Chapter 16. Data Analysis: Frequency Distribution, Hypothesis Testing, and Cross-Tabulation

Frequency Distribution

; Frequency distribution

; A mathematical distribution with the objective of obtaining a count of the number of responses associated with different values of one variable and to express these counts in percentage terms.

; frequency distribution for a variable produces a table of frequency counts, percentages, and cumulative percentages for all the for all the values associated with that variable.

; Conducting Frequency Analysis

; ① Calculating the frequency for each value of the variable

; ② Calculate the percentage and cumulative percentage for each value, adjusting for any missing values

; ③ Plot the frequency histogram

; ④ Calculate the descriptive statistics, measures of location, and variability

; A frequency distribution helps determine the extent of illegitimate responses. (예를 들어 0과 8 값이 나오면 illegitimate response임을 알 수 있다) The presence of outliers can also be detected.

Statistics Associated with Frequency Distribution

Measures of Location

; Measures of location

; Statistics that describe locations within a data set. Measures of central tendency describe the center of the distribution.

; 아래의 Mean, median, mode는 이 central tendency를 서술하는 각기 다른 방법이라 할 수 있다.

Mean

; Mean (평균)

; The average; that value obtained by summing all elements in a set and dividing by the number of elements.

$$\overline{X} = \sum_{i=1}^{n} X_i / n$$

X_i = observed values of the variable X

n = number of observations (sample size)

Mode

; Mode (최빈값)

; A measure of central tendency given as the value the occurs the most in a sample distribution

; Mode is the value that occurs most frequently.

Median

; Median (중위값)

; A measure of central tendency given as the value above which half of the values fall and below which half of the values fall.

; The middle value when the data are arranged in ascending or descending rank order.

; 만약 distribution이 asymmetric하다면, 어떤 measure를 사용해야 하겠는가?

; If the variable is measured on a nominal scale => mode

; If the variable is measured on an ordinal scale => median

; If the variable is measured on an interval / ratio scale => mean is best, median is alternative. 하지만 mean을 사용하게 되면 매우 작거나 큰 outlier값에 예민하게 되므로 주의한다. outlier값이 있으면 mean과 median을 같이 사용하는 것이 낫다.

Measure of Variability

; Measure of variability

; Statistics that indicate the distribution's dispersion

; 여기에는 range and variance or standard deviation 이 있다.

Range

; Range

; The difference between the smallest and largest values of a distribution

; $Range = X_{largest} - X_{smallest}$

Variance and Standard Deviation

; Variance

; The mean squared deviation of all the values from the mean.

; Standard deviation

; The square root of the variance.

; The standard deviation of sample s_x

$$s_x = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}}$$

; n-1이 사용된 이유는 sample이 모집단에서 뽑혀 나왔기 때문.

Introduction to Hypothesis Testing

; examples of hypotheses generated in marketing research: [혹 나올수도 있으니 알아둘 것]

; The average number of stores shopped for groceries is 3.0 per household

; The department store is being patronized by more than 10% of the households

; The heavy and light users of a brand differ in terms of psychographic characteristics

; One hotel has a more upscale image than its close competitor.

; Familiarity with a restaurant results in greater preference for that restaurant.

A General Procedure for hypothesis testing

(1) Formulating the hypothesis

; 우선 Ho와 Ha를 공식화 하여야 한다. Ho는 항상 실험되는 가정이다. Hypothesis는 항상 population parameter의 값으로 서술되어야 하고(예를 들면 μ, σ, π), 표본 통계 값으로 서술되면 안된다(예를 들면 $\overline{X}$)

; 보통 결과는 Ho를 기각하고 Ha를 채택하거나, Ho를 채택하고 Ha를 기각하거나 둘 중 하나인데 Ho를 기각하지 않는다고 하여서 그 결과가 타당한(valid) 것이라고 받아들여서는 안된다. 전통적인 가정 실험에 따르면, Ho가 참이라는 것을 결정하는 방법은 없다.

; Null hypothesis [귀무가설]

; A statement suggesting no expected difference or effect. If the null hypothesis is not rejected, no changes will be made.

; Alternative hypothesis [대립가설]

; A statement suggesting some difference or effect is expected. Accepting the alternative hypothesis will lead to changes in opinions or actions.

; $Ho : \pi \leq 0.40$과 같은 실험은 one-tailed test인데 왜냐하면 대립가설이 지향성을 가지고(directionally) 표현되기 때문이다. 반대로 $Ho : \pi = 0.40$인 경우는 two-tailed test이다.
; 전통적인 MR에서 one-tailed test가 보다 많이 쓰인다.

; One-tailed test
; A test of the null hypothesis where the alternative hypothesis is expressed directionally
; 예) $Ho : \pi \leq 0.40$

; Two-tailed test
; A test of the null hypothesis where the alternative hypothesis is not expressed directionally.
; 예) $Ho : \pi = 0.40$

(2) Select appropriate test
; 즉 population에 대한 inference를 뽑아내는 것이라 생각하면 된다.
; Test statistics
; A measure of how close the sample has come to the null hypothesis. It often follows a well-know distribution, such as the normal, t, or chi-square distribution.

(3) Choose level of significance,
Type-I error
; Type-I error
; Also known as alpha error, it occurs when the sample results lead to the rejection of a null hypothesis that is in fact true.
; 즉 95% 신뢰구간 내에서 측정을 했고 그 결과 유의성이 있다 판단되어 귀무가설을 기각했는데, 사실 그 5%의 확률에 걸린 재수없는 경우이다.
; Level of significance
; The probability of making a type-I error.
; Level of significance(a)% = 100% - confidence level%

Type-II error
; Type-II error
; Also known as beta error, occurs when the sample results lead to nonrejection of a null hypothesis that is in fact false.
; Type-I과는 반대로 유의성이 없다 판단되었는데 사실은 유의성이 있던 경우.
; a와는 달리 b의 크기는 모집단 변수의 실제 값과 관계된다.
; Power of a statistical test
; Complement (1-b) of the probability of a type-II error

Power of a Test
; Power of a test
; The probability of rejecting the null hypothesis when it is in fact false and should be rejected.
; 즉 사실 reject되어야 하는 귀무가설을 기각할 확률 (즉 제대로 짚을 확률)
; 보통 1-b이다. 비록 b가 알려져 있지 않더라도 a와 관계된다. a를 매우 작은 값(0.001)로 취하면 과도하게 높은 b 오차를 낳게 된다. 따라서 2종류의 오류의 균형을 맞추는 것이 중요하다. 보통은 a=0.05로 잡게 된다. 또한 주어진 a의 레벨에서 샘플 크기를 늘리면 b를 줄이게 되며, 따라서 power of the test의 정도를 증가시키게 된다. (보다 정확해진다는 의미)

(4) Collect Data and Calculate Test Statistics
; 주어진 샘플 확률(p) 혹은 샘플 크기(n)를 가지고 모집단의 z-value를 구한다.

(5) Determining the Probability (critical value)

(6)(7) Comparing the probability(critical value) and making the decision
a) Determine probability associated with test statistic (TScal) : Compare with level of significance, a
if probability of TScal 〈 significance level(a), then reject Ho
b) Determine critical value of test statistic (TScr) : Determine if TScr falls into (Non) Rejection Region
if TScal 〉 TScr, then reject Ho
; 하나는 확률(probability)로 계산하고 또다른 하나는 critical value를 이용하여 계산한다. 둘 중 어느 방법을 쓰든지 관계는 없지만 방향을 조심하도록 한다.

; Critical value
; The value of the test statistic that divides the rejection and nonrejection regions. If the calculated value of the test statistic is greater than the critical value of the test statistics, the null hypothesis is rejected.

(8) Marketing research conclusion

Cross-Tabulations
; Cross-tabulation
; A statistical technique that describes two or more variables simultaneously and results in tables that reflect the joint distribution of two or more variables that have a limited number of categories or distinct values.
; 즉 Cross-tab을 하기 위해서는 categorize 작업이 필요함을 알 수 있다.

; Contingency tables
; Cross-tabulation tables; contain a cell for every combination of categories of the two variables.

; Cross-tab은 MR에서 빈번히 쓰이는데 그 이유는 아래와 같다:
; 1) Cross-tabulation analysis and results can be easily interpreted and understood by managers who are not statically oriented.
; 2) The clarity of interpretation provides a stronger link between research results and managerial action
; 3) Cross-tabulation analysis is simple to conduct and more appealing to less-sophisticated researcher.

