5.1 A Permutation Test for Correlation and Slope

Bivariate Sampling
- the pairs are selected at random from a bivariate population of such pairs.

Fixed-X Sampling
- the pairs are obtained in an experiment where the values of X are fixed by the experimenter and one or more values of Y are obtained for each X.

5.1.1 The Correlation Coefficient

Population Correlation Coefficient

$$\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Pearson product-moment correlation

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

, $H_0: \rho = 0$ for either one-sided or two-sided tests

$$t_{corr} = \sqrt{\frac{n-2}{1-r^2}}\, r$$

, t-distribution with $n-2$ degrees of freedom under H_0

* used as the reference distribution for rejecting or not rejecting the null hypothesis.

5.1.2 Slope of Least Squares Line

$$\underbrace{SSE}_{\text{Sum of squares errors}} = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, to test whether or not there is a significant relationship between X and Y, we test $H_0: \beta_1 = 0$, and if ε's are normally distributed

$$t_{slope} = \hat{\beta}_1 \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{MSE}}, \quad MSE = \frac{SSE}{n-2}$$

- For bivariate sampling in which the conditional distribution of Y given X satisfies the linear regression model, it can be shown that the slope and correlation are related by the equation,

$$\beta_1 = \rho \frac{\sigma_Y}{\sigma_X} \rightsquigarrow \hat{\beta}_1 = r \frac{s_Y}{s_X}$$

5.1.3 The Permutation Test

Steps for a Permutation Test for slope or Correlation

1. Compute the slope of the least squares line $\hat{\beta}_{1,obs}$ from the original data.
2. Permute the Y's among the X's in the $n!$ possible ways.
3. For each permutation, compute the slope $\hat{\beta}_1$ of the least squares line.
4. For an upper-tail test, the p-value is,

$$P_{upper\text{-}tail} = \frac{\text{number of } \hat{\beta}_1\text{'s} \geq \hat{\beta}_{1,obs}}{n!}$$

, we may obtain lower-tail and two-tail p-values

* steps for the test of correlation are the same except that r is used
* Use sampling for large n.

Alternative Statistics for the Correlation and Slopes

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i - \sum_{i=1}^{n}(x_i-\bar{x})\bar{Y}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \quad \sum_{i=1}^{n}(x_i-\bar{x})\bar{Y} = 0$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n} x_iY_i - \bar{x}Y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \quad \sum_{i=1}^{n} x_iY_i = S_{xY}$$

5.1.4 Large-Sample Approximation for the Permutation Distribution of r

- Under the assumption of no relationship between X and Y, the mean of the permutation distribution or r is 0 and the variance is given by $Var(r) = \frac{1}{n-1}$

$$Z = r\sqrt{n-1}$$

Derivation of Variance of r

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{n\sigma_X\sigma_Y}, \quad \mathrm{Var}\left\{\sum_{i=1}^{n}(X_i-\bar{X})Y_i\right\} = \sum_{i=1}^{n}(X_i-\bar{X})^2\sigma_Y^2 + \sum_{i\neq j}(X_i-\bar{X})(X_j-\bar{X})\,\mathrm{cov}(Y_i,Y_j),$$

and notice $\mathrm{Cov}(Y_i, Y_j) = -\frac{\sigma_Y^2}{n-1}$, and $\sum_{i\neq j}(X_i-\bar{X})(X_j-\bar{X}) = -n\sigma_X^2$, therefore,

$$\mathrm{Var}\left\{\sum_{i=1}^{n}(X_i-\bar{X})Y_i\right\} = n\sigma_X^2\sigma_Y^2 + \frac{n\sigma_X^2\sigma_Y^2}{n-1} = \frac{n^2\sigma_X^2\sigma_Y^2}{n-1}$$

5.2 Spearman Rank Correlation

- Suppose $(X_i, Y_i) = 1, \ldots, n$, and let $R(X_i)$ and $R(Y_i)$

5.2.1 Statistical Test for Spearman Rank Correlation

- The statistical significance of the Spearman rank correlation is determined by applying the permutation test for correlation with ranked pairs (Section 5.1.3)

