Statistical Data Analysis, Lecture 8

dr. Dennis Dobler

Vrije Universiteit Amsterdam

8 April 2020

Topics in this course

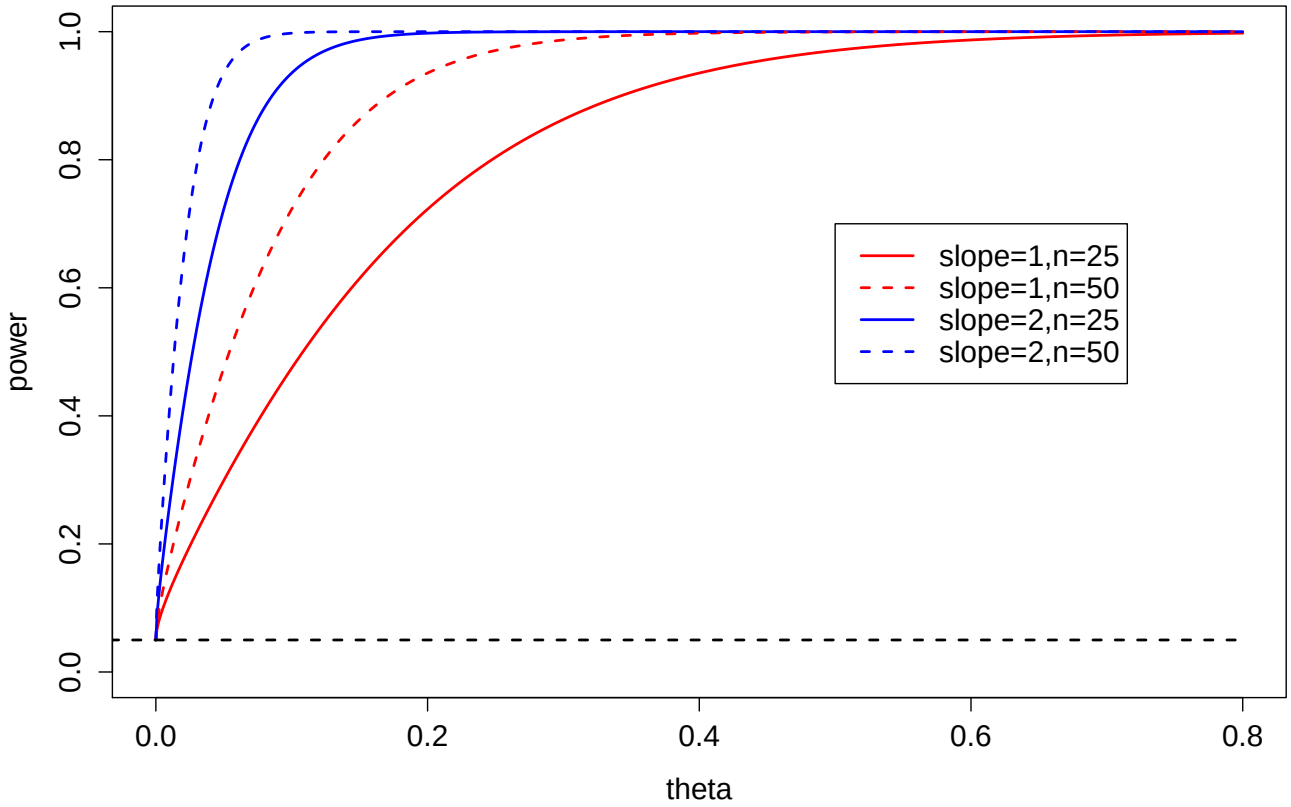
1. Summarizing data
2. Exploring distributions
3. Density estimation
4. Bootstrap methods
5. Nonparametric tests
6. Analysis of categorical data
7. Multiple linear regression

Chapter 6: Nonparametric methods

Contents of Chapter 6:

asymptotic efficiency

Asymptotic efficiency and slope of power of tests



$$\pi_n(\theta) \approx 1 - \Phi(\xi_{1-\alpha} - \sqrt{n}\frac{\mu'(0)}{\sigma(0)}\theta)$$

Difference in slope is compensated by difference in sample size.

