IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Wanyue Xiao
Date: October 6, 2020
Chapter Number: #7
Title of Chapter: Associations between Variables

Covariance and Association

Covariance is a statistical tool that is used to determine the relationship between the movement of two variable. When two variables tend to move together, they are seen as having a positive covariance; when they move inversely, the covariance is negative.

- The idea of a "standard normal" distribution indicates that the mean of a distribution is 0 and its standard deviation is 1.
- Small letter *r to refer to a correlation estimated from a sample* and the word "*rho" to refer to the population value of the correlation coefficient*.
- R studio: cor() command to generate a correlation matrix whose diagonal contains all values of 1, indicating the perfect correlation that always exists between a variable and itself

Categorical Association
Example:

TABLE 7.2. "Null Hypothesis" Contingency Table—No Association

	Down	Up	Row totals
Jelly	15	15	30
Butter	35	35	70
Column totals	50	50	100

But once we have calculated the marginal totals from raw data, ***almost everything about the original contingency table is now fixed.*** In fact, for a 2 × 2 table like this one, once you have calculated the expected frequencies from the marginal totals, ***only one cell in the original table is free to vary—one degree of freedom!***

- More generally, for tables larger than 2 × 2***, the number of degrees of freedom is (rows-1)*(cols-1).***

So, keeping in mind that the expected value of jelly-down is 15, we can conclude that the minimum jelly-down is 0 and the maximum jelly-down is 30, given that we must maintain our exact marginal totals that were used to create the expected value table. Just to make it clear, here are the minimum (Table 7.3) and maximum (Table 7.4) tables.

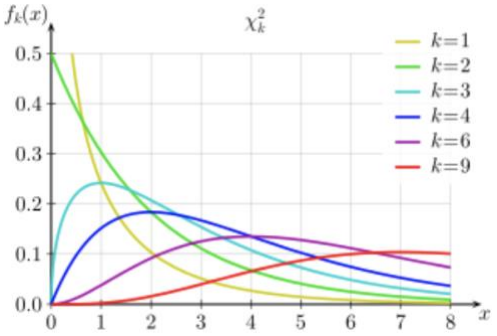
Chi-square Test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.
Calculation: Calculate the *square of the difference between the actual value and the expected value for each cell*. Then, so as to put each cell on a similar footing, let's *divide the squared difference by the expected value*. So*, for each cell we have ((actual - expected) ^2)/expected. Sum the results that we get for each cell* and we have a nice measure of how far any sampled contingency table varies from the expected values table.

- A **very small chi square test statistic** means that your observed data fits your expected data extremely well. In other words, there is a relationship.

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

- A **very large chi square test statistic** means that the data does not fit very well. In other words, there isn't a relationship.
- **R studio: chisq.test(),** performs the standard null hypothesis test.
  The parameter "correct=FALSE" suppresses the so-called Yates correction, which would have made the chi-square test more conservative for small samples. *Only use correct=TRUE if any one of the cells in the contingency table has fewer than five observations.*
- **Bayes Factor: contingencyTableBF():** Bayesian test of contingency tables that produces a Bayes factor and that can optionally produce posterior distributions for the frequencies (or proportions) in the cells of the contingency table.

- **chi-square distribution**
  In probability theory and statistics, the chi-square distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.



Formula for Pearson's Correlation

Formula: $$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Pearson's product–moment correlation (PPMC), aka *r*, which is a "moment" (or "mean") of a set of products. If the *r* ranges from –1 up to 1, signifying that *r* itself is on a kind of standardized scale.

Formula Interpretation: The numerator contains the sum of the cross-products of the deviations from the respective means of the two variables. The denominator contains separate calculations for the sum of squares for each of the two variables. These sums of squares are then multiplied together and then the square root of that product is taken. For all three of the summation symbols, we are summing across all the observations in the sample.

Four Assumption:

1. Your two variables should be measured at the **interval** or **ratio level**.
2. There is a **linear relationship** between your two variables.
3. There should be **no significant outliers**.
4. The variables should be **approximately normally distributed**.

Exercise Review

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

3.

```
> cor.test(rock$area, rock$perm)

        Pearson's product-moment correlation

data:  rock$area and rock$perm
t = -2.9305, df = 46, p-value = 0.005254
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6118206 -0.1267915
sample estimates:
      cor
-0.396637
```

- **t** is the **t-test statistic** value (t = -2.9305).
- **df** is the degrees of freedom (df= 46),
- **p-value** is the significance level of the **t-test** (p-value = 0.005254), which is lower than 0.05. Then we can reject the null hypothesis.
- **conf.int** is the **confidence interval** of the correlation coefficient at 95% (conf.int = [--0.6118206, -0.1267915]); if we repeated this sampling process many times and each time constructed a confidence interval around the calculated value of *r*, ***about 95% of those constructed intervals would contain the true population value,*** rho. In reporting this confidence interval in a journal article, we would simply say that the 95% confidence interval for rho ranged from –0.61 to -0.13.
- **sample estimates** is the correlation coefficient (cor.coeff = -0.396637).