5.2.2 Large-Sample Approximation

$$\mathrm{Var}(r_s) = \frac{1}{n-1}$$

$$z = \frac{r}{\sqrt{\mathrm{Var}(r_s)}} = r_s\sqrt{n-1}$$

Adjustment for Ties

- Average ranks are assigned to ties among the X's, and average ranks are assigned to ties among the Y's
- An alternative formula for the Spearman rank correlation with ties is a variant of an alternative formula without ties. (See Exercise 13 at the end of the chapter)

$$r_s = 1 - \frac{6D}{n(n^2-1)}, \quad D = \sum_{i=1}^{n}[R(X_i) - R(Y_i)]^2 \quad \text{without ties}$$

$$r_s = \frac{1 - 6D/n(n^2-1) - C_1}{C_2}, \quad C_1 = \frac{\sum(S_i^3 - S_i) + \sum(t_i^3 - t_i)}{2n(n^2-1)}$$

$$C_2 = \sqrt{\left[1 - \frac{\sum(S_i^3 - S_i)}{n(n^2-1)}\right]\left[1 - \frac{\sum(t_i^3 - t_i)}{n(n^2-1)}\right]},$$

- where S_i denote the number of observations in the i^{th} group of tied observations among the X's, and t_i denote the number of observations in the i^{th} group of tied observations among the Y's.
- Apply $z = r_s\sqrt{n-1}$

Caution in Using the Spearman Correlation

- Time-dependant data may not work with the use of the Spearman statistic and permutation statistic for correlation because ε's are independent and identically distributed.

5.3 Kendall's Tau

Concordant:

- $X_i < X_j$ implies $Y_i < Y_j$ $\quad (X_i - X_j)(Y_i - Y_j) > 0$

Discordant:

- $X_i < X_j$ implies $Y_i > Y_j$, or vice-versa. $\quad (X_i - X_j)(Y_i - Y_j) < 0$

This idea leads us to define a measure of association between two variables based on counts of concordant and discordant pairs.

Kendall's Tau

$$\tau = 2\underbrace{P[(X_i - X_j)(Y_i - Y_j) > 0]}_{\text{the probability that } (X_i, Y_i) \text{ and } (X_j, Y_j) \text{ are concordant}} - 1$$

If there is no association

$P[(X_i - X_j)(Y_i - Y_j > 0] = \frac{1}{2}$, because $\tau = 0$

5.3.1 Estimating Kendall's Tau with No Ties in the Data

Suppose the data have no ties,

$$U_{ij} = \begin{cases} 1, & (X_i - X_j)(Y_i - Y_j) > 0 \\ 0, & (X_i - X_j)(Y_i - Y_j) < 0 \end{cases}, \text{ and let } V_i = \sum_{j=i+1}^{n} U_{ij}$$

V_i: # of concordant pairs

Then, $\frac{\sum_{i=1}^{n-1} V_i}{\binom{n}{2}}$ is the fraction of concordant pairs in the dataset, so the estimate of τ is $r_\tau = 2\left\{\frac{\sum_{i=1}^{n-1} V_i}{\binom{n}{2}}\right\} - 1$

* same method can be applied using ranks
* To test $H_0: \tau = 0$ against one-sided or two-sided alternative hypotheses, we may carry out a permutation test. Equivalently, the permutation distribution may be determined by permuting the ranks of the Y's among the ranks of the X's

5.3.2 Large-Sample Approximation

Without ties, $E(r_\tau) = 0$, $Var(r_\tau) = \frac{4n+10}{9(n^2-n)}$, $z = \frac{r_\tau}{\sqrt{Var(r_\tau)}}$

