Some Non-Parametric Methods

November 20, 2017

Sign Test

- The oldest of all nonparametric tests
- Used to test whether one random variable in a pair (X, Y) tends to be larger than the other.
- Level of Measure must be at least ordinal
- Sample $(X_1, Y_1), \ldots, (X_n, Y_n)$
- Each pair is **independent** of each other pair
- X and Y are not independent
- Paired t-test

Sign Test

Each pair is classified as

Classification	Condition
"+"	if $X_i < Y_i$
"-"	if $X_i > Y_i$
"0"	if $X_i = Y_i$

Sign Test

The null hypothesis is that the probability of "+" is the same as the probability of "-", i.e. neither variable tends to be larger than the other. Under the null hypothesis, the number of "+" pairs has a binomial distribution with parameters n and $1/2$.

Sign Test - Example

An item A is manufactured using a certain process. Item B serves the same function as A but is manufactured using a new process. The manufacturer wishes to determine whether B is preferred to A. Ten consumers are asked whether they prefer A or B. The results show that 8 consumers preferred B to A, resulting in a "+", 1 consumer preferred A, "-" and one showed no preference. Thus $n = 8 + 1 = 9$. For Y a binomial random variable, the p-value is

$$P(Y \geq 8|p = 1/2, n = 9) = P(Y = 8|p = 1/2, n = 9) + P(Y = 9|p = 1/2, n =$$

We can reject the null hypothesis and conclude that Item B tends to be preferred over A.

McNemar Test

- The data consists of n independent bivariate pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ where each X and Y is either a 0 or a 1.
- There are only four possible (X, Y) pairs, $(0,0), (0,1), (1,0)$ and $(1,1)$. The data can be summarized in a 2×2 contingency table.

a = num of (0,0) pairs	b = num of (0,1) pairs
c= num of (1,0) pairs	d = num of (1,1) pairs

- We can think of the pairs as before and after states. Thus (0,1) is the situation where a person started in state 0 and moved to state 1. The McNemar test tests for a difference in the probability of (0,1) and (1,0).

McNemar Test

To test the null hypothesis: $H_0 : P(0,1) = P(1,0)$, we use test statistic

$$T = \frac{(b-c)^2}{b+c}$$

which has an approximate chi-square distribution with one degree of freedom under the null hypothesis. We reject the null hypothesis if T is too big.

McNemar Test - Example

Prior to a natioinally televised debate between 2 presidential candidates, a random sample of 100 people stated their choice of candidates as 84 for the Democratic candidate and 16 for the Republican. After the debate the same 100 people were polled and 1/4 of the person who formerly favored the Democrat changed their mind while 1/4 of the people who formerly favored the Republican switched to the Democrat. The results can be summarized in the following table

	After	
	Democrat (0)	Republican (1)
Democrat Before (0)	63	21
Republication Before (1)	4	12

McNemar Example cont.

We want to test $H_0 : P(1,0) = P(0,1)$, that is the population voting alignment was not changed by the debate, that is the probability of switching did not depend on which candidate you favored before the debate. The test statistic:

$$T = \frac{(b-c)^2}{b+c} = \frac{(21-4)^2}{21+4} = 11.06$$

p-value:

$$P(\chi_1^2 > 11.56) = .0006739$$

Cox and Stuart Trend Test

Trend

A sequence of numbers $X_1, X_2, \ldots, X_n$ is said to have a *trend* if the later numbers in the sequence tend to be greater than earlier numbers (upward trend) or less than the earlier numbers (downward trend).

The Cox and Stuart trend test involves pairing the later numbers with the earlier numbers and then performing a sign test on the pairs thus formed. If there is a trend, one member of each pair will have a tendency to be higher or lower than the other. If there is no trend, and the observations are independent and identically distributed random variables there will be no tendency for one member of each pair to exceed the other. The null hypothesis is that there is no trend.

Cox and Stuart Example

Total annual precipitation is recorded for 19 years. This record is examined to see if the amount of precipitation is tending to increase or decrease. The null hypothesis is that there is no trend. The alternative is thus two-sided. The precipitation numbers in inches are: 45.25, 45.83, 41.77, 36.26, 45.37, 52.25, 35.37, 58.32, 41.05, 33.72, 45.73, 37.90, 41.72, 36.07, 49.83, 36.24, 39.90.

analysis

Since $n = 19$ is odd, omit the middle number 58.32. Pair up the remaining observations and assign a "+" if the latter observation is greater than the first, "-" if the later observation is less than the first and a tie if the two are equal.

