5 Tests for Trends and Associations

5.1 A Permutation Test for Correlation and Slope

Sample data pairs (X_i, Y_i), for $i = 1, \ldots, n$.

The *population correlation coefficient* is defined by $\rho = \sigma_{x,y}/(\sigma_x \sigma_y)$,

where $\sigma_{x,y} = E(X - \mu_x)(Y - \mu_y)$.

ρ measures the strength of the **linear** relationship between two variables.

The **Pearson** *sample correlation coefficient* is defined by

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{j=1}^{n}(Y_j - \bar{Y})^2}}.$$

Examples (interpreting r): For the following models, generate 100 pairs of observations, determine the sample correlation coefficient, and plot the data.

(a) $X \sim N(0,\ 1)$, $Y = X + \varepsilon$, where the ε are independent $N(0,\ 0.5)$.

(b) $X \sim N(0,\ 1)$, $Y = -X + \varepsilon$, where the ε are independent $N(0,\ 0.5)$.

(c) $X \sim N(20,\ 1)$, $Y = X + \varepsilon$, where the ε are independent $N(30,\ 0.5)$.

(d) $X \sim N(0,\ 1)$, $Y = X + \varepsilon$, where the ε are independent $N(0, 0.2)$.

(e) $X \sim N(0,\ 1)$, $Y = 5X + 37$

(f) $X \sim N(0,\ 1)$, $Y = -5X + 37$

(g) $X \sim N(20,\ 5)$, $Y = X + \varepsilon$, where the ε are independent $N(60,\ 5)$.

How much variability in Y can be explained by the linear relationship between X and Y?

(h) $X \sim N(20,\ 5)$, $Y = \varepsilon$, where the ε are independent $N(60,\ 5)$.

(i) $X \sim Unif(10,\ 30)$, $Y = -5(X-20)^2 + \varepsilon$, where the ε are independent $N(900, 25)$.

□

Problem #5.1.1 (crying babies and IQ scores),

problem5.1.1.txt: (This data set is from Karelitz *et al.* (1964), *Child Development*.) To test whether children who cry more actively as babies later tend to have higher IQs, a cry count was taken for a sample of 38 children aged five days and was later compared with their Stanford-Binet IQ scores at age three with the results shown below:

Cry count	10	20	17	12	12	15	19	12	14	23
IQ score	87	90	94	94	97	100	103	103	103	103

Cry count	15	14	13	27	17	12	18	15	15	23
IQ score	104	106	106	108	109	109	109	112	112	113

Cry count	16	21	16	12	9	13	19	18	19	16
IQ score	114	114	118	119	119	120	120	124	132	133

Cry count	22	31	16	17	26	21	27	13
IQ score	135	135	136	141	155	157	159	162

(a) Plot the data.

```
> z = read.table2( "problem5.1.1.txt", header=TRUE )
```

(b) Determine the sample correlation coefficient.

(c) What is an interesting hypothesis test for this data set?

□

Define $t_{\text{corr}} = r\ \sqrt{(n-2)/(1-r^2)}$ (need not memorize).

If $\rho = 0$, and if the (x, y) are based on a simple random sample from a *bivariate normal* distribution, then t_{corr} is t-distributed with $(n-2)$ degrees of freedom.

Revisit problem #5.1.1 (crying babies and IQ scores),

problem5.1.1.txt: (This data set is from Karelitz *et al.* (1964), *Child Development*.) To test whether children who cry more actively as babies later tend to have higher IQs, a cry count was taken for a sample of 38 children aged five days and was later compared with their Stanford-Binet IQ scores at age three.

(a) State the null and alternative hypotheses.

(b) Determine the value of the standardized test statistic for the *parametric* test, using hand-calculations.

(c) Determine the *asymptotic* p-value of this test using hand-calculations, and state the conclusion.

(d) Plot the *asymptotic* distribution of your standardized test statistic under H_0, and shade in the appropriate region corresponding to the p-value.

(e) Determine the *asymptotic* p-value of this test using `cor.test`.

(f) What is the 95% **lower** confidence **bound** on ρ?

(g) Does association imply causation?

□

5.1.2 Slope of the Least Squares Line

A *simple linear regression model* is defined by

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where the εs are independent and identically distributed random variables with mean zero and positive finite variance.

However, if the ε and X are **normal** and mutually independent, then X and Y are *bivariate normal*.