5.3.3 Adjustment for Ties in the Data

- For ties in either the X's or the Y's in the pairs, we set $U_{ij} = \frac{1}{2}$, and let S_i denote the number of ties in the i^{th} group of ties among the X's and t_i denote the number of ties in the i^{th} group of ties among the Y's,

$$A = \frac{\sum S_i(S_i-1)(2S_i+5) + \sum t_i(t_i-1)(2t_i+5)}{18}$$

$$B = \frac{[\sum S_i(S_i-1)(2S_i-2)][\sum t_i(t_i-1)(t_i-2)]}{9n(n-1)(n-2)}$$

$$C = \frac{[\sum S_i(S_i-1)(2S_i-2)][\sum t_i(t_i-1)(t_i-2)]}{2n(n-1)}$$

, and the variance with ties is

$$Var(r_{\tau\,ties}) = \underbrace{Var(r_\tau)}_{\text{the variance of } r_\tau \text{ without ties}} - \frac{4}{n^2(n-1)^2}(A - B - C)$$

5.4 Permutation Tests for Contingency Tables

5.4.1 Hypotheses to Be Tested and the χ^2 statistic

Consider 2 cases:

1. All n individuals are selected at random from a population and cross-classified according to row and column characteristics.
2. A fixed number $n_{i.}$ is selected according to row characteristic i, $i = 1, 2, 3, \ldots, r$ and classified according to the column characteristic.

- Define the expected cell proportions as $p_{ij} = \frac{E(n_{ij})}{n}$, then the row and column proportions are $p_{i.} = \sum_{j=1}^{c} p_{ij}$, $p_{.j} = \sum_{i=1}^{r} p_{ij}$.

they mean the same thing anyways

- The case 1 above has the null hypothesis $H_0: p_{ij} = p_{i.} p_{.j}$, and the alternative hypothesis is it fails for at least one row and column, indicating statistical dependencies.

- The case 2 above has the null hypothesis $H_0: p_{j|i} = p_{j|i'}$, meaning the conditional probabilities from row to row are the same. The alternative is the conditional probabilities from row to row are not all the same for at least one column.

Both of the null hypotheses mean "no association"

χ^2 - statistic for both cases:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \text{ where } e_{ij} = \frac{n_{i.} n_{.j}}{n}$$

* If the expected cell frequencies e_{ij} are 5 or greater, then χ^2 has an approximate χ^2 distribution with $(r-1)(c-1)$ d.f. However, when the expected cell frequencies are smaller than this, χ^2 approximation may no longer be sufficiently accurate, and this is especially so in tables with a large number of cells for which there are no responses. In such cases, a permutation χ^2 test may be used.

5.4.2 Permutation χ^2 Test

- there are $\binom{n}{r}$ possible permutations. n = # of units, r = # of rows
- compute the χ^2 statistic for each permutation.
- count the number of χ^2 statistic that is greater than or equal to the observed χ^2 statistic.

An Alternative Way to Permute the Data

- Suppose we interchange the roles of the rows and columns.
- the p-values are the same regardless of which way the random assignment is done.

5.4.3 Multiple Comparisons in Contingency Tables

- If the null hypothesis is rejected, then it may be of interest to determine where differences among proportions exist. These comparisons need conditional probabilities.
- For each pair of conditional row probabilities $P_{j|i}$ and $P_{j|i'}$ that we are interested in comparing, we compute the Z statistic for two proportions,

$$Z = \frac{\hat{P}_{j|i} - \hat{P}_{j|i'}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}}, \text{ where } \hat{P}_{j|i} = \frac{n_{ij}}{n_i}, \hat{P}_{j|i'} = \frac{n_{i'j}}{n_{i'}}, \bar{p} = \frac{n_i \hat{P}_{j|i} + n_{i'} \hat{P}_{j|i'}}{n_i + n_{i'}}$$

and suppose there are k number of comparisons of interest, then the Z statistics upon which we base our multiple comparison procedure is, $Q^* = \max_i |Z_i|$

