

## 8 The independence problem

### 8.1.1 Example (Tuna quality)

```
## Hollander & Wolfe (1973), p. 187f.
## Assessment of tuna quality. We compare the Hunter L measure of
## lightness to the averages of consumer panel scores (recoded as
## integer values from 1 to 6 and averaged over 80 such values) in
## 9 lots of canned tuna.

x <- c(44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7, 45.2, 60.1)
y <- c( 2.6,  3.1,  2.5,  5.0,  3.6,  4.0,  5.2,  2.8,  3.8)

## The alternative hypothesis of interest is that the
## Hunter L value is positively associated with the panel score.
plot(x,y)
cor.test(x, y, method = "kendall", alternative = "greater")
## => p=0.05972

cor.test(x, y, method = "kendall", alternative = "greater",
         exact = FALSE) # using large sample approximation
## => p=0.04765

## Compare this to
cor.test(x, y, method = "spearm", alternative = "g")
cor.test(x, y,                alternative = "g")
```

What's the question?

What's the first thing to do?

Kendall tau = 0.4444

What's level.090 test?

Reject if  $K \geq 14$

Data: we obtain  $n$  bivariate observations  $(X_1, Y_1) \dots (X_n, Y_n)$ , one observation on each of  $n$  subjects.

Assumptions:

They are a random sample from a continuous bivariate population.

## 8.2 A distribution free test for independence based on signs

$H_0$ :  $X$  and  $Y$  random variables are independent, i.e.  $F_{X,Y}(x, y) = F_X(x)F_Y(y)$  for all  $(x, y)$ .

### the Kendall population correlation coefficient

$$\tau = 2P\{(Y_2 - Y_1)(X_2 - X_1) > 0\} - 1 = 0 \text{ if } X \text{ and } Y \text{ are independent (why?)}$$

Does  $\tau=0$  imply independence?

### 8.2.1 Procedure

Calculate the values of  $n(n-1)/2$  paired sign statistics  $Q((X_i, Y_i), (X_j, Y_j))$  for  $1 \leq i < j \leq n$ , where  $Q((a, b), (c, d)) = 1[(d - b)(c - a) > 0] - 1[(d - b)(c - a) < 0]$

The Kendall's statistic is then

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^n Q((X_i, Y_i), (X_j, Y_j))$$

One-sided upper tail test: To test  $H_0$  vs the alternative  $H_1$ :  $X$  and  $Y$  are positively correlated, i.e.  $\tau > 0$ , at the  $\alpha$ -level of significance, Reject  $H_0$  if  $K \geq k_\alpha$  (use Table A.30)

### Interpretation of $\tau > 0$ ?

One-sided lower tail test: To test  $H_0$  vs the alternative  $H_2$ :  $X$  and  $Y$  are negatively correlated, i.e.  $\tau < 0$ , at the  $\alpha$ -level of significance, Reject  $H_0$  if  $K \leq -k_\alpha$  (use Table A.30)

Two-sided test: To test  $H_0$  vs the alternative  $H_3$ :  $X$  and  $Y$  are dependent, i.e.  $\tau \neq 0$ , at the  $\alpha$ -level of significance, Reject  $H_0$  if  $|K| \geq k_{\alpha/2}$  (use Table A.30)

### 8.2.2 Large-sample approximation

$$E_0 K = 0$$

$$\text{var}_0 K = \frac{n(n-1)(2n+5)}{18}$$

and the standardized version

$$K^* = \frac{K - E_0 K}{\{\text{var}_0 K\}^{1/2}} \sim N(0,1) \text{ approximately as } n \text{ tends to infinity.}$$

The normal theory approximation of the procedures follow accordingly. (how?)

\*  $K'$  = # of concordant pairs

$K''$  = # of discordant pairs

$K = K' - K''$

\* R output gives  $T = K'$ . 26 for the example. How to compute  $K$  from  $T$ ?

$N(n-1)/2 = 9*8/2 = 36$ .

$K'' = 10$ .  $K = 26 - 10 = 16$ .

\* To compute  $K$ , it's enough to know only ranks of  $X$ s and  $Y$ s. (why?)

\* What's the range of  $K$ ? Maximum and minimum? When do they occur?

\* Power and sample size results: it's there. Complicated.

### 8.3 Estimator (Kendall's tau)

The estimator of the Kendall population correlation coefficient  $\tau$  is

$$\hat{\tau} = \frac{2K}{n(n-1)}$$

called Kendall's sample rank correlation coefficient

\*  $-1 \leq \hat{\tau} \leq 1$

Statistic

#### 8.3.1 Computer implementation in R and Minitab

Well implemented.