Values of Asymptotic Relative Efficiency

	t	s	w
t	1		
s	$\frac{2}{\pi}$	1	
w	$\frac{3}{\pi}$	$\frac{3}{2}$	1

N(0,1)

	t	s	w
t	1		
s	$\frac{\pi^2}{12}$	1	
w	$\frac{\pi^2}{9}$	$\frac{4}{3}$	1

logistic

	t	s	w
t	1		
s	$\frac{1}{3}$	1	
w	1	3	1

uniform

	t	s	w
t	1		
s	2	1	
w	$\frac{3}{2}$	$\frac{3}{4}$	1

Laplace

Table 6.1: Asymptotic relative efficiencies (row-variable with respect to column-variable) of the t-test (t), sign test (s) and Wilcoxon signed rank test (w) for shift alternatives of different F_0.

For example, for a Gaussian sample the a.r.e. of the sign test to t-test is $2/\pi$. That means that the sample size ratio should be:

$$\frac{n_{t-test}}{n_{signtest}} = \frac{2}{\pi} = 0.64$$

to obtain similar power in both tests.

Asymptotic efficiency

Perform the *t*-test, the signed rank test and the sign test B times on a sample of size n from $N(\texttt{theta}, 1)$, and count the fraction of rejected tests (=power) at level 5%.

```
aresimulation=function(B,n,theta)
{
   pvalttest=numeric(B)
   pvalsigntest=numeric(B)
   pvalwilctest=numeric(B)
   for(i in 1:B)
   {
      x=rnorm(n,mean=theta,sd=1)
      pvalttest[i]=t.test(x)[[3]]
      pvalsigntest[i]=binom.test(sum(x>0),size,p=0.5)[[3]]
      pvalwilctest[i]=wilcox.test(x)[[3]]
   }
   powert=sum(pvalttest<0.05)/B
   powersign=sum(pvalsigntest<0.05)/B
   powerwilc=sum(pvalwilctest<0.05)/B
   rbind(c("t","wilc","sign"),c(powert,powerwilc,powersign))
}
```

Asymptotic efficiency

Perform the t-test, the signed rank test, and the sign test 1000 times on a sample of size 100 from $N(0,1)$, $N(0.1,1)$ and $N(0.2,1)$, and count the fraction of rejected tests (=power) at level 5%.

```
> aresimulation(1000,100,0)
     [,1]   [,2]    [,3]
[1,] "t"    "wilc"  "sign"
[2,] "0.06" "0.059" "0.046"
> aresimulation(1000,100,0.1)
     [,1]    [,2]   [,3]
[1,] "t"     "wilc" "sign"
[2,] "0.177" "0.16" "0.094"
> aresimulation(1000,100,0.2)
     [,1]    [,2]    [,3]
[1,] "t"     "wilc"  "sign"
[2,] "0.498" "0.478" "0.301"
```

```
> aresimulation(1000,100,0.2)[,1]
[1] "t"     "0.509"
> aresimulation(1000,round(100*pi/3),0.2)[,2]
[1] "wilc"  "0.504"
> aresimulation(1000,round(100*pi/2),0.2)[,3]
[1] "sign"  "0.486"
```

Apart from some variation adjusting sample sizes yields equal power. The results only hold asymptotically (i.e. `n` should be large).

Asymptotic efficiency

The asymptotic efficiency of T relative to $\tilde{T}$ is

$$\frac{m}{n} = \left(\frac{\mu'(0)/\sigma(0)}{\tilde{\mu}'(0)/\tilde{\sigma}(0)} \right)^2 = \text{are}(T_n, \tilde{T}_m).$$

The a.r.e. of the t-test relative to the Wilcoxon signed rank test is $\pi/3 = 1.05$ (if the data is normal!), thus only slightly larger than 1. Hence, the signed rank test is a powerful alternative to the t-test:

- it is only slightly less efficient for normally distributed data
- it has much higher power and adequate level in case the data do not come from a normal (but still symmetric!) distribution.

median test

Two samples

Two samples can either be paired or independent.

Paired samples: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

Independent samples: $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$.

For paired samples: if interested whether one sample is stochastically larger, consider differences $Z_i = Y_i - X_i$ and apply methods of last week.

Today we discuss tests for independent samples. Next week we will discuss tests for correlation for paired samples.

Median test

Assumption $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} F$, $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} G$, F and G continuous.

Test median test

Hypothesis $H_0 : F = G$

Test statistic
$Z = \#(i = 1, \ldots, m : \ X_i \leq \text{med}(X_1, \ldots, X_m, Y_1, \ldots, Y_n))$.

Distribution Under H_0 we have $Z \sim \text{hypergeom}(m+n, m, p)$ with $p = \lfloor \frac{m+n+1}{2} \rfloor$ draws (without replacement). That is,

$$P_{H_0}(Z = z) = \frac{\binom{m}{z}\binom{n}{p-z}}{\binom{m+n}{p}}$$

This is a nonparametric test.