4.

```
Iterations = 1:10000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean             SD       Naive SE Time-series SE
     -0.342792       0.135615       0.001356       0.001537

2. Quantiles for each variable:

    2.5%      25%      50%      75%    97.5%
-0.61100 -0.43258 -0.34181 -0.25033 -0.08026

Bayes factor analysis
--------------
[1] rhoNot0 : 8.072781 ±0%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

Bayes factor shows the odds in favor of the alternative hypothesis. The odd here is 8:1 which is higher than 3:1 in favor of the alternative hypothesis. The 95% HDI ranges from –0.61 up to -0.08.

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

8.

```
> chisq.test(UCBAdmissions[ , ,1], correct=FALSE)

        Pearson's Chi-squared test

data:  UCBAdmissions[, , 1]
X-squared = 17.248, df = 1, p-value = 3.28e-05
```

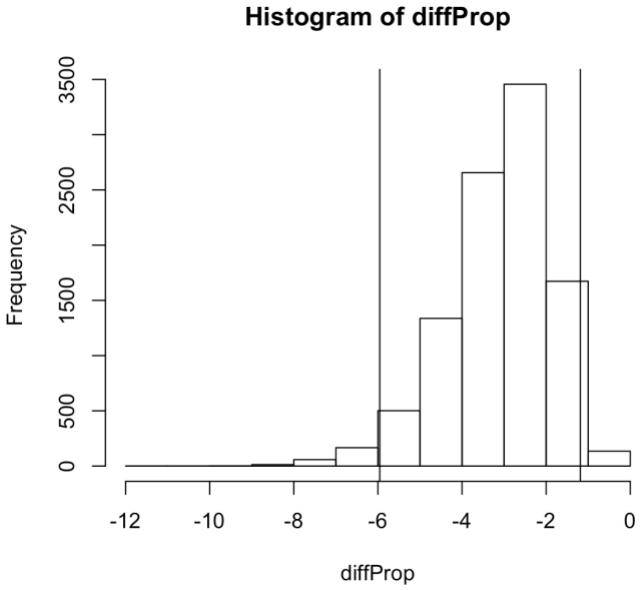
9.

```
Bayes factor analysis
--------------
[1] Non-indep. (a=1) : 1111.64 ±0%

Against denominator:
  Null, independence, a = 1
---
Bayes factor type: BFcontingencyTable, poisson
```

Bayes factor shows the odds in favor of the alternative hypothesis. The odd here is 1111:1 which is higher than 500:1 in favor of the alternative hypothesis. Therefore, the two variables are associated with each other.

10.



R Code Fragment and Explanation

cor() computes the **correlation coefficient**
cor.test() test for association/correlation between paired samples. It returns both the **correlation coefficient** and the **significance level**(or p-value) of the correlation .
cor.test(wood, heat)

- **x, y** numeric vectors of data values. x and y must have the same length.

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

- **alternative** indicates the alternative hypothesis and must be one of "two.sided", "greater" or "less". You can specify just the initial letter. "greater" corresponds to positive association, "less" to negative association.
- **method** a character string indicating which correlation coefficient is to be used for the test. One of "pearson", "kendall", or "spearman", can be abbreviated.

Example:

```
        Pearson's product-moment correlation
data: wood and heat
t = -0.2951, df = 22, p-value = 0.7707
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.4546764 0.3494514
sample estimates:
cor
-0.06279774
```

Interpretation:

- **t** is the **t-test statistic** value (t = -0.2951).
  - Remember that **any *t*-value with an absolute value less than about 2 is unlikely to be significant**
- **df** is the degrees of freedom (df= 22),
- **p-value** is the significance level of the **t-test** (p-value = 0.7707).
  - One way of thinking about this *p*-value is to say that ***there is a 0.7707 chance of observing an absolute value of t this high or higher under the assumption that the population value of rho*** = 0.
- **conf.int** is the **confidence interval** of the correlation coefficient at 95% (conf.int = [-0.4546764, 0.3494514]);
  - if we repeated this sampling process many times and each time constructed a confidence interval around the calculated value of *r*, ***about 95% of those constructed intervals would contain the true population value,*** rho. In reporting this confidence interval in a journal article, we would simply say that the 95% confidence interval for rho ranged from –0.45 to 0.35.
  - Importantly, ***the confidence interval straddles 0***, a result that concurs with the results of the significance test.
- **sample estimates** is the correlation coefficient (Cor.coeff = -0.06279774).

Question for Class

1. How to interpretant this result correctly? What's the meaning of non-indep. and (a=1)?

```
Bayes factor analysis
--------------
[1] Non-indep. (a=1) : 1111.64 ±0%

Against denominator:
  Null, independence, a = 1
---
Bayes factor type: BFcontingencyTable, poisson
```