; Bivariate cross-tabulation
; Cross-tabulation with two variables

General Comments on Cross-tabulation
; 3개 이상의 변수가 cross-tab으로 들어가면 보기도 어렵고 분석도 힘들다. 일반적으로 각 셀에 적어도 5개의 expected observation이 있어야 신뢰성있는 chi-square test를 사용할 수 있다. 또한 몇가지 변수가 있을 경우 cross-tab은 변수간의 관계를 찾아내는 것에 있어서 비효율적이다.

Statistics Associated with Cross-tabulation

Chi-Square

; 두 변수간에 연관성이 있는지를 조사할 때 쓰인다.

; Chi-square statistic을 사용할 경우 Ho=There is no association between the variables가 된다

; Chi-square statistic

; The statistic used to test the statistical significance of the observed association in a cross-tabulation. It assists in determining whether a systematic association exists between the two variables.

; Chi-square distribution

; A skewed distribution whose shape depends solely on the number of degrees of freedom. As the number of degrees of freedom increases, the chi-square distribution becomes more symmetrical.

$f_e = \frac{n_r n_c}{n}$ (f_e = expected cell frequency, f_0 = actual observed frequency)

n_r = total number in the row

n_c = total number in the column

n = total sample size

$$\chi^2 = \sum_{allcells} \frac{(f_0 - f_e)^2}{f_e}$$

$df = (r-1)(c-1)$ (df = degree of freedom)

; 이렇게 해서 나온 $\chi^2_{calc} \rangle \chi^2_{crit}$ 이라면, reject Ho이다.

; chi-square distribution은 skewed distribution 이다. shape는 df의 값에 의존한다. 즉 df값이 커질수록 분포 형태가 보다 대칭적이 되어간다.

; chi-square statistic은 goodness-of-fit test 에도 사용될 수 있다. These tests are conducted by calculating the significance of sample deviations from assumed theoretical(expected) distributions and can be performed of the chi-square statistic and the determination of its significance is the same as illustrated earlier.

; chi-square statistic은 data의 계수(count)에만 쓰일 수 있다. percent의 형태에서는 먼저 절대적인 counts or numbers로 바꾸어 주어야 써먹을 수 있다. 또한 chi-square test의 기저에 있는 가정은 observation이 independently하게 뽑혀 나온다는 것이다. 응답자들이 서로의 응답에 영향을 미치지 않는다는 것을 의미한다. 또한 일반적으로 chi-square analysis는 각 셀의 값이 5보다 작은 경우 사용하면 안된다. 낮은 expected frequency는 type-I error를 유발할 가능성이 있따.

Phi Coefficient

; Phi coefficient

; A measure of the strength of association in the special case of a table with two rows and two columns

$\Phi = \sqrt{\frac{x^2}{n}}$

; phi=0이면 no association임을 의미한다. 만약 변수들이 perfectly associated되어 있다면 phi=1이 된다.

Contingency coefficient

; Contingency coefficient

; A measure of the strength of association in a table of any size.

$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$

; 결과 범위는 0~1인데, c=0 이면 연관성 없음을 의미한다. 반면 1은 얻어질 수 없는 값에 속한다. 오히려 C의 값은 table의 size에 의존적이다.

Cramer's V

; Cramer's V

; A measure of the strength of association used in tables larger than 2 × 2

$V = \sqrt{\frac{\Phi^2}{\min(r-1),(c-1)}}$, $V = \sqrt{\frac{X^2/2}{\min(r-1),(c-1)}}$

; 즉 이는 phi를 개선해서 어느 사이즈에도 적용될 수 있도록 만든 버전이다. 결과 V는 0~1을 가지게 된다. 높은 V값은 높은 연관성이 있음을 의미하지만, 어떻게 변수들이 연관되어 있는가를 나타내지는 않는다.

Cross-tabulation in practice

① Construct the cross-tabulation table

② Test the null hypothesis that there is no association between the variables using the chi-square statistics

③ If you fail to reject the null hypothesis, there is no relationship

④ If Ho is rejected, determine the strength of the association using an appropriate statistic (phi coefficient, contingency coefficient, or Cramer's V)

⑤ If Ho is rejected, interpret the pattern of the relationship by computing the percentages in the direction of the independent variable, across the dependent variable. Draw marketing conclusions.

[Discussion Problems]

1. In each of the following situations, indicate the statistical analysis you would conduct and the appropriate test or test statistic that should be used.

a. Respondents in a survey of 1,000 household were classified as heavy, medium, light, or nonusers of ice cream. They were also classified as being in high-, medium-, or low-income categories. Is the consumption of ice cream related to income level?

; Data가 없기는 하지만, 이는 두 변수 사이의 연관성을 측정하는, χ^2 test를 이용하는 문제이다.

Null Hypothesis(Ho) : "Independence" = No association between Income and Consumption.
Alternative (Ha) : "Interdependence" = There is dependence relationship between income and consumption.

a = 0.05
df = (r-1)(c-1) = 6 {r과 c는 row/column수}
reject Ho if $P(\chi^2_{calc})$ 〈 a = 0.5
reject Ho if $\chi^2_{calc} > \chi^2_{critical}$

b. In a survey using a representative sample of 2,000 households from the Synovate consumer panel, the respondents were asked whether or not they preferred to shop at Sears. The sample was divided into small and large households based on a median split of the household size. Does preference for shopping in Sears vary by household size?

; 역시 Data가 없기는 하지만, two-sample t-test 문제이다. (한 sample이 다른 sample의 response에 영향을 주지 않음)

Ho: $u_1 = u_2$
Ha: $u_1 \neq u_2$
a = 0.05
df = n-2 = 2000 - 2 (1998)
reject Ho if P(t_{calc}) 〈 0.025 = a/2 (왜? two-tailed이기 때문)
reject Ho if t_{calc} 〉 1.96 or t_{calc} 〈 -1.96

2. The current advertising campaign for a major soft drink brand would be changed if less than 30 percent of the consumers like it.

a. Formulate the null and alternative hypotheses.

$Ho : P \geqq 0.30$
$Ha : P < 0.30$

b. Discuss the type-I and type-II errors that could occur in hypothesis testing.

Type I error : Error occurred by bad measurement. Error created by chance.
Type II error : Not rejecting Ho which has be rejected. It can occur by small sample size and bad measuring procedure. You will miss your opportunity because you didn't do anything.

3. A major department store chain is having an end-of-season sale on refrigerators. The number of refrigerators sold during this sale at a sample of 10 stores was 80 110 0 40 70 80 100 50 80 30.

a. Compute the mean, mode, and median. Which measure of central tendency is most appropriate in this case and why?

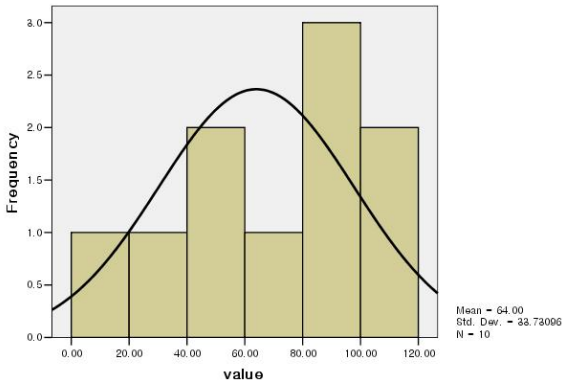
b. Compute the variance and the standard deviation.

[SPSS 수행 결과]

N	Valid	10
	Missing	0
Mean		64.0000
Median		75.0000
Mode		80.00
Std. Deviation		33.73096
Variance		1137.778
Minimum		.00
Maximum		110.00

Mean is more appropriate.

c. Construct a histogram, and discuss whether this variable is normally distributed.



It is not normally distributed.

5. A research project examining the impact of income on the consumption of gourmet foods was conducted. Each variable was classified into three levels of high, medium, and low. The following results were obtained.

		Income		
		Low	Medium	High
Consumption of Gourmet Foods	Low	25	15	10
	Medium	10	25	15
	High	15	10	25

a. Is the relationship between income and consumption of gourmet food significant?