### 8.4 An asymptotically distribution-free confidence interval based on the Kendall statistics

Skip.

### 8.5 A distribution free test for independence based on ranks (Spearman)

Order the  $X$  values and  $Y$  values and call their ranks  $R_i$  and  $S_i$  each.

The Spearman rank correlation coefficient is given by

$$r_s = \frac{12 \sum_{i=1}^n \left\{ \left[ R_i - \frac{n+1}{2} \right] \left[ S_i - \frac{n+1}{2} \right] \right\}}{n(n^2 - 1)} = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2 - 1)}$$

where  $D_i = S_i - R_i$ .

One-sided upper tail test: To test  $H_0$  vs the alternative  $H_1$ : X and Y are positively associated, at the  $\alpha$ -level of significance,

Reject  $H_0$  if  $r_s \geq r_{s,\alpha}$  (use Table A.31)

**Interpretation of  $\tau > 0$ ?**

One-sided lower tail test: To test  $H_0$  vs the alternative  $H_2$ : X and Y are negatively associated, at the  $\alpha$ -level of significance,

Reject  $H_0$  if  $r_s \leq -r_{s,\alpha}$

Two-sided test: To test  $H_0$  vs the alternative  $H_3$ : X and Y are dependent, i.e.  $\tau \neq 0$ , at the  $\alpha$ -level of significance,

Reject  $H_0$  if  $|r_s| \geq r_{s,\alpha}$

### 8.5.1 Large-sample approximation

$$E_0(r_s) = 0$$

$$\text{var}_0(r_s) = \frac{1}{n-1}$$

and the standardized version

$$r_s^* = \frac{r_s - E_0(r_s)}{\{\text{var}_0(r_s)\}^{1/2}} = (n-1)^{1/2} r_s \sim N(0,1) \text{ approximately as } n \text{ tends to infinity.}$$

The normal theory approximation of the procedures follow accordingly. (how?)

\* **Motivation**, in terms of  $D_i$ ??

\* Pearson's product moment sample correlation coefficient.

\* the Spearman rank corr. Coeff. Is simply the Pearson corr. Coeff. Applied to the rank vectors.

## 9 Regression

Skip!

## 10 Comparing two success probabilities

### 10.1.1 Breast cancer data

Patients	Liver Scan		Totals
	Yes	No	
Black	4	8	12
White	1	20	21
Totals	5	28	33

Is there a significance difference between the chance of white patient receiving a scan and the chance of a black patient receiving a scan?

### 10.1.2 Death penalty and gun registration

Gun Registration	Liver Scan		Totals
	Yes	No	
Favor	784	236	1020
Oppose	311	66	377
Totals	1095	302	1397

Is there an association between the attitude toward gun registration and attitude toward the death penalty?

- What's the difference between the two sampling scheme?

$P_1 = \text{Pr}(\text{a black patient receive a liver scan})$

$P_2 = \text{Pr}(\text{a white patient receive a liver scan})$

$H_1: p_1 > p_2$ .

$\hat{p}_1, \hat{p}_2, \hat{p}, \hat{S}_d$ ...

$A = 2.20$

$\alpha = 0.05$  level test- accept?

P-value?

Confirm:  $\chi^2 = 4.85 \sim A^2$

### 10.1.3 Data

We observe the outcomes of  $n_1$  and  $n_2$  independent repeated Bernoulli trials, each with success probability  $p_1$  and  $p_2$  respectively.

	Successes	Failures	Totals
Sample 1	$O_{11}$	$O_{12}$	$n_{1.}$
Sample 2	$O_{21}$	$O_{22}$	$n_{2.}$
Totals	$n_{.1}$	$n_{.2}$	$n_{..}$

### 10.1.4 Assumptions

A1.  $O_{11}$  is the number of successes observed in  $n_1$  independent Bernoulli trials, each with success probability  $p_1$ .

A2.  $O_{21}$  is the number of successes observed in  $n_2$  independent Bernoulli trials, each with success probability  $p_2$ .

A3. The two samples are independent.

Want to test:

$$H_0 : p_1 = p_2 = p$$

$p$  is not specified.

