Stat 415 Review

Chap 9.3 Analysis of Variance (ANOVA)

Analysis of Variance
for comparing all 5 brands

Source	DF	SS	MS	F	P
Brand	4	1174.8	293.7	7.95	0.000
Error	45	1661.7	36.9		
Total	49	2836.5			

1. Source: the source of the variation in the data. The possible choices for a one-factor study, are Factor, Error, and Total. The factor is the characteristic that defines the populations being compared. In the tire study, the factor is the brand of tire.
2. DF: the degrees of freedom in the source.
3. SS: the sum of squares due to the source.
4. MS: the mean sum of squares due to the source.
5. F: the test-statistic which follows a F distribution under H_0.
6. P: the P-value.

One-way Analysis of Variance

Source	DF	SS	MS	F	P
Factor	m-1	SS(Between)	MSB	MSB/MSE	
Error	n-m	SS(Error)	MSE		
Total	n-1	SS(Total)			

From F-distribution with m-1 numerator and n-m denominator d.f.

MSB = SS(Between)/(m-1)
MSE = SS(Error)/(n-m)

n-1 = (m-1) + (n-m)

SS(Total) = SS(Between) + SS(Error)

Chap 8.4 Sign Test and Wilcoxon Test

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random samples obtained from a population with unknown distributions. We want to test $H_0 : m = m_0$ v.s. $H_1 : m \neq m_0$.

Sign Test for One Population:

Let Y be the number of negative signs among $X_i - m_0$, $i = 1, \ldots, n$. $Y \sim Binomial(n, p)$.

- $H_0 : m = m_0$ v.s. $H_1 : m \neq m_0$ is equivalent to $H_0 : p = 0.5$ v.s. $H_1 : p \neq 0.5$.
- $H_0 : m = m_0$ v.s. $H_1 : m < m_0$ is equivalent to $H_0 : p = 0.5$ v.s. $H_1 : p > 0.5$.
- $H_0 : m = m_0$ v.s. $H_1 : m > m_0$ is equivalent to $H_0 : p = 0.5$ v.s. $H_1 : p < 0.5$.

Sign Test for Two Populations:

Suppose we have a paired sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let Z be the number of of negative differences among $X_i - Y_i$, $i = 1, \ldots, n$. $Z \sim Binomial(n, p)$, where $p = P(X < Y)$.

- $H_0 : m_X = m_Y$ v.s. $H_1 : m_X \neq m_Y$ is equivalent to $H_0 : p = 0.5$ v.s. $H_1 : p \neq 0.5$.
- $H_0 : m_X = m_Y$ v.s. $H_1 : m_X < m_Y$ is equivalent to $H_0 : p = 0.5$ v.s. $H_1 : p > 0.5$.
- $H_0 : m_X = m_Y$ v.s. $H_1 : m_X > m_Y$ is equivalent to $H_0 : p = 0.5$ v.s. $H_1 : p < 0.5$.

Wilcoxon Signed Rank Test for One Population:

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random samples from an unknown, symmetric distribution.

1. Compute $X_i - m_0$;
2. Let R_i be the rank of $|X_i - m_0|$;
3. Determine the sign of $X_i - m_0$: 1 if $X_i - m_0 > 0$, and -1 otherwise;
4. Define $W = \sum_i sign(X_i - m_0) \times R_i$, H_0 is supported if W is close to zero.
5. Test statistic: $Z = \frac{W}{\sqrt{n(n+1)(2n+1)/6}} \sim N(0, 1)$ approximately.
6. $H_1 : m \neq m_0$, critical region is $|Z| \geq Z_{\alpha/2}$;
   $H_1 : m < m_0$, critical region is $Z < -Z_\alpha$;
   $H_1 : m > m_0$, critical region is $Z > Z_\alpha$.

If $X_i = m_0$ for some i, that observation is deleted, and the test is performed with a reduced sample size.

If two or more observations are equal, each observation is assigned the average of the corresponding ranks.