Ho : There is no association between income and consumption. "Independent"

Ha : There is association between income and consumption. "Interdependence"

a = 0.05

df = (r-1)(c-1) = 4

χ^2_{crit} = 9.488

reject Ho if χ^2_{calc} 〉 9.488

χ^2_{calc} = (25-16.7)^2/16.7 + (15-16.7)^2/16.7 + (10-16.7)^2/16.7 + (10-16.7)^2/16.7 + (25-16.7)^2/16.7 + (15-16.7)^2/16.7 + (15-16.7)^2/16.7 + (10-16.7)^2/16.7 + (25-16.7)^2/16.7 = 20.95868 〉 9.488

So Ho will be rejected.

b. Is the relationship between income and consumption of gourmet food strong?

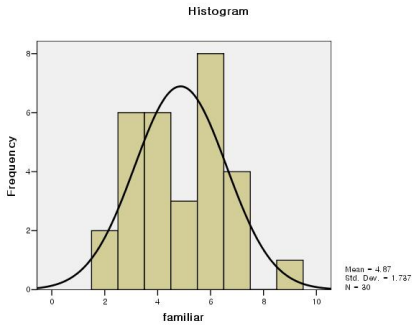
Yes. It has strong relationship.

c. What is the pattern of the relationship between income and consumption of gourmet food?

Higher the income, more gourmet food consumption.

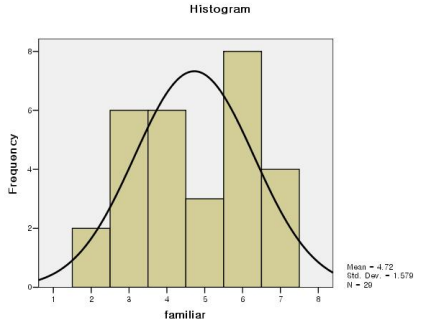
6. A pilot survey was conducted with 30 respondents to examine Internet usage for personal(non-professional) reasons. The following table contains the resulting data given each respondents' sex, familiarity with the Internet, Internet usage in hours per week, attitude toward Internet and toward technology, both measured on a seven-point scale, whether the respondents have done shopping or banking on the Internet.

a. Obtain the frequency distribution of familiarity with the Internet. Calculate the relevant statistics.



There is "Outlier" value 9. It is required to drop this value.

Result of dropping 9 value:



As you can see, the result is more relevant.

b. For the purpose of cross-tabulation, classify respondents as light or heavy users. Those reporting 5 hours or less usage should be classified as light users and the remaining as heavy users. Run a cross-tabulation of sex and Internet usage. Interpret the results. Is Internet usage related to one's sex?

Ho: There is no association between sex and internet usage.

Ha: There is association between sex and internet usage.

a = 0.05

df = (r-1)(c-1) = 1

χ^2_{crit} = 3.841

Reject Ho if $P(\chi^2_{calc}) < 0.05$

[SPSS 수행 결과 - Chi square test]

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.333(b)	1	.068		
Continuity Correction(a)	2.133	1	.144		
Likelihood Ratio	3.398	1	.065		
Fisher's Exact Test				.143	.072
Linear-by-Linear Association	3.222	1	.073		
N of Valid Cases	30				

$P(\chi^2_{calc}) = 0.068 > 0.05$

So Ho will not be rejected. (in 95% confidence interval)

하지만 90%의 신뢰구간 (a = 0.1) 에서는,

$P(\chi^2_{calc}) = 0.068 < 0.1$

Ho will be rejected.

가 된다.

Chapter 17. Data Analysis : Hypothesis Testing Related to Differences

Hypothesis testing related to differences

; Parametric tests

; Hypothesis testing procedures that assume the variables of interest are measured on at least an interval scale.

The t Distribution

; Parametric tests provide inferences for making statements about the means of parent populations.

; t-Test

; A univariate hypothesis test using that t distribution, which is used when the standard deviation is unknown and the sample size is small.

; t statistic

; A statistic that assumes the variable has a symmetric bell-shaped distribution and the mean is known (or assumed to be known), and the population variance is estimated from the sample.

; t distribution

; A symmetric, bell-shaped distribution that is useful for small sample (n〈30) testing.

; t distribution 은 normal distribution과 유사하나 꼬리쪽 영역이 보다 크고 중간쪽 영역이 작다. 이는 모집단 분산이 알려져있지 않고 표본 분산으로부터 유추되기 때문이다. 또한 n의 개수가 커지면 normal distribution과 거의 비슷해진다(n〉=120)

$s_{\overline{X}} = s/\sqrt{n}$

$t = (\overline{X} - \mu)/s_{\overline{X}}$ (t distributed with n-1 degrees of freedom)

Testing hypothesis based on the t Statistic

One-Sample t-Tests

; examples of One-sample t-test

; The market share for the new product will exceed 15%.

; At least 65% of customers will like the new package design.

; The average monthly household expenditure on groceries exceeds $500.

; The new service plan will be preferred by at least 70% of the customers.

; Hypothesis는 아래와 같은 형태를 띤다.

$Ho: \mu \leqq 7.0$

$Ha: \mu > 7.0$

Test for a Single Mean

; Example 1) #page 465: New machine attachment의 도입 여부 (70% 이상) [풀어볼 것]

; mean, STD, a를 이용해 t를 계산한 다음에, df=n-1을 이용해 a=0.05의 t_crit를 계산한다.

; t_{calc} 〉 t_{crit} 이면 Ho reject 이다.

; 또한 모집단의 STD가 알려져 있다면, z-test를 이용할 수도 있다.

; Example 2) 2개의 샘플(10대/20대이상)에 대해 방문전/후 선호도 변화 조사 [풀어볼 것]

; 이 중 10대-방문 전 preference가 5.0인지 검사하는 문제.

; z-Test

; A univariate hypothesis test using the standard normal distribution.

$z = (\overline{X} - \mu)/\sigma_{\overline{X}}$

Test for a Single Proportion

Two-Sample t-Tests

Two Independent Samples

; Examples

; The populations of users and nonusers of a brand differ in terms of their perceptions of the brand.

; The high-income consumers spend more on entertainment than low-income consumers.

; The proportion of brand loyal users in Segment I is more than the proportion in Segment II

; The proportion of households with an Internet connection in the US exceeds that in Germany.

; 보면 알겠지만 각각의 가정들에서 2개의 다른 모집단이 쓰임을 알 수 있다. 표본들은 이들 모집단에서 임의로 뽑혀져 나와 독립적인 표본이 된다.

; Independent samples

; Two samples that are not experimentally related. The measurement of one sample has no effect on the values of the other sample.

Means

; 보통 Hypothesis는 아래와 같은 형태를 띤다.

$Ho: \mu_1 = \mu_2$

$Ha: \mu_1 \neq \mu_2$

[1] (If both populations are found to have the same variance) Pooled variance estimate is computed from the two sample variances:

$$s^2 = \frac{\sum_{i=1}^{n_1}(X_{i_1} - \overline{X}_1)^2 + \sum_{i=1}^{n_2}(X_{i_2} - \overline{X}_2)^2}{n_1 + n_2 - 2}$$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The standard deviation of the test statistic can be estimated as:

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

The appropriate value of t can be calculated as:

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{s_{\overline{X_1} - \overline{X_2}}}$$

이 경우들에 있어서 degrees of freedom = $(n_1 + n_2 - 2)$가 된다.

[2] (If the two populations have unequal variances) an exact t cannot be computed for the difference in sample means.

; 이 경우에는 t의 근사값을 구할 수 밖에 없다. 또한 df값은 정수로 나오지 않기에 반올림하여 사용하여야

한다.

The standard deviation of the test statistic can be estimated as:

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

; F-test

; A statistical test of the equality of the variance of two populations.
; 즉 2개의 모집단이 같은 분산을 가지고 있는지 알지 못할 때 쓰인다.

Hypotheses are:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

; F statistics

; The F statistics is computed as the ratio of two sample variance.

$$F_{(n_1-1, n_2-1)} = \frac{s_1^2}{s_2^2}$$

이 때,

n_1 = size of sample 1
n_2 = size of sample 2
n_1-1 = degrees of freedom for sample 1
n_2-1 = degrees of freedom for sample 2
s_1^2 = sample variance for sample 1
s_2^2 = sample variance for sample 2

; F distribution

; A frequency distribution that depends upon two sets of degrees of freedom: the degrees of freedom in the numerator and the degrees of freedom in the denominator.

; 보면 알겠지만 F_{crit}는 2개의 df를 필요로 한다. 하나는 numerator, 또다른 하나는 denominator.