We define the *least squares* estimates of β_0 and β_1 by:

$$\hat{\beta}_1 = r s_y / s_x \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

(memorize these two formulas).

The *least squares line*, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, is the line which minimizes the sum of the squares of the vertical distances between the observations and the line.

For large n, what are the *asymptotic* values in the formula $\hat{\beta}_1 = r s_y / s_x$?

Revisit problem #5.1.1 (crying babies and IQ scores),

problem5.1.1.txt: (This data set is from Karelitz *et al.* (1964), *Child Development*.) To test whether children who cry more actively as babies later tend to have higher IQs, a cry count was taken for a sample of 38 children aged five days and was later compared with their Stanford-Binet IQ scores at age three.

(a) Which variable should be X, and which should be Y?

(b) Determine the least squares equation using hand-calculations.

(c) Plot the least squares line on the scatterplot.

(d) Predict the child's IQ at age three years if the cry count at age five days was 25.

(e) Determine the least squares equation using the function `lm`.

(f) Verify that the least squares equation goes through the sample means.

(g) Regarding the *simple linear regression model*, what is an interesting hypothesis test?

□

5.1.3 The Permutation Test

Specifically, this **permutation correlation test** is for *nonzero*, *negative*, or *positive* Pearson correlation (or slope).

If the simple linear regression model holds, and the null hypothesis of $\beta_1 = 0$ holds, what can we say about X and Y?

How should a *permutation* test be performed?

Problem 5.1.2: In the data set below, test for negative *Pearson* correlation.

x	11	15	21
y	139	143	87

(a) State the null and alternative hypotheses.

(b) Determine the *permutation distribution* of $\hat{\beta}_1$ and r.

$X_1 = 11$	$X_2 = 15$	$X_3 = 21$		
Y_1	Y_2	Y_3	$\hat{\beta}_1$	r
87	139	143	5.21	0.839
87	143	139	4.74	0.763
139	87	143	1.11	0.178
139	143	87	-5.53^*	-0.890^*
143	87	139	0.316	0.051
143	139	87	-5.84	-0.941

$\hat{\beta}_1$	r	probability
-5.84	-0.941	1/6
-5.53^*	-0.890^*	1/6
0.316	0.051	1/6
1.11	0.178	1/6
4.74	0.763	1/6
5.21	0.839	1/6
sum		1

(c) Determine the p-value of the *permutation correlation* test, using hand-calculations.

(d) What would be the p-value for a two-sided test?

(e) Obtain the *simulated* p-value using the function `perm.cor.test`.

(f) What assumptions were needed to perform this test from part (e)?

(g) Using the one-sided hypothesis test and the **parametric** approach, determine the p-value.

(h) What assumptions were needed to perform this test from part (g)?

(i) Are the assumptions needed to perform this test reasonable?

□

In general, for n data pairs, how many permutations (groupings) exist when performing the permutation correlation test?

Homework p. 189: Exercise 5.1*

Hints for homework exercise 5.1*: Let x be *height* and y be *weight.* Show **all** *R*-code and *R*-output used for generating the permutation distribution. You do **NOT** need to state H_0, H_a, p-value, or conclusion. Introduce the question number as a comment using "#" or in red using .html code; e.g., `<span style="color: red"> Exercise 5.1 </span>`.

5.1.4 Large-Sample Approximation for the Permutation Distribution of r

Suppose n independent pairs of observations (X, Y) are sampled from distributions with positive finite variances such that X and Y are **independent**.

What is ρ, the *population* Pearson correlation coefficient?

What is r, the *sample* Pearson correlation coefficient?

Consider the permutation distribution of r.

What is Er (where the expectation is taken conditionally on the permutation distribution)?

$\text{var}(r) \approx 1/(n-1)$, conditional on the permutation distribution.

Example: Sample n independent pairs of observations (X, Y) from distributions with positive finite variances such that X and Y are **independent**, where n is large.

(a) Determine r_n (a fixed function of n) such that there is approximately a 95% chance that r (the sample correlation coefficient) will be between $-r_n$ and r_n.

(b) Set $n = 100$ in part (a).

(c) Set $n = 400$ in part (a).

(d) Set $n = 1000$ in part (a).

(e) Set $n = 10,000$ in part (a).

□

5.2 Spearman Rank Correlation

Problem 5.2.1: Consider the following data set.