Steps in the Multiple Comparison Permutation Test for Proportions

1. Identify the comparisons of conditional row probabilities that we wish to make and compute the corresponding Z statistics for two proportions, which we denote $Z_{1\,obs}$, $Z_{2\,obs}$, ... $Z_{k\,obs}$.
2. Permute the observations as we did with the permutation X^2 statistic. For each permutation, find the Z statistics, $Z_1, Z_2, \ldots Z_k$, that correspond to the row proportions we wish to compare. Compute $Q^* = \max_i |Z_i|$.
3. Obtain the permutation distribution of Q^*. Let $q^*(\alpha)$ denote the $100\alpha\%$ point of the distribution of Q^*. Declare the two proportions involved in the computation of $Z_{i\,obs}$ to be statistically significantly different at level of significance α if
$$|Z_{i\,obs}| \geq q^*(\alpha)$$
4. The p-value for the i^{th} comparison is the fraction of the permutation distribution of Q greater than or equal to $|Z_{i\,obs}|$.

5.5 Fisher's Exact Test for a 2×2 Contingency Table

Fisher's Exact Test:

- special case of the permutation test applied to 2×2 contingency table.
- Compute the hypergeometric probabilities of X's possible assignments

5.5.1 Probability Distribution Under the Null Hypothesis

- In the 2×2 table, knowledge of one cell entry along with knowledge of the row and column totals completely determines the entries in the table.

pdf 189-190 for further explanation

5.6 Contingency Tables with Ordered Categories

- We consider alternatives to χ^2 statistic when there is an ordering among the categories

5.6.1 Singly Ordered Tables

- run wilcoxon rank-sum test or Kruskal-Wallis test.

- The large difference between the p-value for the Kruskal-Wallis statistic and the χ^2 statistic is due to the type of alternatives that the two statistics are designed to detect.
- χ^2 statistic is designed to detect "any" association, whereas the Kruskal-Wallis statistic is designed to detect an ordering effect.

5.6.2 Doubly Ordered Tables

- apply Jonckheere-Terpstra statistic with ties.

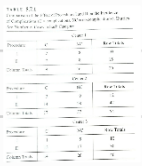
5.7 Mantel-Haenszel Test

- the responses are independent from a stratum to another.

5.7.1 Hypotheses to Be Tested

H_0: $P_{j|i}(k) = P_{j|i'}(k)$, $k = 1, 2, \ldots S$, the number of strata

H_a: $P_{j|i}(k) \geq P_{j|i'}$ or $P_{j|i}(k) \leq P_{j|i'}$, for at least one stratum.



example

There are three ways of stating the hypotheses

1. Independence $\Rightarrow H_0 : p_{11}(k) = p_{1\cdot}(k)p_{\cdot 1}(k), k = 1, \ldots, s$
2. Homogeneity $\Rightarrow H_0 : p_{1|1}(k) = p_{1|2}(k), k = 1, 2, \ldots, s.$
3. Odds ratio $\Rightarrow H_0 : \theta(k) = \frac{p_{11}(k)p_{22}(k)}{p_{12}(k)p_{21}(k)} = 1, k = 1, 2, \ldots, s.$

$$H_0: \theta(k) = 1, \quad k = 1, 2, \ldots, s$$
$$H_a: \theta(k) \geq 1, \quad k = 1, 2, \ldots, s \quad \text{or}$$
$$\theta(k) \leq 1, \quad k = 1, 2, \ldots, s$$

5.7.2 Testing Hypotheses for Stratified Contigency Tables

the number of possible permutations

$\prod_{k=1}^{s} \binom{N_k}{r_{1k}}$, N_k = # of observations in k^{th} stratum, r_{ik} = # of observations in i^{th} row in k^{th} stratum

C_{jk} = # of observations in j^{th} column in k^{th} stratum

Expected value of X_k

$$E(X_k) = \frac{r_{1k} c_{1k}}{N_k}$$

Expected Variance of X_k

$$Var(X_k) = \frac{r_{1k} r_{2k} c_{1k} c_{2k}}{N_k^2 (N_k - 1)}$$

Mantel-Haenszel Statistic

$$MH = \frac{\left(\sum_{k=1}^{s} [X_k - E(X_k)] \right)^2}{\sum_{k=1}^{s} Var(X_k)}$$

* MH is a measure of how much the total number of observations in row 1 and column 1 deviates from what one would expect under the assumption of independence.