; example 3) F test를 이용하여 2개 집단 간을 비교하는 예제 p470 [풀어볼 것]

; one-tailed와 two-tailed를 언제 어느 때 사용하는지 정확하게 알아두도록 할 것. 대개 two-tailed test가 보다 conservative하다. 즉 two-tailed test에서 Ho가 기각된다면, one-tailed test에서도 Ho는 기각된다.

Proportions

; Example) US, Hong Kong 청바지 예제 [p472] [풀어볼 것]

$$z = \frac{P_1 - P_2}{S_{P_1 - P_2}}$$

$$S_{P_1 - P_2} = \sqrt{P(1-P)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}, \quad P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

The t-test is equivalent to a chi-square test for independence in a 2x2 contingency table. The relationship:

$$\chi^2_{(1)} = t^2_{n_1 + n_2 - 2}$$

Paired Samples

; Examples
; Shoppers consider brand name to be more important than price when purchasing fasion clothing
; Households spend more money on pizza than on hamburgers.
; The proportion of households who subscribe to a daily newspaper exceeds the proportion subscribing to magazines.
; The proportion of a bank's customers who have a checking account exceeds the proportion who have a savings account.

; Paired samples

; In hypothesis testing, the observations are paired so that the two sets of observations relate to the same respondents.
; 즉 같은 사람이 2번 응답한 것이라고 생각하면 된다.

Means

; Paired-samples t-test

; A test for differences in the means of paired samples.
; 아래와 같은 공식을 세울 수 있다.

$Ho : \mu_D = 0$ (이 때 μ_D는 paired-difference이다.)
$Ha : \mu_D \neq 0$

$$t_{n-1} = \frac{\overline{D} - \mu_D}{S_{\overline{D}}} \quad \text{(The degrees of freedom : n-1)}$$

$$t_{n-1} = \frac{\overline{D} - \mu_D}{\frac{S_D}{\sqrt{n}}}$$

Where

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

$$S_D = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1}}$$

$$s_{\overline{D}} = \frac{s_D}{\sqrt{n}}$$

; Example) Disney case를 이용하여 계산하는 예제 [p474-475] [풀어볼 것]
; 이 때 D=preference after the visit-preference before the visit 이 된다.

Proportions

; 이는 chi-square로 테스트 될 수 있다!

Testing hypotheses for more than two samples

; Analysis of variance(ANOVA)

; A statistical technique for examining the differences among means for two or more populations.
; 이 때 보통 Ho는 모든 mean이 equal함을 가정한다.

; examples
; Do the various segments differ in terms of their volumes of product consumption?
; Do the brand evaluations of groups exposed to different commercials vary?
; Do retailers, wholesalers, and agents differ in their attitudes toward the firm's distribution policies?
; Do the users, nonusers, and former users of a brand differ in their attitudes toward the brand?

Dependent and Independent Variables

; 보통 ANOVA는 dependent variables(metric, 즉 ratio scale이어야 한다)를 가져야 한다. 또한 1개나 2개의 independent variable을 가진다. 이들 independent variables들은 모두 categorical(nonmetric)해야 한다.

; One-way analysis of variance

; An ANOVA technique in which there is only one factor.

; 즉 하나의 categorical variable 혹은 single factor가 다른 그룹 샘플을 정의하게 된다.

; Factor

; Categorical independent variable; the independent variable must be categorical (nonmetric) to use ANOVA.

; Treatment

; In ANOVA, a particular combination of factor levels or categories.

; Independent variable=X, dependent variable = Y. 이 때 X는 c개의 categori를 가지고, Y에는 n개의 observation이 존재하게 된다. 따라서 총 sample size ,$N = n \times c$가 된다.

Decomposition of the Total Variation

; Decomposition of the total variation

; In one-way ANOVA, separation of the variation observed in the dependent variable into the variation due to the independent variables plus the variation due to error.

; SS_y

; The total variation in Y

$SS_y = SS_{between} + SS_{within}$

; $SS_{between}$

; Also denoted as SS_x, the variation in Y related to the variation in the means of the categories of X. This represents variation between the categories of X, or the portion of the sum of squares in Y related to X.

; SS_{within}

; Also referred to as SS_{error}, the variation in Y due to the variation within each of the categories of X. This variation is not accounted for by X.

; 따라서 아래와 같이 표현하는 것이 가능하다.

$SS_y = SS_x + SS_{error}$

$SS_y = \sum_{j-1}^{c}\sum_{i=1}^{n}(Y_{ij-}\overline{Y})^2$

or $SS_y = \sum_{i=1}^{n}(Y_i - \overline{Y})^2,\ SS_x = \sum_{j=1}^{c}\sum_{i=1}^{n}(Y_j - \overline{Y})^2$

or $SS_x = \sum_{j=1}^{c} n(Y_i - \overline{Y})^2,\ SS_{error} = \sum_{j=1}^{c}\sum_{i=1}^{n}(Y_{ij} - \overline{Y_j})^2$

Y_i = individual observation

$\overline{Y_j}$ = mean for category j

$\overline{Y}$ = mean over the whole sample, or grand mean

Y_{ij} = i^{th} observation in the j^{th} category

Measurement of Effects

; X의 Y에 대한 효과(effect)는 SS_x로 측정된다. SS_x가 X의 카테고리의 평균 간의 분산에 연관되어 있기에 SS_x의 상대적 양은 X 카테고리 내의 Y의 평균 간의 차이가 증가할수록 증가하게 된다. 따라서 X의 Y에 대한 효과는 아래와 같이 측정될 수 있다.

; eta2(η^2)

; The strength of the effects of X(independent variable or factor) on Y(dependent variable) is measured by eta2(η^2). The value of η^2 varies between 0 and 1.

$\eta^2 = SS_x/SS_y = (SS_y - SS_{error})/SS_y$

Significance Testing

; one-way ANOVA에서 Ho는 아래와 같은 형태로 설정된다.

$Ho: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_c$

; The estimate of the population variance of Y

$S_y^2 = SS_x/(c-1)$ = Mean Square due to X = MS_x

혹은

$S_y^2 = SS_{error}/(N-c)$ = Mean square due to error = MS_{error}

와 같이 나타낼 수 있다.

; Mean square

; The sum of squares divided by the appropriate degrees of freedom.

; Significance of the overall effect

; A test to determine whether some differences exist between some of the treatment groups.

$$F = \frac{SS_x/(c-1)}{SS_{error}/(N-c)} = \frac{MS_x}{MS_{error}}$$

; 이는 (c-1), (N-c) df를 가진 F distribution을 따른다.

Illustrative Applications of One-way Analysis of Variance

; Example) Effect of In-store promotion on sales [p479-p480] [풀어볼 것]

Sample	Test/comments
One sample	
Means	t-test, if variance is unknown z-test, if variance is known
Proportions	z-test
Two independent samples	
Means	Two-group t-test F-test for equality of variances
Proportions	z-test Chi-square test
Paired samples	
Means	Paired t-test
Proportions	Chi-square test
More than two samples	
Means	One-way analysis of variance
Proportions	Chi-square test

[Discussion Problems]

2. The current advertising campaign for a major automobile brand would be changed if fewer than 70% of the consumers like it.

a. Formulate the null and alternative hypotheses.

Ho : $\pi \geqq 0.70$

Ha : $\pi < 0.70$

Ho will be rejected if

b. Which statistical test would you use? Why?

One-sample t-test will be used. Because we test for a single proportion.

$$Z = \frac{p - \pi}{\sigma_p}$$

c. A random sample of 300 consumers was surveyed, and 204 respondents indicated that they liked the campaign. Should the campaign changed? Why?

a = 0.05

Ho will be rejected if TS_{CAL} 〉 TS_{CRIT} = 1.645

$$Z = \frac{p - \pi}{\sigma_p} = \frac{0.7 - 204/300}{\sqrt{\frac{204/300 \times (1 - (204/300))}{300}}} = 0.74$$

TS_{CAL} = 0.74 〈 1.645

So Ho will not be rejected. The campaign will not be changed.

하지만 sample의 수가 300보다 더 많다면 더 괜찮은 결과를 얻어낼 수 있다. 즉 보다 많은 sample number가 필요함.

