Statistical Data Analysis, Lecture 7

dr. Dennis Dobler

Vrije Universiteit Amsterdam

1 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

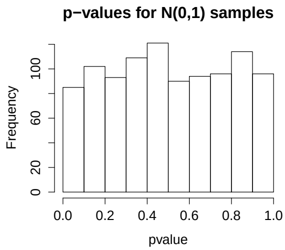
Contents of Chapter 6:

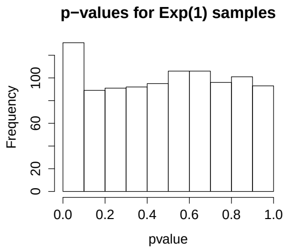
1. One sample problems
   - sign test
   - signed rank test
2. Asymptotic efficiency
3. Two sample problems
4. Tests for correlation

Example

Suppose we apply the t-test to exponentially distributed data.....

```
> n=20
> pvalexp=numeric(1000)
> pvalnorm=numeric(1000)
> for (i in 1:1000){
+   x=rnorm(n,mean=1)
+   y=rexp(n)
+   pvalnorm[i]=t.test(x,mu=1)[[3]]
+   pvalexp[i]=t.test(y,mu=1)[[3]]
+ }
> hist(pvalnorm,main="...")
> hist(pvalexp,main="...")
> sum(pvalnorm<0.05)/m
[1] 0.045
> sum(pvalexp<0.05)/m
[1] 0.09
```





Actual level is 0.09 instead of 0.05.

Idea

Nonparametric tests make no (parametric) assumptions about the underlying distribution F of the data.
E.g. no normality assumption.

These tests are applicable for broad classes of distributions, and have actual level α. The distribution of the test statistic under H_0 is the same for each distribution F belonging to H_0.

Nonparametric tests are robust with respect to the level:
they have the intended level α for a large class of distributions.

Nonparametric tests are more efficient (have higher power) than parametric tests when the (normality) assumptions are not fulfilled.

Test on location

Assume we have a sample $X_1, \ldots, X_n$ from an unknown distribution F, and we want to test the location of F.

Which test would you use?

t-test

Assumption $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$

Test t-test

Null hypothesis $H_0 : \mu = \mu_0$

Test statistic $T = \sqrt{n}\frac{\overline{X} - \mu_0}{S_X}$.

Distribution Under H_0, we have $T \sim t_{n-1}$.

This is a parametric test (assumes normality) for a composite H_0, consisting of all normal distributions with expectation μ_0.

sign test

Sign test

Assumption Underlying distribution F has a unique median m, such that $\mathrm{P}(X_i < m) = \mathrm{P}(X_i > m) = \frac{1}{2}$.

Test sign test

Hypothesis $H_0 : m = m_0$. This is a composite null hypothesis. Which class of distributions?

Test statistic $T = \sum_{i=1}^{n} 1_{X_i > m_0}$.

Distribution Under H_0, we have $T \sim \mathrm{bin}(n, \frac{1}{2})$. This is a nonparametric test, since T has this distribution for all F in H_0.

In case k of the X_i's are equal to m_0, delete these k values and perform the test conditionally on k values equal to m_0, and $T \sim \mathrm{bin}(n-k, \frac{1}{2})$ under H_0.

Example sign test (1)

Example We have measured the grades of 13 students in order to test the difficulty of an exam. We want to test whether the location is smaller than 6 versus the alternative that the exam is too easy.

```
> grades <- c(3.7,5.2,6.9,7.2,6.4,9.3,4.3,8.4,6.5,8.1,7.3,6.1,5.8)
```

Estimate the location

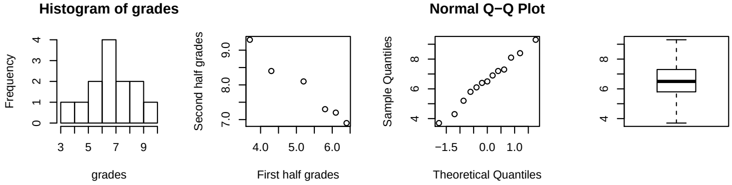
```
> mean(grades)
[1] 6.553846
> median(grades)
[1] 6.5
```

Based on these numbers the exam looks alright, but we need a test.