* see example 5.7.1 at pdf 198

5.7.3 Estimation of the Odds Ratio

Mantel-Haenszel Estimate of the common odds ratio

$$\hat{\theta} = \frac{A}{B}$$,

$$A = \sum_{k=1}^{s} \frac{n_{11k} n_{22k}}{N_k}$$

$$B = \sum_{k=1}^{s} \frac{n_{12k} n_{21k}}{N_k}$$

} n_{ijk} denote the observation in row i and column j of stratum k.

- The variable $\log(\hat{\theta})$ has an approximate normal distribution with a mean of $\log(\theta)$ and a variance given by ,

$$\text{Var}(\log\hat{\theta}) = \frac{\sum_{k=1}^{S}(n_{11k}+n_{22k})(n_{11k}n_{22k})/N_k^2}{2A^2} + \cdots$$

$$+ \frac{\sum_{k=1}^{S}(n_{11k}+n_{22k})(n_{12k}n_{21k}) + (n_{12k}+n_{21k})(n_{11k}n_{22k})/N_k^2}{2AB} + \cdots$$

$$+ \frac{\sum_{k=1}^{S}(n_{12k}+n_{21k})(n_{12k}n_{21k})/N_k^2}{2B^2}$$

The normal approximation can be used to make a confidence interval for $\log(\theta)$ and then exponentiated to obtain a confidence interval for θ.

5.8 McNemar's Test

- Used for binary responses
- We are willing to know if a treatment affects the binary responses of 2 groups.

5.8.1 Notation and Hypotheses

- We are interested in testing $P_{AA} + P_{AB} = P_{AA} + P_{BA}$, or $H_0: P_{AB} = P_{BA}$.

5.8.2 Test Statistic and Permutation Distribution

- We consider 4 test statistics based on X_{AB} and X_{BA}

$T_1 = X_{AB}$

$T_2 = X_{AB} - X_{BA}$ $\quad T_2 = T_1 - X_{BA}$

$T_3 = \dfrac{X_{AB} - X_{BA}}{\sqrt{X_{AB} + X_{BA}}}$ $\quad T_3 = \dfrac{T_2}{\sqrt{X_{AB}+X_{BA}}}$

$T_4 = T_3^2$ McNemar's Test statistic using χ^2 approximation (1 d.f.)

The conditional distribution of T_1 will be obtained given that the number of individuals who switch is fixed. Now the number who switch is $n = X_{AB} + X_{BA}$. If n is fixed and if $P_{AB} = P_{BA}$, then among the individuals who switch, there is an equal chance that the switch is from A to B or B to A. If we assign a plus sign to individuals who are classified (A, B) and a minus sign to those who are classified (B, A), then X_{AB} is the number of plus signs. The outcome of whether an individual is (A, B) or (B, A) under the null hypotheses is equivalent to assigning pluses and minuses with probability $p = .5$ to the individuals who switch. The conditional distribution of X_{AB}, given n, is a binomial distribution with mean $.5n$ and variance $.25n$ under the null hypotheses. It follows that the test based on X_{AB} is just the sign test as discussed in Section 4.3.1.

$X_{AB} - X_{BA} = X_{AB} - (n - X_{AB}) = 2X_{AB} - n$, T_1 and T_2 are equivalent, and

$T_3 = \dfrac{X_{AB} - 0.5n}{\sqrt{0.25n}}$, has an approximate standard normal distribution under H_0.

* Using T_4 is equivalent to doing a two-tail test with T_3. It has an approximate χ^2 distribution with 1 d.f., and this case is called McNemar's Test.