3. A major computer manufacturer is having an end-of-season sale on computers. The number of computers sold during this sale at a sample of 10 stores was 800 1100 0 400 700 800 1000 500 800 300

a. Is there evidence that an average of more than 500 computers per store were sold during this sale? Use a = 0.05

$Ho : \mu \leqq 500$

$Ha : \mu > 500$

a = 0.05

Ho will be rejected if $P(t_{calc}) < \alpha = 0.05$

[SPSS 수행 결과]

	Test Value = 500					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Sales	1.313	9	.222	140.000	-101.30	381.30

$P(t_{calc}) = 0.222 > 0.05$

So Ho will not be rejected.
(혹은 95% Confidence interval을 보아도 바로 알 수 있다. 범위에 0이 들어가면 뽑아낸 값이 큰지 작은지 신뢰할 수 없기 때문에 Ho를 기각하지 못하게 된다)

이 외에도 Test value를 변경하거나(val=50) 신뢰구간을 변경(65%)하면 Ho를 기각할 수 있게 된다.

b. What assumption is necessary to perform this test?

4. After receiving complaints from readers, your campus newspaper decided to redesign its front page. Two new formats, B and C, are developed and tested against the current format, A. A total of 75 students are randomly selected, and 25 students are randomly assigned to each of three format conditions. The students are asked to evaluate the effectiveness of the format on a 11-point scale.

a. State the null hypothesis.

$Ho: \mu_A = \mu_B = \mu_C$

Ha : At least two means differ

b. What statistical test should you use?

c. What are the degrees of freedom that are associated with the test statistic?

One-way analysis of variance testing will be used (because more than 2 variables are involved)

a = 0.05

Reject Ho if P(F) 〈 0.05

Reject Ho if $F_{calc} > F_{critical}$ = 3.10

* 이 때 $F_{critical}$ 계산하는 방법 :

"How many groups do we have?" = c = 3

"df between groups"= df1 = (c-1) = 3-1 = 2

"df within groups" = df2 = (k-c) = 75-3 = 72

$F_{2,72} = 3.10$

5. A marketing researcher wants to test the hypothesis that, within the population, there is no difference in the importance attached to shopping by consumers living in the northern, southern, eastern, and western United States. A study is conducted and analysis of variance is used to analyze the data. The results obtained are presented in the following table.

Source	df	Sum of squares	Mean squares	F ratio	F probability
Between groups	3	70.212	23.404	1.12	0.3
Within groups	996	20812.416	20.896		

a. Is there sufficient evidence to reject the null hypothesis?

b. What conclusion can be drawn from the table?

$Ho: \mu_A = \mu_B$

$Ha: \mu_A \neq \mu_B$

a = 0.05

Reject Ho if $P(F_{calc}) < 0.05$

Reject Ho if $F_{calc} > F_{critical} = F_{3,996} = 2.60$

결과를 분석해 보면

$F_{calc} = 1.12 < 2.60$

So do not reject Ho. There is no sufficient evidence to reject Ho.

c. If the average importance was computed or each group, would you expect the sample means to be similar or different?

It will be similar because $SS_{between} \ll SS_{within}$.

d. What is the total sample size in this study?

df1 = (c-1) = 3. So c=4.

df2 = (k-c) = 996. So k=100.

n = c+k = 4+996 = 1000 observations.

6. In a pilot study examining the effectiveness of three commercials (A, B, and C), 10 consumers were assigned to view each commercial and rate it on a nine-point Likert scale. The data obtained are shown in the following table. These data should be analyzed by doing hand calculations.

a. Calculate the category means and the grand means.

[SPSS] Compare means를 수행한다.

comm	Mean	N	Std. Deviation
1.00	4.0000	10	.81650
2.00	5.0000	10	1.05409
3.00	7.0000	10	1.05409
Total	5.3333	30	1.58296

b. Calculate SS_y, SS_x, and SS_{error}.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	46.667	2	23.333	24.231	.000
Within Groups	26.000	27	.963		
Total	72.667	29			

$SS_y = SS_{total} = 72.667$

$SS_x = SS_{between} = 46.667$

$SS_{error} = SS_{within} = 26.000$

c. Calculate η^2.

$\eta^2 = SS_x/SS_y = 46.667/72.667 = 0.642$

d. Calculate the value of F.

$F = MS_x/MS_{error} = 23.333/0.963 = 24.231$

e. Are the three commercials equally effective?

$Ho: \mu_A = \mu_B = \mu_C$

Ha : At least two means differ

a = 0.05

Reject Ho if $F_{calc} > F_{critical} = F_{2,7} = 4.74$

결과를 분석해보면

$F_{calc} = 24.231 > 4.74$

So Ho will be rejected.

추가적으로, Post-hoc test에 Scheffe를 함께 수행하면 그룹간 비교가 가능하게 된다. 아래는 결과 테이블이다.

comm	N	Subset for alpha = .05	
		1	2
1.00	10	4.0000	
2.00	10	5.0000	
3.00	10		7.0000
Sig.		.093	1.000

즉 1과 2는 같은 그룹으로 묶을 수 있고, 3은 다른 그룹으로 묶어야 함을 알 수 있다.

8. In a pretest, respondents were asked to express their preference for an outdoor lifestyle(V1) using a seven point scale. They were also asked to indicate the importance of the following variables on a seven-point scale.

a. Does the mean preference for an outdoor lifestyle exceed 3.0?

[SPSS] 단일 비교이므로 One-sample t-test를 사용하면 된다.

$Ho: \mu \leqq 3.0$
$Ha: \mu > 3.0$
a = 0.05

Reject Ho if $P(t_{calc}) < 0.05$

[SPSS수행결과] Test value=3을 집어넣고 계산하면 됨.

	Test Value = 3					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
preferen	2.893	29	.007	1.03333	.3029	1.7638

$P(t_{calc}) = 0.007 < 0.05$

So Reject Ho. ('cause zero is not included!)
또한 95% 신뢰구간 내의 Lower~Upper 사이에 0이 포함되는지를 보는 것이 빠르다.

b. Does the mean importance of enjoying nature exceed 3.5?

[SPSS] 단일 비교이므로 One-sample t-test를 사용하면 된다.

$Ho: \mu \leqq 3.5$
$Ha: \mu > 3.5$

	Test Value = 3.5					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
nature	3.225	29	.003	1.10000	.4025	1.7975

역시 Reject Ho. ('cause zero is not included!)

c. Does the mean preference for an outdoor lifestyle differ for males and females?

[SPSS] 그룹이 서로 다른 category에 속해 있으므로, two independent T-test를 수행한다. (결과에 Levene's test 포함시킴)

$Ho: \mu_M = \mu_W$
$Ha: \mu_M \neq \mu_W$
a = 0.05
Reject Ho if $t_{calc} > t_{critical}$ = 2.0484 (df=28, a=0.05/2)

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
preferen	Equal variances assumed	2.746	.109	.092	28	.928	.06667	.72681	-1.42214	1.55547
	Equal variances not assumed			.092	25.980	.928	.06667	.72681	-1.42737	1.56070

결과 분석
$t_{calc} = 0.092 < 2.0484$
Do not reject Ho. Keep Ho.('cause zero is included!)

d. Does the importance attached to V2 through V6 differ for males and females?

[SPSS] 각각에 대해서 수행하며, 다른 category respondent를 대상으로 하므로 Independent samples t-test 수행한다.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
nature	Equal variances assumed	,072	,790	-8,025	28	,000	-3,06667	,38214	-3,84945	-2,28389
	Equal variances not assumed			-8,025	26,537	,000	-3,06667	,38214	-3,85140	-2,28194
weather	Equal variances assumed	1,067	,310	-,207	28	,837	-,13333	,64390	-1,45230	1,18563
	Equal variances not assumed			-,207	27,742	,837	-,13333	,64390	-1,45285	1,18618
environm	Equal variances assumed	1,485	,233	-3,038	28	,005	-1,60000	,52675	-2,67899	-,52101
	Equal variances not assumed			-3,038	26,787	,005	-1,60000	,52675	-2,68119	-,51881
exercise	Equal variances assumed	1,495	,232	-1,095	28	,283	-,66667	,60893	-1,91400	,58067
	Equal variances not assumed			-1,095	27,926	,283	-,66667	,60893	-1,91415	,58081
people	Equal variances assumed	,009	,925	-3,117	28	,004	-1,86667	,59894	-3,09354	-,63979
	Equal variances not assumed			-3,117	27,989	,004	-1,86667	,59894	-3,09356	-,63977

[결과 분석]

$|t_{calc}| > t_{critical} = 2.0484$ 인 것들은 모두 Significance 한 것이다. 혹은 95% CI에 0이 포함되지 않으면 Significance 한 것.

nature : SIG

weather : N/S

environment : SIG

exercise : N/S

people : SIG

* SIG : Significant (0 is not included)

* N/S : Not Significant (0 is included)

- 이 때 Interval에서 negative 값이 나온다는 것은 여성의 값이 남성의 값보다 크다는 의미이다. 따라서 nature, environment, people에서 여성의 value가 남성보다 크다는 것을 알 수 있다.