## 10.2 Approximate tests and confidence intervals for $p_1 - p_2$ (Pearson)

Let

$$D = \hat{p}_1 - \hat{p}_2 \text{ where}$$

$$\hat{p}_1 = \frac{O_{11}}{n_{1.}}, \hat{p}_2 = \frac{O_{21}}{n_{2.}}$$

It's SD is estimated by

$$SD = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_{1.}} + \frac{\hat{p}(1-\hat{p})}{n_{2.}}}$$

where

$$\hat{p} = \frac{O_{11} + O_{21}}{n_{1.} + n_{2.}} \text{ (why?)}$$

Under  $H_0$ , the standardized version of  $D$

$$A = \frac{\hat{p}_1 - \hat{p}_2}{SD} \sim N(0,1) \text{ approximately as both sample sizes goes to infinity.}$$

Approximate one-sided upper-tail, lower-tail, two-sided test

To test  $H_0$  versus  $H_1: p_1 - p_2 > 0, < 0, \neq 0$  at the approximate  $\alpha$  level of significance, Reject  $H_0$  if  $A \geq z_\alpha, A \leq -z_\alpha, |A| \geq z_{\alpha/2}$ .

Approximate CI for  $p_1 - p_2$  is given by

$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \tilde{SD}$  where

$$\tilde{SD} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Why??

### 10.2.1 2x2 chi-squared test of homogeneity

The equivalent of the large-sample two-sided test.

If  $H_0$  is true, the best estimator of  $p$  is

$\hat{p} = \frac{n_{1.}}{n_{..}}$  and the expected values of the random quantities  $O_{11}, \dots, O_{22}$  are estimated by

$E_{11} = n_{1.} \times \hat{p} = \frac{n_{1.} \times n_{.1}}{n_{..}}$

and  $E_{12}, E_{21}, E_{22}$ ?

A measure of discrepancy between the observed frequencies and the their estimators under  $H_0$ , is the chi-squared statistic given by

$$\chi^2 = \sum_{i,j=1,2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Interpretation? Motivation?

It can be shown that

$A^2 = \chi^2$ .

A simple formula is given by

$$\chi^2 = \frac{n_{..} (O_{11}O_{22} - O_{12}O_{21})^2}{n_{.1}n_{.2}n_{1.}n_{2.}}$$

The two sided test is equivalent to the test which

Reject  $H_0$  if  $\chi^2 \geq \chi_{\alpha,1}^2$  (why?)

### 10.2.2 The 2x2 chi-squared test of independence

Data: Neither of the four sums  $n_{.1}$  etc are fixed but each observation from a general population is cross-classified on the basis of two characteristics.

	C	Not C	Totals
D	$O_{11}$	$O_{12}$	$n_{1.}$

Not D	$O_{21}$	$O_{22}$	$n_{2.}$
Totals	$n_{.1}$	$n_{.2}$	$n_{..}$

$p_{11}=P(C \text{ and } D)$

$p_{12}=P(\text{not } C \text{ and } D)$

$p_{21}=P(C \text{ and not } D)$

$p_{22}=P(\text{not } C \text{ and not } D)$

$H_0$ : the occurrences of the two characteristics are independent.

(More formally,  $H_0: p_{ij}=p_{i.}p_{.j}$ )

Then we have the same procedure as above!

- \* Different sampling scheme. For homogeneity, row columns are fixed. For independence (**cross-sectional sampling**), only the total is fixed.
- Yates' correction for continuity
- Sample size determination (p.467)
- The degree of association is measured by, e.g. the odds ratio

### 10.2.3 Computer Implementation

R: chi.sq

Minitab:

### 10.3 An exact test for the difference between two success probabilities (Fisher)

The conditional distribution:

$$P(O_{11} = x | n_{1.}, n_{2.}, n_{.1}, n_{.2}) = \frac{\binom{n_{1.}}{x} \binom{n_{2.}}{n_{.1} - x}}{\binom{n_{..}}{n_{.1}}}$$

where

$$\max(0, n_{1.} - n_{.1} - n_{.2}) \leq x \leq \min(n_{1.}, n_{.1}).$$

Hypergeometric distribution.

Fisher's exact test judges whether  $O_{11}$  is significantly small or significantly large with respect to the conditional distribution defined by this equation.

To test  $H_0$  vs  $H_1: p_1 < p_2$ ,

Reject  $H_0$  if  $O_{11} \leq q_\alpha$  where  $q_\alpha$  is chosen so that  $\Pr(O_{11} \leq q_\alpha | n_{1.}, n_{.1}, n_{2.}, n_{.2}) = \alpha$ .