Example median test (1)

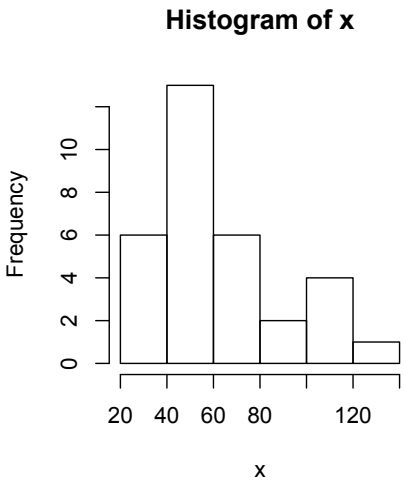
Example We have measured thromboglobulin data of Raynaud patients without organ defects (x) and of patients with other (CTRP) auto-immune disease (y). This concerns two independent samples.

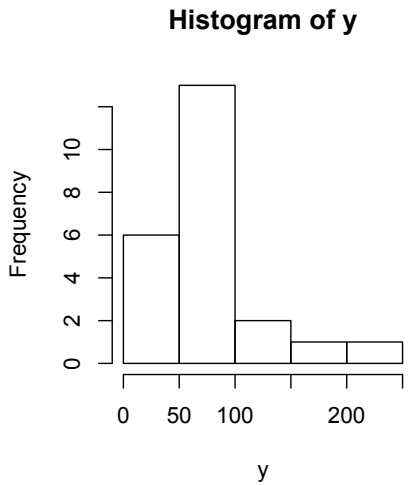
```
> x
 [1]  22.0  25.0  27.0  30.5  32.5  34.0  41.0  41.0  43.5  43.5  44.5
[12]  44.5  44.5  48.5  50.5  53.0  55.5  58.5  58.5  63.5  68.5  68.5
[23]  68.5  73.5  80.0  89.5  92.0 101.5 104.0 119.0 119.0 124.5
> y
 [1]  20.0  23.5  27.0  32.0  41.0  44.0  51.0  53.0  55.5  58.5  62.5
[12]  62.5  65.0  67.0  69.5  72.0  80.0  88.5  91.0 138.0 146.5 160.5
[23] 219.0
```

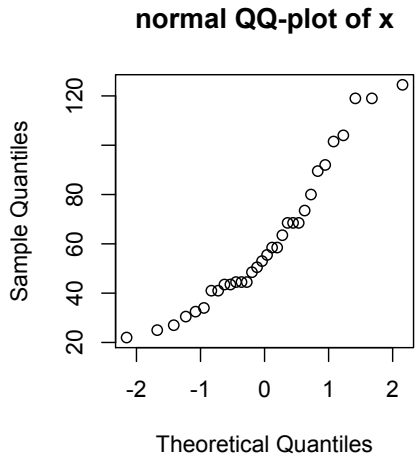
The question is: is sample x stochastically smaller than sample y, i.e. $F_X(u) > F_Y(u)$ for every u? This is what we would like to confirm in the alternative hypothesis.

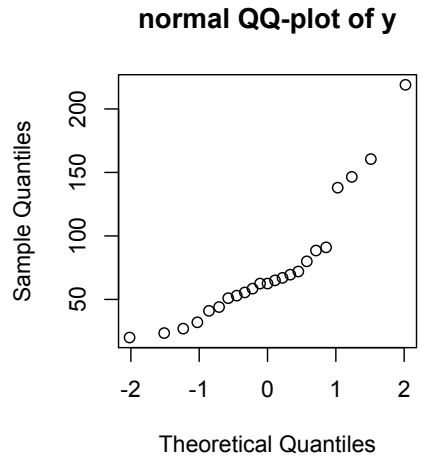
This is a one-sided hypothesis. (We will use $\alpha = 5\%$.)