[e-f] 이하 e~f는 같은 동일 응답자에 대해 비교하는 것이므로 Paired t-test를 사용해서 분석한다. 결과 테이블은 아래와 같다.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	nature - weather	1,00000	2,33415	,42616	,12841	1,87159	2,347	29	,026
Pair 2	weather - people	-,26667	1,98152	,36178	-1,00658	,47325	-,737	29	,467
Pair 3	environm - exercise	,93333	1,94641	,35536	,20653	1,66013	2,626	29	,014

e. Do the respondents attach more importance to enjoying nature than they do to relating to the weather?

결과 95% CI에 0이 포함되지 않는다. 따라서 귀무가설 기각.

⇒ Yes.

f. Do the respondents attach more importance to relating to the weather than that they do to meeting other people?

결과 95% CI에 0이 포함된다. 따라서 귀무가설 기각하지 않는다.

⇒ No. (no difference in preference)

g. Do the respondents attach more importance to living in harmony with the environment than they do to exercising regularly?

결과 95% CI에 0이 포함되지 않는다. 따라서 귀무가설 기각.

⇒ Yes.

Chapter 18. Data Analysis : Correlation and Regression

Product Moment Correlation

; Product moment correlation

; A statistic summarizing the strength of association between two metric variables.

; 즉 이는 2개의 계량 변수(interval or ratio) 간 연관성의 세기와 방향을 요약하는데 사용되는 변수이다. 이는 때때로 Pearson correlation coefficient, simple correlation, bivarate correlation, 혹은 correlation coefficient 라고도 불린다.

; Examples

; How strongly are sales related to advertising expenditures?

; Is there an association between market share and size of the sales force?

; Are consumers' perceptions of quality related to their perceptions of prices?

; n개의 observation 하에서 X와 Y, 그리고 product moment correlation r은 아래와 같이 계산될 수 있다.

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

Division of the numerator and denominator by (n-1) gives

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

$$= \frac{COV_{xy}}{S_x S_y}$$

; 이 때 COV_{xy}는 X와 Y 간의 covariance(공분산)을 의미한다. covariance값은 + 혹은 -가 될 수 있고 SxSy로 나누기 때문에 r은 -1.0 ~ +1.0의 값을 가질 수 있게 된다.

; Covariance

; A systematic relationship between two variables in which a change in one implies a corresponding change in the other (COV_{xy})

; Example) Correlation 계산 예제 [p499] [풀어볼 것]

; 결과 값으로 r=1.0에 가깝게 나오면 강한 상관관계가 있음을 의미한다. (-1.0은 상관관계가 있으나 역의 상관관계가 있음을 의미)

; Since r indicates the degree to which variation in one variable is related to variation in another, it can also be expressed in terms of the decomposition of the total variation.

r^2 = Explained variation / Total variation = $\frac{SS_x}{SS_y}$

= (Total variation - Error variation) / Total variation = $\frac{SS_y - SS_{error}}{SS_y}$

; r과 r2 모두 연관성을 측정하는 대칭적(symmetric) 방법이다. 다른 말로 하자면 X에 대한 Y의 correlation이나 Y에 대한 X의 correlation이나 서로 같은 이야기라는 의미이다. 여기에서 어떤 변수가 종속 변수이고 독립 변수인지는 중요하지 않다. 또한 PMC(product moment coefficient)는 선형 관계의 정도를 측정하도록 되어있지 비선형 관계를 측정하도록 설계되어 있지 않다. 즉 r=0이라는 것은 '선형 관계가 없다'이지 '둘 사이에 아무런 관계가 없다'라는 의미는 아니라는 것이다. (예를들어 y=|-2x|과 같은 연관관계도 있을 수 있다) 이는 r로 잡아내는 것이 불가능하다.

; 표본이 아닌 모집단으로 계사되었을 시, PMC는 ρ로 표시된다. 계수 r은 p의 예측자(estimator)이다. 또한 이 때 X와 Y의 분산이 같은 형태여야 하지, 그렇지 않으면 r은 deflate되어 ρ를 underestimate하게 된다.

; The statistic significance of the relationship between two variables measured by using r can be conveniently tested.

; The hypotheses are

$Ho : \rho = 0$

$Ha : \rho \neq 0$

; The test statistic is

$t = r\left[\frac{n-2}{1-r^2}\right]^{1/2}$ (t distribution with df=n-2)

; Example) t test [p502] [풀어볼 것]

Regression Analysis

; Regression analysis (회귀 분석)

; A statistical procedure for analyzing associative relationships between a metric-dependent variable and one or more independent variables.

; Usage of Regression analysis

; ① Determine whether the independent variables explain a significant variation in the dependent variable: whether a relationship exists.

; ② Determine how much of the variation in the dependent variable can be explained by the independent variables: strength of the relationship

; ③ Determine the structure or form of the relationship: the mathematical equation relating the independent and dependent variables

; ④ Predict the values of the dependent variable

; ⑤ Control for other independent variables when evaluating the contributions of a specific variable or set of variables

; 다만 주의할 것은, criterion variable이 인과적인 관계로 independent variable에 의존적인 것은 아니라는 점이다.

Bivariate Regression (이변량 회귀)

; Bivariate regression

; A procedure for deriving a mathematical relationship, in the form of an equation, between a single metric-dependent variable and a single metric-independent variable.

; Bivariate regression model

; An equation used to explain regression analysis in which one independent variable is regressed onto a single dependent variable.

; Example

; Can variation in sales be explained in terms of variation in advertising expenditure? What is the structure and form of this relationship, and can it be modeled mathematically by an equation describing a straight line?

; Can the variation in market share be accounted for by the size of the sales force?

; Are consumers' perceptions of quality determined by their perceptions of price?

Conducting Bivariate Regression Analysis

Scatter Diagram

; Scatter diagram

; A plot of values of two variables for all the cases or observations.

; 보통 종속 변수를 수직축에, 독립 변수를 수평축에 plot 한다.

; 이를 이용하여 연구자는 자료상에 어떠한 패턴이나 문제가 있는지 알아낼 수 있다.

; Least-squared procedure

; A technique for fitting a straight line to a scattergram by minimizing the vertical distances of all the points from the line.

; 직선(straight line)이 얼마나 자료에 적합(fitting)한지를 검사하는 방법이다. 이렇게 찾아낸 best-fitting line을 regression line이라고 부른다.

; Sum of squared errors

; The sum of the squared vertical differences between the actual data point and the predicted one on the regression line.

Bivariate Regression Model

; The general form of a straight line is

$Y = \beta_0 + \beta_1 X$

Y = dependent or criterion variable

β_0 = intercept of the line

β_1 = slope of the line

X = independent or predictor variable

; 그러나 때로는 error term을 더하는 경우도 있다. 따라서 The basic regression equation은 아래와 같다.

$Y = \beta_0 + \beta_1 X + e_i$

e_i = the error term associated with the i^{th} observation.

Estimation of Parameters

; 대개의 경우 β_0과 β_1이 알려져 있지 않기에 표본 관측값(Sample observation)을 통해 아래와 같이 예측된다.

$\hat{Y}_i = a + bx_i$

; 이 때 $\hat{Y}_i$는 Y_i의 Estimated or predicted value로 불리고, a와 b는 β_0과 β_1의 estimator로 불린다. 또한 상수 b는 주로 nonstandardized regression coefficient로 불린다.

; Estimated or predicted value

; The value $\hat{Y}_i = a + bx_i$ where Y_i is the estimated or predicted value of Y_i and a and b are estimators of β_0 and β_1 respectively.

; Nonstandardized regression coefficient

; The weight or multiplier of the independent variable when it is regressed onto a single dependent variable.

; a와 b를 구하는 공식

$$b = \frac{COV_{xy}}{S_x^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}$$

; 이렇게 b를 구하면 a는 아래와 같이 구할 수 있다.

