Statistical Data Analysis, Lecture 10

dr. Dennis Dobler

Vrije Universiteit Amsterdam

15 April 2020

Topics in this course

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

Chapter 6: Nonparametric methods

Contents of Chapter 6:

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

Chapter 7: Analysis of categorical data

Contents of Chapter 7:

1. Fisher's exact test
2. Chisquare test
3. Extreme values
4. Bootstrap methods for contingency tables

tests for correlation

Ranks of the two samples

Given a paired sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ we define

- $S_1, \ldots, S_n$ the ranks of $X_1, \ldots, X_n$ in the ordered sample $X_{(1)}, \ldots, X_{(n)}$
- $R_1, \ldots, R_n$ the ranks of $Y_1, \ldots, Y_n$ in the ordered sample $Y_{(1)}, \ldots, Y_{(n)}$

If the two samples are independent, then the ranks $S_1, \ldots, S_n$ are independent of $R_1, \ldots, R_n$.

If the two samples are positively dependent, then the ranks $S_1, \ldots, S_n$ will run approximately in parallel with $R_1, \ldots, R_n$.

If the two samples are negatively dependent, then the ranks $S_1, \ldots, S_n$ will run approximately in opposite order as $R_1, \ldots, R_n$.

Spearman's rank correlation test

Assumption Given a paired sample $(X_1, Y_1), \dots, (X_n, Y_n)$

Test rank correlation test of Spearman

Hypothesis $H_0 : X_i$ and Y_i are independent $i = 1, \dots, n$
vs. $H_1 : X_i$ and Y_i are dependent $i = 1, \dots, n$.

Test statistic $r_s = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\left[\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (S_i - \bar{S})^2\right]^{\frac{1}{2}}}$

Distribution Either exact distribution or a normal approximation

This is a nonparametric test.

Kendall's rank correlation test

Assumption Given a paired sample $(X_1, Y_1), \ldots, (X_n, Y_n)$

Test rank correlation test of Kendall

Hypothesis $H_0 : X_i$ and Y_i are independent $i = 1, \ldots, n$
vs. $H_1 : X_i$ and Y_i are dependent $i = 1, \ldots, n$.

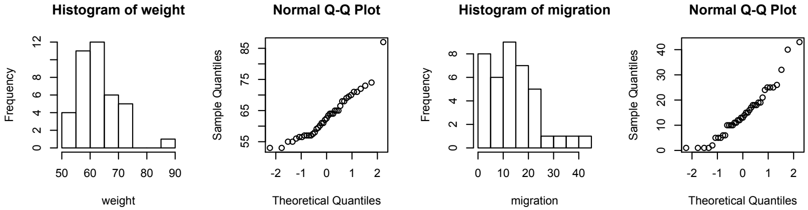
Test statistic $\tau = \frac{\sum\sum_{i \neq j} \mathrm{sgn}(R_i - R_j)\mathrm{sgn}(S_i - S_j)}{n(n-1)} = \frac{4N_\tau}{n(n-1)} - 1$, where the statistic N_τ is equal to the number of pairs (i, j) with $i < j$ for which either $X_i < X_j$ *and* $Y_i < Y_j$, or $X_i > X_j$ *and* $Y_i > Y_j$.

Distribution Either exact distribution or a normal approximation

This is a nonparametric test.

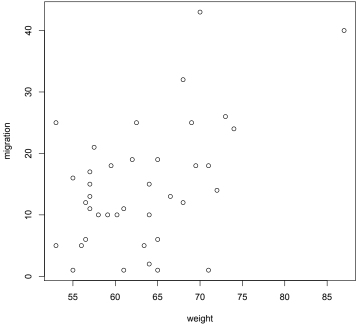
Example rank correlation test (1)

From a sample of 39 Peruvian men that had moved from a native culture to a modern society, the following variables were measured (amongst others): years since migration, systolic and diastolic blood pressure, heart rate, weight, length.



Example rank correlation test (2)

We investigate dependence between years since `migration` and `weight` using Spearman's test.




```
> cor.test(migration,weight,method="s")

Spearman's rank correlation rho

data:  migration and weight
S = 6415.128, p-value = 0.02861
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3506956

Warning message:
In cor.test.default(migration, weight, method = "s") :
  Cannot compute exact p-values with ties
```

R uses a normal approximation for the *p*-value because of ties.

Conclusion?