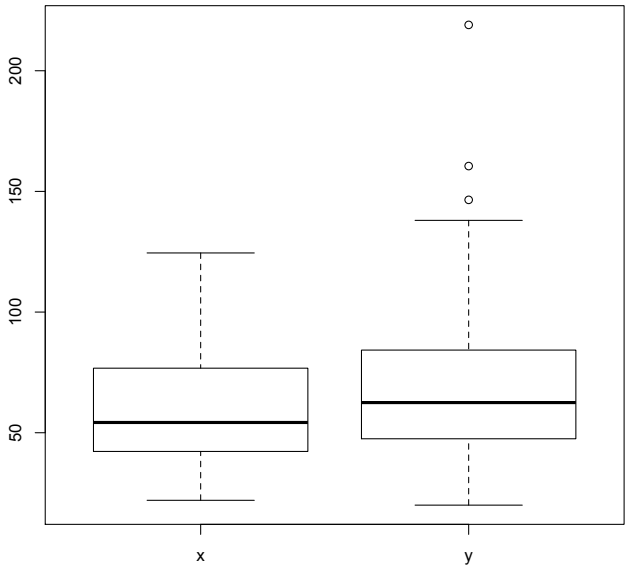
Example median test (2)

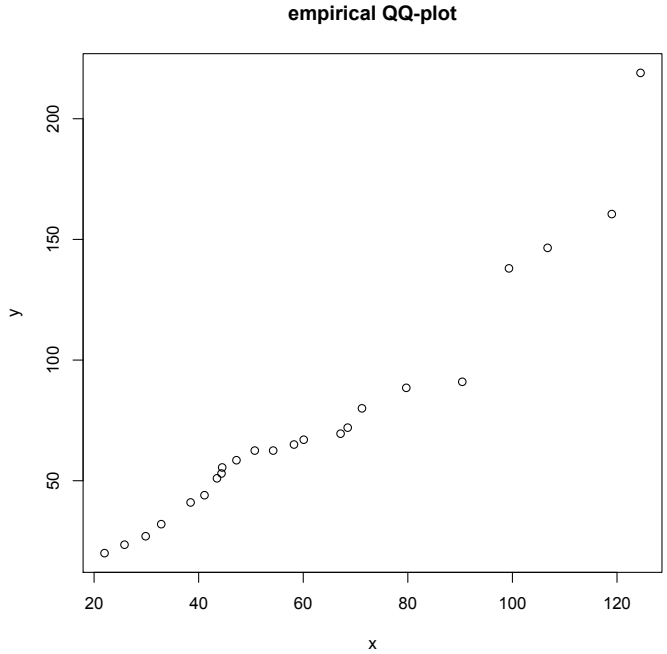












Example median test (3)

The median test is not standard in R: you have to program it yourself, using `phyper`.

```
> data=c(x,y)
> t=sum(x<=median(data))
> t
[1] 19
> nx=length(x)
> ny=length(y)
> 1-phyper(t-1,nx,ny,floor((nx+ny+1)/2))
[1] 0.1134224
```

We only look at the right p-value, since under H_1 we expect Z to be large, that is, we expect more X_i's smaller than the median of the combined sample.

Conclusion H_0 is not rejected at level $\alpha = 5\%$, we cannot say with statistical certainty that x is stochastically smaller than y.

Wilcoxon two sample test

Wilcoxon two sample test (1)

Assumption $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} F$ and $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} G$

Test Wilcoxon two sample test

Hypothesis $H_0 : F = G$

Test statistic $W = \sum_{i=1}^{n} R_i$, where R_i is the rank of Y_i in the combined sample.

Distribution Under H_0 we have for the ordered ranks $R_{(i)}$

$$P_{H_0}((R_{(1)}, \ldots, R_{(n)}) = (r_1, \ldots, r_n)) = \frac{1}{\binom{n+m}{n}},$$

for every subset $r_1 < r_2 < \cdots < r_n$ of $\{1, 2, \ldots, n+m\}$. The distribution of the test statistic is given by

$$P_{H_0}(W = w) = \frac{\#(\text{arrangements } r_1 < r_2 < \cdots < r_n \text{ with } \sum_{i=1}^{n} r_i = w)}{\binom{n+m}{n}}.$$

Wilcoxon two sample test (2)

This is a nonparametric test.

Equivalent test statistics for this test are:

$$
\begin{aligned}
W &= \sum_{i=1}^{n} R_i \\
U &= \sum_{i=1}^{m}\sum_{j=1}^{n} 1_{\{X_i<Y_j\}} = W - \frac{1}{2}n(n+1) \\
\tilde{W} &= \sum_{i=1}^{m} S_i,
\end{aligned}
$$

where S_i is the rank of X_i in the combined sample,

$$
\tilde{U} = \tilde{W} - \frac{1}{2}m(m+1)
$$

This one is used in R

Wilcoxon two sample test (3)

Each of the test statistics is asymptotically normally distributed (under H_0, after appropriate centering and scaling).