X	0	2	3	4	6
Y	0	16	81	256	1296

(a) Determine the *Pearson* correlation coefficient.

```
> x = c( 0, 2, 3, 4, 6 )
> y = c( 0, 16, 81, 256, 1296 )
> cor( x, y )
```

(b) Is r, the *Pearson* correlation coefficient, a reasonable measure?

(c) Compare the ranks of X with the ranks of Y.

□

5.2.1 Statistical Test for Spearman Rank Correlation

The **Spearman rank correlation**, r_s, is the *Pearson* correlation coefficient applied to ranks.

Thus, the *Spearman* rank correlation is not heavily influenced by outliers.

The *Spearman* rank correlation measures the *association* between two variables.

Revisit problem 5.2.1:

(a) Determine the *Spearman* rank correlation for this data set.

```
> x = c( 0, 2, 3, 4, 6 )
```

```
> y = c( 0, 16, 81, 256, 1296 )
```

(b) Test if the association between the two variables is positive, using the *Spearman* rank correlation and hand-calculations.

(c) Test if the population *Spearman* rank correlation is positive, using the function `cor.test`.

□

Revisit problem 5.1.2: In the data set below, test for negative *Spearman* correlation.

x	11	15	21
y	139	143	87

rank(x)	1	2	3
rank(y)	2	3	1

(a) State the null and alternative hypotheses.

(b) Determine the *permutation distribution* of r_s.

rank(X_1) = 1	rank(X_2) = 2	rank(X_3) = 3	
rank(Y_1)	rank(Y_2)	rank(Y_3)	r_s
1	2	3	1
1	3	2	0.5
2	1	3	0.5
2	3	1	-0.5^*
3	1	2	-0.5
3	2	1	-1

r_s	probability
-1	$1/6$
-0.5^*	$2/6$
0.5	$2/6$
1	$1/6$
sum	1

(c) Determine the p-value of the *Spearman correlation* test, using hand-calculations.

(d) Determine the p-value of the *Spearman correlation* test, using the function `cor.test`.

(e) What would be the p-value for a two-sided test?

(f) Obtain the *simulated* p-value using the function `perm.cor.test`.

(g) What assumptions were needed to perform this test?

□

Homework p. 189: Exercises 5.2 (Spearman only), 5.3 (Pearson and Spearman only)

5.2.2 Large-Sample Approximation

Recall from section 5.1.4 that for large n (and finite positive variances) and independent X and Y, the distribution of $Z = r\sqrt{n-1}$ is approximately standard normal, where r is the sample *Pearson* correlation coefficient.

Adjustment for Ties

Two options:

⊙ Apply the *Pearson* correlation to the ranks adjusted for *ties*.

⊙ Use the normal approximation.

Revisit problem #5.1.1 (crying babies and IQ scores), problem5.1.1.txt: (This data set is from Karelitz *et al.* (1964), *Child Development*.) To test whether children who cry more actively as babies later tend to have higher IQs, a cry count was taken for a sample of 38 children aged five days and was later compared with their Stanford-Binet IQ scores at age three.

(a) Determine the sample *Spearman* correlation coefficient.

(b) State the null and alternative hypotheses.

(c) Estimate the *exact* p-value of this test using simulations, and state the conclusion.

(d) Determine the value of the standardized test statistic for the *large-sample* test, using hand-calculations.

(e) Determine the *asymptotic* p-value of this test using hand-calculations.

(f) Plot the *asymptotic* distribution of your standardized test statistic under H_0, and shade in the appropriate region corresponding to the p-value.

(g) Determine the *asymptotic* p-value of this test using `cor.test`.

□

Caution in Using the Pearson or Spearman Correlation

Caution: Time-dependent data may invalidate the independence between the (X, Y) data pairs.

Example: Suppose x is the year, and y is the average ocean temperature for the year.

□

5.4 Permutation Tests for Contingency Tables

Scenario: We have two categorical variables, and we enter the data into a table.

5.4.1 Hypotheses to be Tested and the Chi-Square Statistic

Problem 5.4.1 (gender and handedness; hypothetical data): The following hypothetical data are based on gender and the preferred hand for writing among 10-year-old children.

Observed	Left	Right	total
Girls	5	83	88
Boys	7	65	72
total	12	148	160

The hypothesis test will be a **test for association**.

(a) What is the hypothesis test of interest?