Example rank correlation test (3)

Kendall's rank correlation test yields:

```
> cor.test(migration,weight,method="k")

Kendall's rank correlation tau

data:  migration and weight
z = 2.1864, p-value = 0.02879
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.2505268

Warning message:
In cor.test.default(migration, weight, method = "k") :
  Cannot compute exact p-value with ties
```

R uses a normal approximation for the *p*-value because of ties.

Conclusion?

Permutation test

Assumption Given a paired sample $(X_1, Y_1), \ldots, (X_n, Y_n)$

Test permutation test for paired samples

Hypothesis $H_0 : X_i$ and Y_i are independent $i = 1, \ldots, n$
vs. $H_1 : X_i$ and Y_i are dependent $i = 1, \ldots, n$.

Test statistic Some test statistic T that expresses dependence between the two samples

Distribution The right p-value is

$$P_{H_0}\left(T \geq t \middle| X_1, \ldots, X_n, Y_{(1)}, \ldots, Y_{(n)}\right)$$
$$= \frac{\#(\text{permutations } \pi \text{ with } T(X_1, \ldots, X_n, Y_{(\pi_1)}, \ldots, Y_{(\pi_n)}) \geq t)}{n!}$$

The left p-value is computed likewise.

This is a nonparametric test.

Example permutation test for paired samples

We can verify the p-value for Kendall's rank correlation test, using a permutation test, in a bootstrap fashion (i.e. considering B permutations, instead of all $n!$ permutations)

```
> B=1000
> t=cor.test(migration,weight,method="k")[[1]]
> permutationtval = numeric(B)
> for(i in 1:B)    ...
> pl= ...;  pr= ...
> p=2*min(pl,pr)
> p
[1] 0.029
```

Compare to p-value output by Kendall's test: $p = 0.02879$.

Remark Smaller values of B yield more variation in the p-value.

Remark Any test statistic that expresses dependence between the two samples can be used in a permutation test for paired samples.

categorical data

Idea (1)

Question Are kind of study and gender independent?
Consider the following data (numbers given are counts):

	exact	arts
men	23	17
women	7	13

Notation

	exact	arts	total
men	N_{11}	N_{12}	$N_{1\cdot}$
women	N_{21}	N_{22}	$N_{2\cdot}$
total	$N_{\cdot 1}$	$N_{\cdot 2}$	$n = N_{\cdot\cdot}$

Idea (2)

Question: Does frequency of nucleotides in DNA depend on its position in the DNA sequence?
Consider the following data of 100 DNA sequences of length 5:

position	1	2	3	4	5	total
A	33	34	19	20	21	127
G	22	27	23	24	21	117
C	31	18	34	30	25	138
T	14	21	24	26	33	118
total	100	100	100	100	100	500

Model I and Fisher's exact test

2 x 2 tables (Model I)

	exact	arts	total
men	23	17	$N_{1\cdot} = 40$
women	7	13	$N_{2\cdot} = 20$
total	$N_{\cdot 1} = 30$	$N_{\cdot 2} = 30$	$n = 60$

2×2 Model We assume the row and column totals are fixed.

Null hypothesis No dependence between row and column variable

Test statistic N_{11} — this number determines the entire table (given the values of $N_{\cdot 1}$, $N_{1\cdot}$, and n).

Distribution Under H_0 we have $N_{11} \sim \text{hypergeom}(n, N_{1\cdot}, N_{\cdot 1})$.

2 x 2 tables (Model I)

	exact	arts	total
men	23	17	$N_{1\cdot} = 40$
women	7	13	$N_{2\cdot} = 20$
total	$N_{\cdot 1} = 30$	$N_{\cdot 2} = 30$	$n = 60$

Null hypothesis No dependence between row and column variable

Test statistic N_{11} — this number determines the entire table

What if we want directed alternative? For instance:

Alternative hypothesis Studying arts is more common among women than men.

Fisher's exact test – directed alternative

If necessary, rephrase the alternative in terms of categories corresponding to N_{11}:

Alternative hypothesis Studying exact sciences is more common among men than women.

```
> study
     [,1] [,2]
[1,]   23    7
[2,]   17   13
> fisher.test(study,alternative="greater")

Fisher's Exact Test for Count Data

data:  study
p-value = 0.08511
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.8634461       Inf
sample estimates:
odds ratio
   2.47347
```

Result 2×2 Model

	exact	arts	total
men	23	17	40
women	7	13	20
total	30	30	60

Under H_0 we expect N_{11} to be $30 \times 40/60 = 20$. We have found a value of 23.