$a = \overline{Y} - b\overline{X}$

; Example) 이변량 회귀분석 구하기 [p506-507] [풀어볼 것]

Standardized Regression Coefficients

; Standardization

; The process by which the raw data are transformed into new variables that have a mean of 0 and a variance of 1.

; 이는 z value를 구하는 것과 유사하다.

; Beta coefficient (or beta weight)

; Also known as beta weight, used to denote the standardized regression coefficient

$B_{yx} = B_{xy} = r_{xy}$

; B_{yx} = the slope obtained by the regression of Y on X

; B_{xy} = the slope obtained by the regression of X on Y

; Relationship between standardized and nonstandardized regression coefficients

$B_{yx} = b_{yx}(S_x / S_y)$

; example) beta coefficient 구하기 [p508-509] [풀어볼 것]

Significance Testing

; X와 Y의 통계적 유의성 검정(significance testing)을 위해서는 아래의 가정을 설정한다.

$Ho : \beta_1 = 0$

$Ha : \beta_1 \neq 0$

; 이 때 귀무가설은 X와 Y간에 아무런 선형 관계가 없음을 암시한다. 테스트를 위해서는 아래의 t stastistic을 사용한다. (df=n-2)

$$t = \frac{b}{SE_b}$$

; 이 때 SE_b는 b의 standard deviation임.

; Standard error

; SE_b denotes the standard deviation of b and is called the standard error.

Strength and Significance of Association

; 두 변수간의 연관성의 강도(strength of association)은 coefficient of determination(r^2)을 통해 측정될 수 있다.

; Coefficient of determination

; The proportion of variance in one variable associated with the variability in a second variable.

; Total variation : SS_y

; Explained Variation : SS_{reg}

; Residual Variation : SS_{res}

; 아래와 같은 공식이 성립한다.

$$SS_y = SS_{reg} + SS_{res}$$

$$SS_y = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

$$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

; The strength of association :

$$r^2 = \frac{SS_{reg}}{SS_y} = \frac{SS_y - SS_{res}}{SS_y}$$

; example) r^2의 계산 문제 [p510-511] [풀어볼 것]

; X와 Y의 선형 관계의 중요성을 관측하는 또다른 방법은 coefficient of determination의 중요성을 검증하는 것이다.

; 가설

$H_0 : R^2_{pop} = 0$

$H_1 : R^2_{pop} > 0$

; 이 때는 F test가 사용되고 아래와 같이 계산할 수 있다. (df=1, n-2)

$$F = \frac{SS_{reg}}{SS_{res}/(n-2)}$$

Prediction Accuracy

; Standard error of estimate (SEE)

; The standard deviation of the actual Y values from the predicted Y values.

; "The larger the SEE is, the poorer the fit of the regression"

Examination of Residuals

; Residual

; The difference between the observed value of Y and the value predicted by the regression equation Y.

Multiple Regression (다변량 회귀)

; Multiple Regression

; A statistical technique that simultaneously develops a mathematical relationship between two or more independent variables and an interval-scaled dependent variable.

; Multiple regression model

; An equation used to explain the results of multiple regression analysis.

; examples

; Can variation in sales be explained in terms of variation in advertising expenditures, prices, and level of distribution?

; Can variation in market shares be accounted for by the size of the sales force, advertising expenditures, and sales promotion budgets?

; Are consumers' perceptions of quality determined by their perceptions of prices, brand images, and brand attributes?

; How much of the variation in sales can be explained by advertising expenditures, prices, and level of distribution?

; What is the contribution of advertising expenditures in explaining the variation in sales when the levels of prices and distribution are controlled?

; What levels of sales may be expected given the levels of advertising expenditures, prices, and level of distribution?

; General form of the multiple regression model :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k + e$$

; 이는 아래의 수식에 의해 예측된다.

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

Conducting Multiple Regression Analysis Partial Regression Coefficients

Partial regression coefficient

; 부분 회귀 변수 b_1을 해석하자면 X_1이 1단위 변화될 때 X_2는 통제되거나 변화되지 않는(constant) 경우의 Y의 예측된 변화량을 나타낸다. b2, b3도 마찬가지이다.

; Partial regression coefficient

; Also known as b_1, denotes the change in the predicted value of Y when X_1 is changed by one unit but the other independent variables, X_2 to X_k are held constant.

; The relationship of the standardized to the nonstandardized coefficients:

$$B_1 = b_1(S_{x1}/S_y)$$

...

$$B_k = b_k(S_{xk}/S_y)$$

; 다음의 상황 아래에서는 equation을 풀 수 없다:

1) The sample size n 〈= The number of independent variables k
2) One independent variable is perfectly corrrelated with another.

Strength of Association

; 다음을 통해 계산될 수 있다.

; Coefficient of multiple determination

; In multiple regression, the strength of association is measured by the square of the multiple correlation coefficient, R2, which is called the coefficient of multiple determination.

$$SS_y = SS_{reg} + SS_{res}$$

$$SS_y = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

$$SS_{reg} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

$$SS_{res} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$R^2 = \frac{SS_{reg}}{SS_y}$$

; R^2는 가장 큰 이변량 r^2 값보다 작을 수 없다. 또한 독립 변수들간의 상관관계가 낮을 때 커진다. 만약 독립 변수들이 통계적으로 독립이라면 R^2는 각 이변량 r^2값의 합계가 된다. R^2는 보다 많은 독립 변수들이 들어올 때 마다 커지게 된다. 여하튼 이러한 이유로 adjusted R^2를 사용한다.

; Adjusted R2

; The value of R^2 adjusted for the number of independent variables and the sample size.

Adjusted R^2 = $R^2 - \frac{k(1-R^2)}{n-k-1}$

; Example) R^2, Adjusted R^2 계산 예제 [p516] [풀어볼 것]

Significance Testing

; 이를 수행할 때 전체 회귀 수식과 부분 회귀 계수의 중요도를 고려하여야 한다. 전체적인 검정에 대한 귀무 가설은 모집단의 다변량 결정 계수 R^2_{pop} =0임을 가정한다.

$$H_0 : R^2_{pop} = 0$$

; 이는 아래와 동일하다

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_k = 0$$

; 전체적인 검정은 아래와 같이 F statistics를 이용하여 수행할 수 있다.

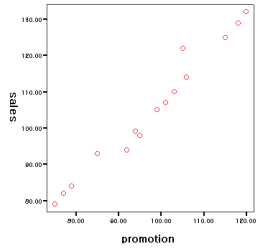
$$F = \frac{SS_{reg}/k}{SS_{reg}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

; example) R^2 검정 [p517] [풀어볼 것]

[Discussion Problems]

1. A major supermarket chain wants to determine the effect of promotion on relative competiveness. Data were obtained from 15 states on the promotional expenses relative to a major competitor (competitor expenses = 100) and on sales relative to this competitor (competitor sales = 100). You are assigned the task of telling the manager whether there is any relationship between relative promotional expense and relative sales.

a. Plot the relative sales (Y-axis) against the relative promotional expense (X-axis), and interpret this diagram.



more promotion, more sales. 선형 관계가 있음을 알 수 있다.

b. Which measure would you use to determine whether there is a relationship between the two variables? Why?
Bivariate Regression을 사용하여야 한다.

c. Run a bivariate regression analysis of relative sales on relative promotional expense.
d. Interpret the regression coefficients

[내가 돌린 것] : 교수님 정답과 약간 차이가 있으나, 무엇을 잘못 잡아서 그런지는 잘 모르겠음.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.984[a]	.969	.967	3.12817

a. Predictors: (Constant), promotion

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3972.523	1	3972.523	405.963	.000[a]
	Residual	127.210	13	9.785		
	Total	4099.733	14			

a. Predictors: (Constant), promotion
b. Dependent Variable: sales

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-9.768	5.747		-1.700	.113	-22.183	2.646
	promotion	1.175	.058	.984	20.149	.000	1.049	1.300

a. Dependent Variable: sales

Sales = f(const, promo) = -9.768 + 1.175(promo)
R^2 = 0.969

[교수님 정답]
Sales = f(const, promo)
Sales = -7.9 + 1.149(promo)
R^2 = 0.986

e. Is the regression relationship significant?
* 수업 안들어서 모르겠음.

f. If the company matched the competitor in terms of promotional expense (if the relative promotional expense was 100), what would the company's relative sales be?
* 수업 안들어서 모르겠음.

g. Interpret the resulting r^2.
R^2 = 0.969; so there is strong linear relationship between two variables.