(b) Estimate the *expected* (i.e., average) number of *left*-handed *girls* under H_0.

(c) Estimate the *expected* (i.e., average) number of *right*-handed *girls* under H_0.

(d) Estimate the *expected* (i.e., average) number of *left*-handed *boys* under H_0.

(e) Estimate the *expected* (i.e., average) number of *right*-handed *boys* under H_0.

Expected	Left	Right	total
Girls	6.6	81.4	88
Boys	5.4	66.6	72
total	12	148	160

(f) What is the general rule for computing these expected cell frequencies under H_0?

Formula: The *chi-square* test statistic is defined as:

$$V = \sum_{i=1}^{r} \sum_{j=1}^{c} (n_{ij} - e_{ij})^2 / e_{ij}.$$

(g) Compute the *chi-square* test statistic for this data set.

Asymptotic result: For large sample sizes (i.e. if $e_{ij} \geq 5$ for all cells) and

independent observations, the test statistic V has approximately a χ^2 distribution with degrees of freedom equal to (number of rows $-$ 1) $\times$ (number of columns $-$ 1).

(h) How many degrees of freedom are associated with this test?

(i) Determine the *asymptotic* p-value associated with this test.

(j) Plot the *asymptotic* distribution of your test statistic under H_0, and shade in the appropriate region corresponding to the p-value.

(k) Determine the *asymptotic* p-value using the function `chisq.test`.

□

Example (gender and handedness; real data): In a survey of Scottish school children, aged approximately ten to twelve years, the teacher observed

whether the pupil wrote with the left or right hand, with the following results (Clark, 1957, *Left Handedness*, University of London Press, London).

Observed	Left	Right	Percentage left
Girls	1,478	25,045	5.57
Boys	991	12,629	7.28

□

Problem 5.4.2 (Roosevelt), problem5.4.2.txt: The following results were obtained in a 1948 study of the 1944 Presidential election in Elmira, New York (McCarthy, 1957, *Introduction to Statistical Reasoning*, McGraw-Hill).

	Individual Interviewed on			Percentage reached
1944 Pres. vote	First call	Second or later call	Total	on first call
Roosevelt	138	217	355	38.9
Dewey	124	200	324	38.3
Did not vote	90	142	232	38.8
Other or too young	39	78	117	33.3
Total	391	637	1028	38.0

Test the hypothesis that the distribution of responses is the *same* for individuals reached on the first call as for those interviewed on the second or later calls.

□

5.4.2 Permutation Chi-Square Test

Herein, we use the same *chi-square* test statistic,

$$V = \sum_{i=1}^{r} \sum_{j=1}^{c} (n_{ij} - e_{ij})^2 / e_{ij}.$$

However, we do NOT use the χ^2-distribution to approximate the p-value.

Instead, the **permutation distribution** of the test statistic is determined exactly or is simulated.

Determining the permutation distribution:

Revisit problem 5.4.1 (gender and handedness; hypothetical data): The following hypothetical data are based on gender and the preferred hand for writing among 10-year-old children.

Observed	Left	Right	total
Girls	5	83	88
Boys	7	65	72
total	12	148	160

⊙ Fix all of the margin totals: 12, 148, 88, 72, 160.

Is our observed table *rare* under H_0?

Under these fixed margins, what are other possible values of the table? Specifically, what are the possible values for X, the number of *left*-handed *girls*?

Based on a permutation involving, say, 10 *left*-handed *girls*, complete the rest of the table.

⊙ How many degrees of freedom are associated with this test?

⊙ The chi-square test statistic, V, may be determined for each value of X, the number of *left*-handed *girls*, for $X = 0, 1, \ldots, 12$.

⊙ To obtain the permutation distribution of V under these fixed margins, the probabilities of V (or X) may be determined based on the *hypergeometric* distribution. A *hypergeometric* distribution is similar to a *binomial* distribution, except that a *hypergeometric* distribution is based on sampling withOUT replacement.

Example (aside): Suppose a classroom has 10 *female* students and 7 *male* students. Sample 5 students (withOUT replacement) at random, and let W be the number of *female* students in the sample. Then W has a *hypergeometric* distribution. □

⊙ The p-value for this *permutation* distribution is based on either **all** possible permutations of V or a **large** number of **simulated** *permutations* of V.