Under H_0 $N_{11} \sim \text{hypergeom}(60, 40, 30)$.
Note that `phyper` in R takes parameters $(N_{1\cdot}, n - N_{1\cdot}, N_{\cdot 1})$.

```
> pl=phyper(23,40,20,30)
> pr=1-phyper(23-1,40,20,30)
> c(pl,pr)
[1] 0.97305140 0.08511085
```

Conclusion? Neither is less than $2.5\% = \alpha/2$. H_0 is not rejected. We cannot conclude that exact studies are chosen more frequently by men.

Fisher's exact test

This exact test is called Fisher's exact test for 2 x 2 tables.

```
> study=matrix(c(23,17,7,13),nrow=2,ncol=2)
> study
     [,1] [,2]
[1,]   23    7
[2,]   17   13
> fisher.test(study)

Fisher's Exact Test for Count Data

data:  study
p-value = 0.1702
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.7290557 9.0441625
sample estimates:
odds ratio
   2.47347
```

Remark This two-sided p-value is different from $2 \cdot \min\{P(N_{11} \geq 23), P(N_{11} \leq 23)\}$.

Model(s) II and χ^2-test

General contingency table – Models II

	B_1	$\cdots\cdots\cdots$	B_j	$\cdots$	B_c	total
A_1	N_{11}	$\cdots\cdots\cdots$	N_{1j}	$\cdots$	N_{1c}	$N_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	N_{i1}	$\cdots\cdots\cdots$	N_{ij}	$\cdots$	N_{ic}	$N_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	N_{r1}	$\cdots\cdots\cdots$	N_{rj}	$\cdots$	N_{rc}	$N_{r\cdot}$
total	$N_{\cdot 1}$	$\cdots\cdots\cdots$	$N_{\cdot j}$	$\cdots$	$N_{\cdot c}$	$n = N_{\cdot\cdot}$

The general form of a contingency table, with row variable A (r categories) and column variable B (c categories).

Model II A – more details next week!

1 sample of size n. 1 rc-nomial distribution with probabilities p_{ij},

$$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = 1$$

Null hypothesis No dependence between row & column variable,
$p_{ij} = p_{i\cdot}p_{\cdot j}$ for $i = 1, \ldots, r, j = 1, \ldots, c$.
($p_{ij}, p_{i\cdot}, p_{\cdot j}$ all unspecified!)

Test statistic

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(N_{ij} - n\hat{p}_{ij}\right)^2}{n\hat{p}_{ij}} \qquad \text{with } \hat{p}_{ij} = \frac{N_{i\cdot}N_{\cdot j}}{n^2}$$

Distribution $X^2 \sim \chi^2_{(r-1)(c-1)}$ under H_0, approximately.

Model II B – more details next week!

r samples of size $N_{i\cdot}$ each. r c-nomial distributions with probabilities p_{ij},

$$\sum_{j=1}^{c} p_{ij} = 1 \qquad \text{for } i = 1, \ldots, r$$

Null hypothesis The r samples are homogeneous, i.e.
$p_{1j} = p_{2j} = \cdots = p_{rj} \equiv p_j$ for $j = 1, \ldots, c$.
(all p_j unspecified!)

Test statistic

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(N_{ij} - n\hat{p}_{ij}\right)^2}{n\hat{p}_{ij}} \qquad \text{with } \hat{p}_{ij} = \frac{N_{i\cdot} N_{\cdot j}}{n^2}$$

Distribution $X^2 \sim \chi^2_{(r-1)(c-1)}$ under H_0, approximately.

Model II C – more details next week!

c samples of size $N_{\cdot j}$ each. c r-nomial distributions with probabilities p_{ij},

$$\sum_{i=1}^{r} p_{ij} = 1 \qquad \text{for } j = 1, \ldots, c$$

Null hypothesis The c samples are homogeneous, i.e.
$p_{i1} = p_{i2} = \cdots = p_{ic} \equiv p_i$ for $i = 1, \ldots, r$.
(all p_i unspecified!)

Test statistic

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(N_{ij} - n\hat{p}_{ij}\right)^2}{n\hat{p}_{ij}} \qquad \text{with } \hat{p}_{ij} = \frac{N_{i\cdot} N_{\cdot j}}{n^2}$$

Distribution $X^2 \sim \chi^2_{(r-1)(c-1)}$ under H_0, approximately.

to finish

To wrap up

Today we discussed

1. Two sample problems: Tests for correlation
2. Fisher's exact test
3. Chisquare test (beginning)

Next week More on categorical data, and linear regression