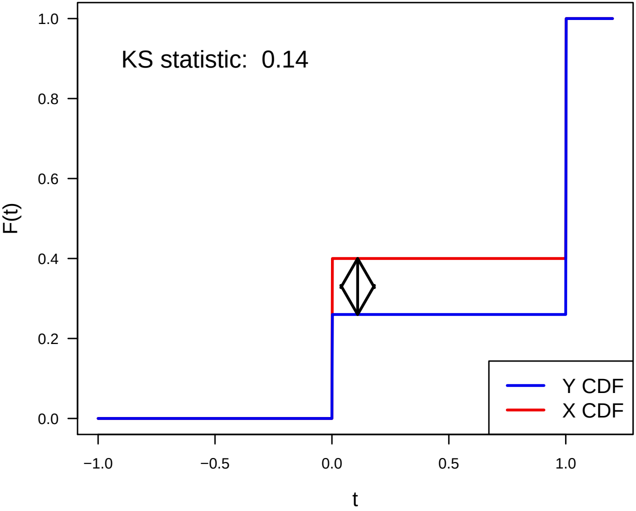
Answer: The tests are the same! When the the two distributions under comparison are both Bernoulli, the null of the KS test is, $H_0: \; P_x = P_y$ and the test statistic becomes.

$$D = \widehat{P_x} - \widehat{P_y}$$

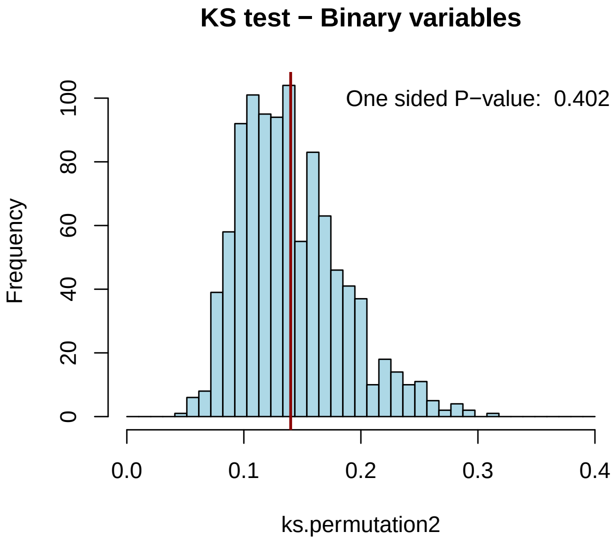
Example:

```
set.seed(14)
x=rbinom(50,size=1,prob=0.5)
y=rbinom(100,size=1,prob=0.7)
```

Kolmogorov-Smirnov: Binary variables



Kolmogorov-Smirnov: Binary variables



Kolmogorov-Smirnov: Binary variables

```
Welch Two Sample t-test

data:  x and y
t = -1.6926, df = 88.688, p-value = 0.09404
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -0.30435636  0.02435636
sample estimates:
mean of x mean of y
     0.60      0.74
```

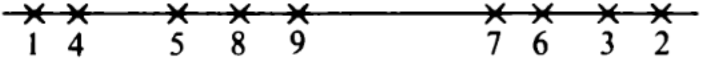
Siegel-Tukey test (ST)

- The ST is used to test for differences in scale between two groups, i.e., equality of variance between two groups.
- There is an implemented in **R** using the package "DescTools", and the command "SiegelTukeyTest()".
- What is the standard test for equality of variance? Levene's test, and it can be implemented using "levene.test()" in **R** , using the package "lawstat".
- Example code:

  ```
  library(DescTools)
  x <- c(12, 13, 29, 30)
  y <- c(15, 17, 18, 24, 25, 26)
  SiegelTukeyTest(x, y)
  ```

- When would we want to test for equality of variance between groups? One example, is to validate a constant treatment effect model.

Siegal-Tukey test statistic



Constructing the test statistic step-by-step:

1. Assign rank 1 to the smallest observation, rank 2 to the highest observation, rank 3 to the second highest observation, rank 4 to the second smallest observation and so on.
2. The ST test statistic is the sum of the ranks in treated group,

$$t(Z, r) = \sum_{i=1}^{N} s_i$$

where $s_1, \dots s_n$ are the ranks of the observations using the ranking scheme above.