Example sign test (2)

First make some graphics to judge which test is applicable (look for symmetry, normality, outliers, etc.)

```
> par(mfrow=c(1,4))
> hist(grades)
> symplot(grades)
> qqnorm(grades)
> boxplot(grades)
```



These plots show that F could very well be symmetric or even normal, but the sample size is small, so no strong conclusions!

Example sign test (3)

Perform the sign test for $H_0 : m \leq 6$ at significance level $\alpha = 5\%$.

```
> length(grades)
[1] 13
> sum(grades>6)
[1] 9
> sum(grades==6)
[1] 0
> binom.test(9,13,alt="g")

        Exact binomial test

data:  9 and 13
number of successes = 9, number of trials = 13, p-value = 0.1334
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.4273807 1.0000000
sample estimates:
probability of success
             0.6923077
```

Conclusion?

Example sign test (3)

Compare this result to the t-test for $H_0 : \mu \leq 6$ at significance level $\alpha = 5\%$:

```
> t.test(grades,mu=6,alt="g")

        One Sample t-test

data:  grades
t = 1.2569, df = 12, p-value = 0.1164
alternative hypothesis: true mean is greater than 6
95 percent confidence interval:
 5.768463      Inf
sample estimates:
mean of x
 6.553846
```

Conclusion?

(Wilcoxon) signed rank test

Signed rank test (1)

Assumption Underlying distribution F is continuous and symmetric around m.

Test signed rank test

Hypothesis $H_0 : m = m_0$. This is a composite null hypothesis. Which class of distributions?

Test statistic V is based on the ranks R_i of the absolute differences $|X_i - m_0|$. $V = \sum_{i=1}^{n} R_i \operatorname{sgn}(X_i - m_0)$.

Distribution Relatively large values of V indicate that m is larger than m_0. Under H_0, V is distributed as $\sum_{i=1}^{n} Q_i \tilde{R}_i$ with

- Q_i random variable, $\mathrm{P}(Q_i = -1) = \mathrm{P}(Q_i = 1) = \frac{1}{2}$.
- $(\tilde{R}_1, \ldots, \tilde{R}_n)$ a random permutation of $\{1, \ldots, n\}$.

Since this distribution is the same for all distributions under H_0, this is a nonparametric test.

Signed rank test (2)

Distribution $V = \sum_{i=1}^{n} R_i \operatorname{sgn}(X_i - m_0)$ is distributed as $\sum_{i=1}^{n} Q_i \tilde{R}_i$. This follows from Theorem 6.1 in the syllabus:

Let $Z_1, \dots, Z_n$ be independent random variables, with a distribution that is symmetric around 0 and with a continuous distribution function. Let $(R_1, \dots, R_n)$ be the vector of ranks of $|Z_1|, \dots, |Z_n|$ in the corresponding vector of order statistics $(|Z|_{(1)}, \dots, |Z|_{(n)})$. Then the following three properties hold.

- The vectors $(R_1, \dots, R_n)$ and $(\operatorname{sgn}(Z_1), \dots, \operatorname{sgn}(Z_n))$ are independent.
- $P(R_1 = r_1, \dots, R_n = r_n) = 1/n!$ for every permutation $(r_1, \dots, r_n)$ of $\{1, 2, \dots, n\}$.
- The variables $\operatorname{sgn}(Z_1), \dots, \operatorname{sgn}(Z_n)$ are independent and identically distributed with $P(\operatorname{sgn}(Z_i) = -1) = P(\operatorname{sgn}(Z_i) = 1) = \frac{1}{2}$.

For large n a normal approximation can be made for the distribution of V.