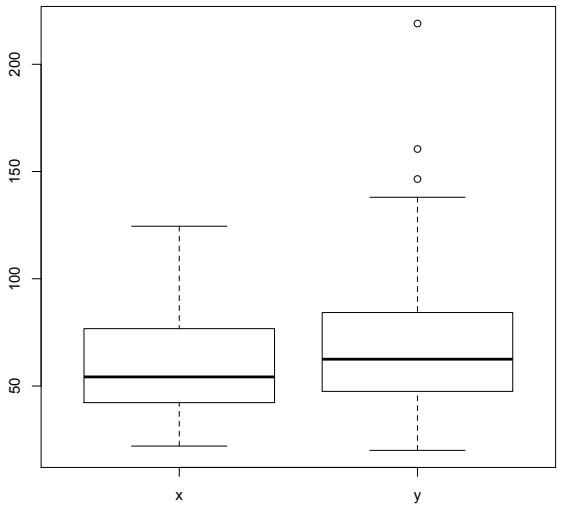
In case of ties (groups of identical measurements) the test is based on pseudoranks and the test is performed conditionally on the pattern of ties. In such case R uses a normal approximation for the distribution of the test statistic. This test is still nonparametric.

Other names for this test: Mann-Whitney test, Wilcoxon rank sum test.

The thromboglobulin data

```
> x
 [1]  22.0  25.0  27.0  30.5  32.5  34.0  41.0  41.0  43.5  43.5  44.5
[12]  44.5  44.5  48.5  50.5  53.0  55.5  58.5  58.5  63.5  68.5  68.5
[23]  68.5  73.5  80.0  89.5  92.0 101.5 104.0 119.0 119.0 124.5
> y
 [1]  20.0  23.5  27.0  32.0  41.0  44.0  51.0  53.0  55.5  58.5  62.5
[12]  62.5  65.0  67.0  69.5  72.0  80.0  88.5  91.0 138.0 146.5 160.5
[23] 219.0
```

H_1: distribution of x is shifted to the left of distribution of y



Example Wilcoxon two sample test

The Wilcoxon two sample test applied to the thromboglobulin data yields:

```
> wilcox.test(x,y,alternative="less")

Wilcoxon rank sum test with continuity correction

data:  x and y
W = 319, p-value = 0.2039
alternative hypothesis: true location shift is less than 0

Warning message:
In wilcox.test.default(x, y, alternative = "less") :
  cannot compute exact p-value with ties
```

There is a warning because of the ties: R uses a normal approximation in the computation of the p-value.

Remark We use `alternative="less"` because R uses as test statistic $\tilde{U}$ (so ranks of the X_i's) which is smaller under H_1.

Kolmogorov Smirnov test

Kolmogorov Smirnov two sample test (1)

Assumption $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} F$ and $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} G$, F and G continuous.

Test Kolmogorov Smirnov two sample test

Hypothesis $H_0 : F = G$

Test statistic $D = \sup_{-\infty<x<\infty} |\hat{F}_m(x) - \hat{G}_n(x)|$ equivalent to $D = \max_{1\leq i\leq n} \max\left\{ |\frac{1}{m}(R_{(i)} - i) - \frac{i}{n}|, |\frac{1}{m}(R_{(i)} - i) - \frac{i-1}{n}| \right\}$ with $R_{(1)}, \ldots, R_{(n)}$ the ordered ranks of $Y_1, \ldots, Y_n$ in the combined sample (see syllabus).

Distribution The distribution of D under H_0 depends on m and n, but is independent of F and G. It can be approximated by a normal distribution.

This is a nonparametric test.

Example Kolmogorov Smirnov test

The Kolmogorov Smirnov test applied to the thromboglobulin data yields:

```
> ks.test(x,y,alternative="greater")

Two-sample Kolmogorov-Smirnov test

data:  x and y
D^+ = 0.2079, p-value = 0.3146
alternative hypothesis: the CDF of x lies above that of y

Warning message:
In ks.test(x, y, alternative = "greater") :
  cannot compute exact p-values with ties
```

There is a warning because of the ties: *R* uses a normal approximation for the *p*-value. In case of large ties this test is unreliable.

Remark We use `alternative="greater"` such that *R* uses `"the CDF of x lies above that of y"` as alternative hypothesis.

permutation test

Permutation test (general procedure!)