Siegal-Tukey test statistic

- What are possible problems with the construction of this test statistic? We can construct a symmetric test statistic by starting to rank from the highest observation. In general this will not matter, but in some cases it can lead to different conclusions.
- A possible solution is to calculate the test statistic as the average of starting from both directions.
- When using the Siegal-Tukey test statistic what null hypothesis are we testing? The sharp null of no treatment effect! This is always the null hypothesis that we test using permutation inference.
- Homework assignment: Write **R** code to calculate the Siegal-Tukey test statistic.

Example:

- Next we compare Siegal-Tukey to Levene's test in terms of power.
- The DGP is,

```
rm(list=ls())
set.seed(12345)
library(DescTools)
library(lawstat)
n=200
x = rnorm(n)
y= rnorm(n,sd=0.8)
outcome = c(x,y)
tr = c(rep(1,n),rep(0,n))
```

- Using Levene's test is the inference based on permutation inference or asymptotic theory? Asymptotic theory. Suggest a way to use Levene's test and make inference using permutation inference.

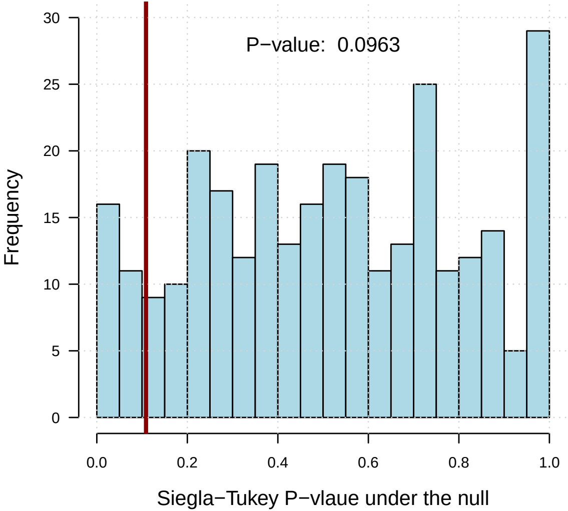
Example:

```
> SiegelTukeyTest(outcome[tr==1], outcome[tr==0])

Siegel-Tukey-test for equal variability

data:  outcome[tr == 1] and outcome[tr == 0]
ST = 2228, p-value = 0.1087
alternative hypothesis: true ratio of scales is not equal to 1

> levene.test(outcome,factor(tr))

modified robust Brown-Forsythe Levene-type test based on the absolu

data:  outcome
Test Statistic = 24.963, p-value = 8.767e-07
```

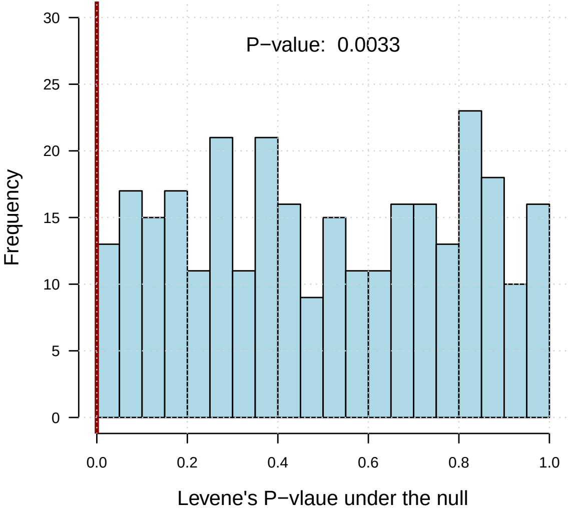
Levene’s test using permutation inference

```
### Levene’s test with permutation inference:
S=300
st <- levene <- rep(NA,S)
for (s in c(1:S)){
  tr0 = sample(tr,length(tr),replace=FALSE)
  st[s] <- SiegelTukeyTest(outcome[tr0==1], outcome[tr0==0])$p.value
  levene[s] <- levene.test(outcome,factor(tr0))$p.value
}

st.obs <- SiegelTukeyTest(outcome[tr==1], outcome[tr==0])$p.value
levene.obs <- levene.test(outcome,factor(tr))$p.value
```

Siegal-Tukey



Levene's



P−value: 0.0033
Frequency
30
25
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0
Levene's P−vlaue under the null

Examples of permutation inference

- Is there an association between paling against each other in the World-Cup and military conflict?
- The paper in the link bellow tries to answer this question using permutation inference. This is a nice and simple example of permutation inference is,
  `http://www.andrewbertoli.org/wp-content/uploads/2013/02/Direct-Sports-Competition.pdf`