2. To understand the role of quality and price in influencing the patronage of drugstores, 14 major stores in a large metropolitan area were rated in terms of preference to shop, quality of merchandise, and fair pricing. All the ratings were obtained on an 11-point scale, with higher numbers indicating more positive ratings.

a. Run a multiple regression analysis explaining store preference in terms of quality of merchandise and pricing.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	.535	.471		1.136	.280		
	Quality	.976	.097	.798	10.096	.000	.719	1.392
	Price	.251	.071	.278	3.522	.005	.719	1.392

a. Dependent Variable: Preference

Pref = f(const, qual, price)
Pref = .535 + .976 (qual) + .251 (price)

b. Interpret the partial regression coefficients.
c. Determine the significance of the overall regression.
d. Determine the significance of the partial regression coefficients.
quality is more important than price.
결과를 분석하여 보면 quality가 price보다 중요하다는 것을 알 수 있다.

3. Imagine that you've come across a magazine article reporting the following relationship between annual expenditure on prepared dinners (PD) and annual income (INC)

PD = 23.4 + 0.003 INC

The coefficient of the INC variable is reported as significant.

a. Does this relationship seem plausible? Is it possible to have a coefficient that is small in magnitude and yet significant?
b. From the information given, can you tell how good the estimated model is?

It is plausible.
But the problem is that it has "0.003" in coefficient. It is problematic that we don't know the range and drivers.

c. What are the expected expenditures on PDs of a family earning $30,000?
PD = 113.4

즉 얼마나 range가 되느냐에 따라서 결과를 해석하는 것이 달라질 수 있다.
we don't know the range of the model.

* magnitude problem, ...

d. If a family earning $40,000 spent $130 annually on PDs, what is the residual?

Residual = $Y - \hat{Y}$ = 130- (23.4+0.003*4000) = 94.6

e. What is the meaning of a negative residual?

They spend less money than expected.

5. Conduct the following analyses for the preference of the outdoor-lifestyle data.

a. Calculate the simple correlations between V1 to V6 and interpret the results.

Correlations

		preferen	nature	weather	environm	exercise	people
preferen	Pearson Correlation	1	,126	,787**	-,124	,647**	,416*
	Sig. (2-tailed)		,506	,000	,513	,000	,022
	N	30	30	30	30	30	30
nature	Pearson Correlation	,126	1	,162	,501**	,223	,448*
	Sig. (2-tailed)	,506		,393	,005	,237	,013
	N	30	30	30	30	30	30
weather	Pearson Correlation	,787**	,162	1	-,068	,395*	,398*
	Sig. (2-tailed)	,000	,393		,721	,031	,030
	N	30	30	30	30	30	30
environm	Pearson Correlation	-,124	,501**	-,068	1	,308	,047
	Sig. (2-tailed)	,513	,005	,721		,098	,807
	N	30	30	30	30	30	30
exercise	Pearson Correlation	,647**	,223	,395*	,308	1	,115
	Sig. (2-tailed)	,000	,237	,031	,098		,547
	N	30	30	30	30	30	30
people	Pearson Correlation	,416*	,448*	,398*	,047	,115	1
	Sig. (2-tailed)	,022	,013	,030	,807	,547	
	N	30	30	30	30	30	30

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Sig 〈 0.05, 즉

Preference-weather,
Preference-exercise,
Preference-people,
Nature-environment,
Nature-people,
Weather-exercise,
Weather-people에 correlation 관계가 있음을 알 수 있다.

b. Run a bivariate regression with preference for an outdoor lifestyle (V1) as the dependent variable and the importance of enjoying nature (V2) as the independent variable. Interpret the results.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,126[a]	,016	-,019	1,97477

a. Predictors: (Constant), nature

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3,424	,972		3,522	,001
	nature	,132	,196	,126	,675	,506

a. Dependent Variable: preferen

R^2=0.016으로서, 별다른 선형 관계가 없음을 알 수 있다.
beta값은 .132로서 + 의 관계이다.

c. Run a multiple regression with preference for an outdoor lifestyle as the dependent variables. Interpret the results. Compare the coefficients for V2 obtained in the bivariate and the multiple regressions.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,909[a]	,826	,790	,69711

a. Predictors: (Constant), people, environm, exercise, weather, nature

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	,563	,654		,861	,398
	nature	-,031	,118	-,029	-,258	,799
	weather	,566	,117	,502	4,850	,000
	environm	-,288	,128	-,240	-2,250	,034
	exercise	,594	,117	,508	5,086	,000
	people	,191	,109	,182	1,743	,094

a. Dependent Variable: preferen

Pref = .563 - .031(nature) + .566(weather) - .288(environ) + .594(exercise) + .191(people)
이 때 nature가 앞과는 다르게 - 값을 가지는 것을 알 수 있다.

6. In a pretest, data were obtained from 20 respondents on preferences for sneakers on a seven-point scale, 1 = not at all preferred, 7 = greatly preferred (V1). The respondents also provided their evaluations of the sneakers on comfort (V2), style (V3), and durability (V4), also on seven-point scales, 1 = poor, and 7 = excellent. The resulting data follow.

a. Calculate the simple correlations between V1 to V4 and interpret the results.

Correlations

		preferen	comfort	style	durabili
preferen	Pearson Correlation	1	,573**	,642**	,559*
	Sig. (2-tailed)		,008	,002	,010
	N	20	20	20	20
comfort	Pearson Correlation	,573**	1	,560*	,534*
	Sig. (2-tailed)	,008		,010	,015
	N	20	20	20	20
style	Pearson Correlation	,642**	,560*	1	,364
	Sig. (2-tailed)	,002	,010		,114
	N	20	20	20	20
durabili	Pearson Correlation	,559*	,534*	,364	1
	Sig. (2-tailed)	,010	,015	,114	
	N	20	20	20	20

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

이 때 나머지는 모두 significant 하지만 durability - style는 NOT important 하다는 것을 알 수 있다.

b. Run a bivariate regression with preference for sneakers (V1) as the dependent variable and evaluation on comfort (V2) as the independent variable. Interpret the results.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,573[a]	,328	,291	1,599

a. Predictors: (Constant), comfort

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	,028	1,337		,021	,983
	comfort	,921	,310	,573	2,967	,008

a. Dependent Variable: preferen

강한 상관관계 존재.

c. Run a bivariate regression with preference for sneakers (V1) as the dependent variable and evaluation on style (V3) as the independent variable. Interpret the results.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,642[a]	,412	,380	1,496

a. Predictors: (Constant), style

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1,078	,848		1,271	,220
	style	,739	,208	,642	3,554	,002

a. Dependent Variable: preferen

강한 상관관계 존재.

d. Run a bivariate regression with preference for sneakers (V1) as the dependent variable and evaluation on durability (V4) as the independent variable. Interpret the results.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.559[a]	.312	.274	1.619

a. Predictors: (Constant), durabili

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.307	.961		1.361	.190
	durabili	.598	.209	.559	2.857	.010

a. Dependent Variable: preferen

강한 상관관계 존재.

[교수님 해설] (b-d) 까지 종합하면

		R2	P(f)
pref = f	(comfort)	.291	.008
	(style)	.380	.002
	(durability)	.274	.010

와 같음을 알 수 있다.

e. Run a multiple regression with preference for sneakers (V1) as the dependent variable and V2 to V4 as the independent variables. Interpret the results. Compare the coefficients for V2, V3, and V4 obtained in the bivariate and the multiple regressions.

[SPSS] 다 집어넣고 Enter로 분석하면 style이 제일 중요하다고 나오고, Stepwise로 분석하면 역시 style 빼고 나머지는 다 빠지게 된다.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-.539	1.183		-.455	.655	-3.048	1.970		
	durabili	.336	.214	.313	1.571	.136	-.117	.789	.709	1.410
	comfort	.258	.360	.161	.717	.484	-.506	1.022	.561	1.781
	style	.504	.234	.438	2.152	.047	.007	1.001	.681	1.469

a. Dependent Variable: preferen

[Enter로 집어넣은 결과]

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	1.078	.848		1.271	.220	-.704	2.861		
	style	.739	.208	.642	3.554	.002	.302	1.176	1.000	1.000

a. Dependent Variable: preferen

[Stepwise로 집어넣은 결과]

pref = 1.078 + .739 (style)