Assumption $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} F$ and $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} G$

Test permutation test for 2 independent samples

Hypothesis $H_0 : F = G$

Test statistic Some sensible test statistic T that expresses differences between the two underlying distributions

Distribution The right p-value is $P_{H_0}\left(T \geq t \middle| Z_{(1)}, \ldots, Z_{(m+n)}\right)$

$$= \frac{\#(\text{permutations } \pi \text{ with } T(Z_{(\pi_1)}, \ldots, Z_{(\pi_m)}, Z_{(\pi_{m+1})}, \ldots, Z_{(\pi_{m+n})}) \geq t)}{(m+n)!}$$

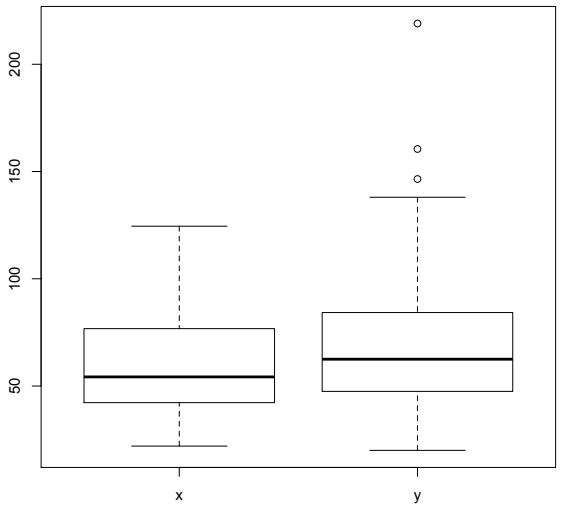
where $Z_{(1)}, \ldots, Z_{(m+n)}$ is the ordered combined sample. The left p-value is computed likewise. The p-value is derived under H_0 conditionally on all observations!

This is a nonparametric test.

The thromboglobulin data

```
> x
 [1]  22.0  25.0  27.0  30.5  32.5  34.0  41.0  41.0  43.5  43.5  44.5
[12]  44.5  44.5  48.5  50.5  53.0  55.5  58.5  58.5  63.5  68.5  68.5
[23]  68.5  73.5  80.0  89.5  92.0 101.5 104.0 119.0 119.0 124.5
> y
 [1]  20.0  23.5  27.0  32.0  41.0  44.0  51.0  53.0  55.5  58.5  62.5
[12]  62.5  65.0  67.0  69.5  72.0  80.0  88.5  91.0 138.0 146.5 160.5
[23] 219.0
```

H_1: distribution of x is shifted to the left from distribution of y



Example permutation test for thromboglobulin samples

A permutation test applied to the thromboglobulin data using $T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = \text{med}(X) - \text{med}(Y)$ yields:

```
> data=c(x,y)
> nx=length(x)
> ny=length(y)
> myteststatistic=function(z,m,n) {median(z[1:m])-median(z[(m+1):(m+n)])}
> B=1000
> permutationtval=numeric(B)
> for (i in 1:B)  permutationtval[i]=myteststatistic(sample(data),nx,ny)
> t=myteststatistic(data,nx,ny)
> sum(permutationtval<=t)/B
[1] 0.182
```

Remark We only look at the left p-value because under H_1 we expect a small $\text{med}(X) - \text{med}(Y)$.

Remark We have made a bootstrap test approximation here, considering only $B = 1000$ permutations instead of all $55! = 1.26964 \cdot 10^{73}$ possible permutations of data.

Permutation test (general procedure!)

Assumption $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} F$ and $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} G$

Test permutation test for 2 independent samples

Hypothesis $H_0 : F = G$

Warning: If only the location parameters in both samples are the same but $F \neq G$, the permutation test does in general not keep the nominal level α anymore!

E.g., for $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} N(0, \sigma_X^2)$ and $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma_Y^2)$ one should not use the test statistic $\bar{X}_m - \bar{Y}_n$ in combination with the permutation approach if $\sigma_X \neq \sigma_Y$!
Better choice: permutation version of the two-sample t-test that assumes unequal variances. (Exact only for large samples.)

Overview of the 2 sample tests

t-test assumes normality, only for shift alternatives

median test nonparametric, especially suited (i.e. efficient) for shift alternatives, not very powerful

Wilcoxon two sample test nonparametric, especially suited (i.e. efficient) for shift alternatives, uses more information from the data

Kolmogorov Smirnov test nonparametric, suited for all alternatives, uses a lot of information from the data. Careful in case of large ties.

permutation test nonparametric, power depends on the test statistic chosen. Careful in case of no exchangeability.

to finish

To wrap up

Today we discussed

1. One sample problems
2. Asymptotic efficiency
3. Two sample problems
4. Tests for correlation

Next week tests for correlation and categorical data