In R: wilcox.test which uses the equivalent test statistic $V_+ = \sum_{i:X_i > m_0} R_i$.

Example signed rank test

Let m be the point of symmetry of the underlying distribution of the grades. We apply the Wilcoxon signed rank test to test $H_0 : m \leq 6$ at significance level $\alpha = 5\%$.

```
> wilcox.test(grades,mu=6,alt="g")

Wilcoxon signed rank test

data:  grades
V = 64, p-value = 0.1082
alternative hypothesis: true location is greater than 6
```

Conclusion?

Which test to choose?

We have performed three tests:

- t-test: $p = 0.12$
- signed rank test: $p = 0.11$
- sign test: $p = 0.13$

Which one is best?

Based on the QQ-plot, normality is plausible, as is symmetry. The sample size is rather small to be 'sure' about normality, hence the best option here is probably the Wilcoxon signed rank test.

confidence intervals

Confidence intervals for the location (1)

Using the relationship between statistical tests and confidence intervals we can determine 95% confidence intervals for the location m based on the sign test, the signed rank test and the t-test.

```
> t.test(grades,mu=6)
        .....
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 5.593730 7.513962

> wilcox.test(grades,mu=6,conf.int=T)
        .....
alternative hypothesis: true location is not equal to 6
95 percent confidence interval:
 5.50 7.55
```

For the sign test one has to do this manually: check for which values m_0 the hypothesis $H_0 : m = m_0$ versus $H_1 : m \neq m_0$ is not rejected at level α. Those values together form a $(1 - \alpha)$-confidence interval.

Confidence intervals for the location (2)

H_0 is rejected when $P_{H_0}(T \leq t) \leq \frac{\alpha}{2}$ or $P_{H_0}(T \geq t) \leq \frac{\alpha}{2}$.
Here t is the observed value and $T \sim \mathrm{bin}(n, \frac{1}{2})$ under H_0.

```
> rbind(0:13,round(pbinom(0:13,size=13,p=0.5),3))
[1,]    0 1.000 2.000 3.000 4.000 5.000 ...    13 ## t
[2,]    0 0.002 0.011 0.046 0.133 0.291 ...     1 ## p-left = P(T<=t)
> rbind(0:13,round(1-pbinom((0:13)-1,size=13,p=0.5),3))
[1,]    0 ... 8.000  9.000  10.000  11.000  12.000   13  ## t
[2,]    1 ... 0.291  0.133   0.046   0.011   0.002    0  ## p-right = P(T>=t)
```

Note: only the small p-values are relevant for rejecting H_0 (see the top of this slide).

If $T = \#(X_i > m_0) \in \{3, 4, \ldots, 9, 10\}$ H_0 is not rejected.

```
> sort(grades)
[1]  3.7  4.3  5.2  5.8  6.1  6.4  6.5  6.9  7.2   7.3   8.1   8.4   9.3
```

If $T = 3$, then $8.1, 8.4, 9.3$ exceed m_0, so $m_0 < 8.1$.
If $T = 10$, then $5.8, 6.1, \ldots$ exceed m_0, but 5.2 does not (we could not reject $T = 11$ then). Hence $5.2 < m_0$.

Confidence intervals for the location (3)

Are 5.2 and 8.1 in the confidence interval?

So far we tested possible values of m_0 different than elements of the sample (otherwise X_i's equal to m_0 should be removed).

Check borders separately, in a conditional test with $n = 12$, e.g. for $H_0 : m = 5.2$ and $T = 10$, check $P(T \geq 10) = 1 - P(T \leq 9)$:

```
> 1-pbinom(10-1,12,0.5)
[1] 0.01928711
```

Hence, the value 5.2 is not in the confidence interval. For this case the resulting interval is the open interval $(5.2, 8.1)$.

to finish

To wrap up

Today we discussed

1. One sample problems
   - sign test
   - signed rank test
2. Asymptotic efficiency
3. Two sample problems
4. Tests for correlation

Next week Asymptotic efficiency and tests for two samples