Cox and Stuart Example cont

Pair	Sign	Pair	Sign
(45.25, 41.05)	-	(45.83, 33.72)	-
(41.77, 45.73)	+	(36.26, 37.90)	+
(45.37, 41.72)	-	(52.25, 36.07	-
(35.37, 49.83)	+	(57.16, 36.24)	-
(35.37, 39.90)	+		

analysis

There are no ties and 9 pairs so $n = 9$, the test statistic, T, the number of "+"'s is 4. Under the null hypothesis of no trend, T has a binomial distribution with parameters $n = 9$ and $p = 1/2$ so $E(T) = 9(1/2) = 4.5$ almost exactly what we got. There is no evidence to support a trend. The two sided p-value is actually 1.

Measures of Correlation

For random variables X and Y, the correlation between them is

$$\rho = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Correlation

The correlation measures the strength of the linear relationship between X and Y. We can show that $-1 \leq \rho \leq 1$. if $\rho = \pm 1$ there is a perfect linear relationship between X and Y.

Pearson's correlation coefficient

Suppose we have a random sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. The Pearson's correlation coefficient is an estimate of ρ given by

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

Pearson's correlation coefficient

- $-1 \leq r \leq 1$
- $r = \pm 1$ indicates a perfect linear relationship.
- If (X_i, Y_i) for $i = 1, \ldots, n$ are a random sample from a bivariate normal distribution with correlation parameter ρ, we can test $H_0 : \rho = 0$ using test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

  which has a t-distribution with $n-2$ degrees of freedom under H_0.
- If the data are not from a bivariate normal distribution the test statistic does not have a t-distribution and results for the test may be way off.

Spearman's Rank Correlation Coefficient

When data is not from a bivariate normal distribution or is at the ordinal level of measure, there are non-parametric estimates of correlation. **Spearman's rank correlation** is just Pearson's correlation coefficient computed on the ranks of X_i's and Y_i's. Assign ranks to the X's and Y's separately. In case of ties, assign midranks. Let $u_i = rank(X_i)$ and $v_i = rank(Y_i)$. If there are no ties then u_i's and v_i's are permutations of the integers $\{1, 2, \ldots, n\}$.

Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient is

$$r_S = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2 \sum_{i=1}^{n}(v_i - \bar{v})^2}}$$

If there are no ties, then this formula becomes

$$r_S = 1 - \frac{6\sum_{i=1}^{n}(u_i - v_i)^2}{n(n^2 - 1)}$$

Spearman's Rank Correlation Coefficient

- $-1 \leq r_S \leq 1$
- $r_s = \pm 1$ if there is a perfect monotone association between X and Y
- If X and Y are independent, $E(r_S) = 0$ and $Var(r_S) = \dfrac{1}{n-1}$.
- For large n ($n \geq 10$) and we can use the test statistic $z = r_s\sqrt{n-1}$ which tends to a standard normal distribution if X and Y are independent.

Spearman's Example

The data below is information collected by country on wine consumption and heart disease deaths:

Country	Alcohol from wine	u_i	Heart Disease Deaths	v_i	$u_i - v_i$
Australia	2.5	11	211	12.5	-1.5
Austria	3.9	15	167	6.5	8.5
Belgium	2.9	13.5	131	5	-8.5
Canada	2.4	10	191	9	1
Denmark	2.9	13.5	220	14	-0.5
⋮	⋮	⋮	⋮	⋮	⋮
W Germany	2.7	12	172	8	-4

Spearman's Example

Estimate

From this data we compute $r_S = -0.8289$ indicating that countries with higher wine consumption have lower number of heart disease deaths.

Test

To test whether the variables are independent, the test statistic $z = -0.8289\sqrt{18} = -3.517$ resulted in a p-value of 0.000218. The null hypothesis of independence is thus rejected. We can conclude that wine consumption and heart deaths are not independent.

Kendall's Rank Correlation Coefficient

Concordant and Discordant Pairs

Suppose we have two bivariate observations (X_i, Y_i) and (X_j, Y_j). The pairs are **concordant** if $(X_i - X_j)(Y_i - Y_j) > 0$ and **discordant** if $(X_i - X_j)(Y_i - Y_j) < 0$.

Let π_C be the probability of a concordant pair and π_D be the probability of a discordant pair. Assuming the probability of a tie is 0, $\pi_C + \pi_D = 1$. Kendall's rank correlation coefficient is

$$\tau = \pi_C - \pi_D$$

It is easy to see that $-1 \leq \tau \leq 1$.

Kendall's Rank Correlation Coefficient

To estimate τ, suppose we have a sample of size n bivariate pairs. Define $N = \binom{n}{2}$ to be the number of pairwise comparisons. Let N_c be the number of concordant pairs in the sample and N_d be the number of discordant pairs in the sample. ($N_d = N - N_c$). Estimate

$$\hat{\pi}_C = \frac{N_c}{N} \quad \hat{\pi}_D = \frac{N_d}{N}$$

and

$$\hat{\tau} = \hat{\pi}_C - \hat{\pi}_D$$