Wilcoxon Signed Rank Test for Two Population:

Let $X_1, X_2, \ldots, X_{n_1}$ be i.i.d. samples from an unknown distribution with median m_X.

Let $Y_1, Y_2, \ldots, Y_{n_2}$ be i.i.d. samples from an unknown distribution with median m_Y.

$H_0 : m_X = m_Y$.

1. Assign ranks to the combined samples $(x_1, x_2, \ldots, x_{n_1}, y_1, y_2, \ldots, y_{n_2})$.
2. Let W be the sum of ranks from $(y_1, y_2, \ldots, y_{n_2})$.

   Under H_0, $W \sim N(\frac{n_2(n_1+n_2+1)}{12}, \frac{n_1 n_2(n_1+n_2+1)}{12})$.

   Test statistics: $Z = \frac{W - n_2(n_1+n_2+1)/12}{\sqrt{n_1 n_2(n_1+n_2+1)/12}} \sim N(0,1)$.
3. $H_1 : m \neq m_0$, critical region is $|Z| \geq Z_{\alpha/2}$;
   $H_1 : m_X < m_Y$, critical region is $Z > Z_\alpha$;
   $H_1 : m_X > m_Y$, critical region is $Z < -Z_\alpha$.

Chap 9.1 Chi-Square Goodness-of-Fit Test

$X = (X_1, X_2, \cdots, X_K) \sim$ Multinomial(n, p) where $p = (p_1, p_2, \cdots, p_K)$ are unknown.

H_0: $p_1 = p_{10}, p_2 = p_{20}, \cdots, p_K = p_{K0}$ v.s. H_1: at least one equation does not hold.

Under H_0, $Q = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(K-1)}$.

Critical region: $Q \geq \chi^2_\alpha(K-1)$.

Need $E_j < 5$ for no more than 20% of the cells in the table and none of $E_j < 1$.

If the above condition is not satisfied, we often combine cells until all E_j are large enough.

Sometimes you are interested in testing whether a data set fits a probability model with d parameters left unspecified.

1. Estimate the d parameters e.g. using the maximum likelihood method.
2. Calculate the chi-square statistic Q using the obtained estimates.
3. Critical region: $Q \geq \chi^2_\alpha(K-1-d)$.

Chap 9.2 Contigency Tables

A contingency table is a type of table that displays the distribution of the combination of multiple categorical variables. For instance, y_{ij} is the frequency of obsering ith category of the first variable and jth category of the second variable.

Example: University admissions officers were concerned that males and females students applying for the four different schools (business, engineering, liberal arts, and science) at the university. They collected the following data on the acceptance of 1200 males and 800 females who applied to the university.

Accepted	Business School	Engineering School	Liberal Arts School	School of Science	Total
Male	240	480	120	360	1200
Female	240	80	320	160	800
Total	480	560	440	520	2000

Homogeneity Test: Test whether two multinomial distributions are equal.

$$H_0 : p_{11} = p_{21}, p_{12} = p_{22} \ldots, p_{1k} = p_{2k}$$

$E_{ij} = y_{i.} \times \hat{p}_{.j}$ where

- $y_{i.}$ is the sample size of the ith population,
- $\hat{p}_{.j}$ is the pooled sanple proportion for the jth category.

Under H_0, $Q = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(h-1)(k-1)}$.

Critical region: $Q \geq \chi^2_{\alpha}(h-1)(k-1)$.

Independence Test: Whether gender and major are independent?

$$H_0 : p_{ij} = p_{i.} \times p_{.j} \text{ for all i and j.}$$

$E_{ij} = n \times \hat{p}_{i.} \times \hat{p}_{.j}$ where

- $\hat{p}_{i.}$ is the pooled sanple proportion for the ith category of the row variable,
- $\hat{p}_{.j}$ is the pooled sanple proportion for the jth category of the column variable.

$Q = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(h-1)(k-1)}$.

Critical region: $Q \geq \chi^2_{\alpha}(h-1)(k-1)$.