□

5.5 Fisher's Exact Test for a 2 × 2

Contingency Table

Fisher's exact test is similar to the *permutation chi-square* test, except the permutations are based on X (say, the number of *left*-handed *girls*), rather than V, where

$$V = \sum_{i=1}^{r} \sum_{j=1}^{c} (n_{ij} - e_{ij})^2 / e_{ij}.$$

The permutation probabilities of X are again determined by the *hypergeometric* distribution under the fixed margin totals.

Revisit problem 5.4.1 (gender and handedness; hypothetical data): The following hypothetical data are based on gender and the preferred hand for writing among 10-year-old children.

Observed	Left	Right	total
Girls	5	83	88
Boys	7	65	72
total	12	148	160

Let X (the test statistic) be the number of *left*-handed *girls*.

What does a *large* value of X suggest?

What does a *small* value of X suggest?

Test if the proportion of *girls* who are *left*-handed is **smaller** than the proportion of *boys* who are *left*-handed.

□

Example (Bush vs. Gore, Election of 2000):

Summary: The Presidential Election between George W. Bush and Albert Gore took place on November 7, 2000. The vote was quite close in Florida, the winner of which would win the election. On November 8 the count resulted in a small lead for Bush. Gore sought a manual recount in several Florida counties, so the process of recounting votes began, as permitted by the Florida Supreme Court. Bush argued that recounting only certain counties violated the *equal protection* clause of the fourteenth amendment to the U.S. Constitution, and Bush also argued that Florida's electors should be selected by the December 12 deadline.

On December 11, a 5–4 majority of the U.S. Supreme Court ruled that no constitutionally-valid recount could be completed by the December 12 deadline, effectively ending the recounts.

Is there statistically significant evidence that the U.S. Supreme Court Justices tended to favor their own *political parties* (i.e., according to the political party of the President who appointed the Justice)?

Use *Fisher's* exact test.

Justice	Appointed by President	Decision
William Rehnquist	Reagan	End recount
Sandra Day O'Connor	Reagan	End recount
Antonin Scalia	Reagan	End recount
Anthony Kennedy	Reagan	End recount
Clarence Thomas	G. H. W. Bush	End recount
John Paul Stevens	Ford	Continue recount
David Souter	G. H. W. Bush	Continue recount
Ruth Ginsburg	Clinton	Continue recount
Stephen Breyer	Clinton	Continue recount

Is the proportion of *Republican*-appointed Justices who voted to *end* the recount significantly *large*?

□

Example (sampling and the U.S. Census, 1999):

The U.S. Census contains error, in that some individuals are not counted, quite often in regions dominated by Democrats. The statistical technique of **sampling** could greatly reduce this error, and consequently could affect the apportionment in the House of Representatives in favor of the Democrats. The Clinton administration wanted to use **sampling**, but the Republicans opposed the use of **sampling**, for determining seats in the House of Representatives. On January 25, 1999, the U.S. Supreme Court ruled 5 to 4 **against** the use of **sampling** in the Census for the purpose of apportioning seats in the House of Representatives among the states.

Is there statistically significant evidence that the U.S. Supreme Court Justices tended to favor their own *political parties* (i.e., according to the political party of the President who appointed the Justice)?

Use *Fisher's* exact test.

Justice	Appointed by President	Use sampling?
William Rehnquist	Reagan	no
Sandra Day O'Connor	Reagan	no
Antonin Scalia	Reagan	no
Anthony Kennedy	Reagan	no
Clarence Thomas	G. H. W. Bush	no
John Paul Stevens	Ford	yes
David Souter	G. H. W. Bush	yes
Ruth Ginsburg	Clinton	yes
Stephen Breyer	Clinton	yes

□

p. 189

Exercises

1 Find the permutation distribution of the slope of the least squares line for the height and weight data in the table.

Height	68	70	74
Weight	145	155	160

2 Generate the permutation distributions of Spearman's r_s (see Section 5.2) and Kendall's τ (see Section 5.3) for the data in Exercise 1.

3 The data in the table are the ages (in days) of concrete cylinders and the compressive strengths of the cylinders.

Age	3	7	15	24	85	180	360
Strength	2500	3200	4300	5300	5900	6700	6900

a Plot the data to show a nonlinear relationship. Compute Pearson's correlation, Spearman's correlation, and Kendall's tau.

b Test for significant association using each of the measures of association in part a.