**실험설계법**

**감성공학과**

**201434037 이정년**

**목차**

1. Typical Format of a Journal Article…………………………………………………………………………..………….3p
2. Descriptive Statistics : The Univariate Case……………………………………………………………..………….6p
3. Bivariate Correlation…………………………………………………………………………………………………………….8p
4. Foundations of Inferential Statistics……………………………………………………………………………………10p
5. Estimation……………………………………………………………………………………………………………………………13p
6. Hypothesis Testing……………………………………………..………………………………………………………………13p
7. Effect Size, Power, Cis, and Bonferroni………..………………………………………………..……………………13p
8. Logistic Regression………………………………..……………………………………………………………………………15p
9. Kruskalwallis (Nonparametric method for One-way Anova).............................................................

* **Typical Format of a Journal Article**
* **Typical format of a journal article can easily share information in professional fields**
* **The outline form of this book for writing article, as follows**
  1. **Abstract**
     + - Including information: Purpose, or Objective of the investigation, Participants description, Study Method, Summary of important findings
       - Eventually, You should write WHY(study was conducted), HOW(trying to answer the questions of interest, WHAT( Discovery Truth and study’s data were analyzed )
       - Why 1문장 -> How(어떤 종류의 논문인지 언급해도 좋음, 독립변수 종속변수 피험자정보정도까지만) 2문장 -> What ( 비율이 제일 많아야함)3문장
  2. **Introduction (가설세우는 것이 목적이면 기존문헌에서 Missing Point를 찾아냄)**

**2-1. Background**

* + - * Write connection between their studies and others’ previously published work.
      * Descript previous study for supporting your study’s hypotheses
      * **Although , However**

**2-2. Statement of purpose** ( one of the most important part of a research report)

* + - * Usually, State specific purpose or goal of the investigation
      * The statement of purpose can be as short a single sentence or as long as one or two full paragraphs.

**2-3 Hypotheses**

* + - * we can see this part in scientific research article
  1. **Method ( an author explains how the study was conducted, Must make about description between Independent variable and dependent variable)**

**3-1 Participants**

* + - * Describe who participated in the study
      * How many participants, Who they were, How they were selected
      * State three statistical concepts : n, M(mean), and SD(Standard Deviation)

**3-2 Measures**

* + - * What kinds of measuring instruments were used to collect the data
      * Including information: Materials, Equipment, Apparatus, Instrument, or Measures
      * What instrument was used (by indicating form, model number, publication date, etc.).
      * The dependent variable is closely connected to the measuring instrument used to collect data
      * What is dependent variable ? 1) interest to the researcher , 2)

**3-3 Procedure**

* + - * This section must be to give another readers who want know information for their studies (they can replicate your study )
      * Where was the study conducted, Who conducted the study, In what sequence did events take place? Did any of the subjects drop out prior to the study’s completion?

**3-4 Statistical Plans**

* + - * This part can present end of the method section or beginning of the report’s result section
  1. **Results (The results section of a journal article contains a technical report of how the statistical analyses turned out, 가설의관련내용만)**
     + - Can be presented within the **text** of the article
       - Can be summarized in one or more **tables**
       - can be displayed by means of a graph (technically called a **figure**).
  2. **Discussion**
     + - To explain what the results mean in regard to the central purpose of the study
       - Usually contains an underlying or obvious research question; the discussion section ought to provide a direct answer to that question.
       - Sometimes an author uses the discussion section to suggest ideas for further research studies
  3. **References**

**-** A research report normally concludes with a list of the books, journal articles, and other source material referred to by the author.

* **Descriptive Statistics : The Univariate Case**
* **We can say univariate analysis that can be summarized using picture techniques such as frequency distribution, Stem-and0leaf display, histogram and bar graphs.**
* **Next, the types of distributional shape are normal, skewed, bimodal or rectangular**
* **The researchers use four types of range which have standard deviation, variance, Z-score, and T-score**

1. Picture Techniques

* Frequency Distribution : Two types are simple frequency distribution(ungrouped frequency distribution) – Excerpt 2.1 and grouped frequency distribution – Excerpt 2.2
* Stem and-Leaf Display : Stem( score intervals which have lowest order) , Leaf( final digit is given for each observed score )
* Histograms and Bar Graphs: The difference point between Histogram and bar graphs is horizontal axis that histogram has frequency and bar graph has quantitative variable. 히스토그램은 바차트의 한 형태인데 x축이 변수 하나로 이루어지면 히스토그램, 바그래프는 두 컬럼으로 구성됨

2. Distributional Shape (Don’t enough space in paper, so we have to know meaning of a few terms that researchers use to describe the distributional shape of their data) –결측값 처리 필요

* Normal distribution: clustered near the middle of the continuum of observed scores, if value is high that meaning is data set has middle shape while low value is away from the middle area of scores.
* Skewed distribution: most of the scores end up being high or low, with a small percentage of scores strung out in one direction away from the majority.
* Bimodal/ Trimodal : two or three grouped together
* Rectangular (or uniform): if scores are fairly evenly distributed along the score continuum without any clustering at all
* Kurtosis: you may encounter the terms leptokurtic and platykurtic, which denote distributional shapes that are more peaked and less peaked (as compared with the normal distribution), respectively.
* most researchers consider data to be approximately normal in shape if the skewness and kurtosis values turn out to be anywhere from -1.0 to +1.0 to decide statistical method

3. Measures of Central Tendency

* The mode, median, and mean are the most popular measures of central tendency
* In a true normal distribution (or in any unimodal distribution that is perfectly symmetrical), the values of the mode, median, and mean are identical.

4. Measures of variability

* A measure of variability simply indicates the degree of this dispersion among the scores
* If the scores are very similar, there is little dispersion and little variability (**homogeneous**). If the scores are very dissimilar, there is a high degree of dispersion (**heterogeneous)**.
* The range is the simplest measure of variability
* the interquartile range indicates how much spread exists among the middle 50 percent of the scores
* The standard deviation and the variance, are usually better indices of dispersion
* In reporting their standard deviations, authors may use the abbreviation SD, the symbol s or σ simply write out the word sigma

5. Standard Scores

* A z-score close to 0, of course, indicates that the original raw score is near the group mean
* a Z score is an ordinary score transformed so that it better describes the location of that score in a distribution
* T scores are used to tell individuals how far their score is from the mean
* **Bivariate Correlation**
* **This chapter’s points have issue such as (whether) there is a relationship between the two sets of scores and (how) strong or weak that relationship is, presuming that a relationship does, in fact, exist**

1. Scatter plot : Correlational continuum

2. Correlation Coefficient

* Symbolized as r, a correlation coefficient is normally exported as a decimal number somewhere between +1.00 and -1.00.
* **positive correlation;** this indicates a **direct relationship(***high–high, low–low)*
* **negative correlation,** and this indicates an **indirect,** or **inverse, relationship(***high–low*, *low–high*)

3. **Pearson’s product–moment correlation**, and is designed for the situation in which (1) each of the two variables is **quantitative in nature** and (2) each variable is measured so as to **produce raw scores**

4. **Spearman’s rho(= rank–order correlation)**. This kind of correlation is similar Pearson’s in that it is appropriate for the situation in which both variables are **quantitative in nature**. With Spearman’s technique, however, each of the two variables is measured in such a way as to **produce ranks**

5. Python

-------------------------------------------------------------------------------------------------------

>>> x = [1,2,3,4,5]

>>> y = [2,3,4,5,6]

>>> import scipy.stats as ss

>>> ss.stats.pearsonr(x,y)

(1.0, 0.0)

>>> ss.stats.spearmanr(x,y)

(1.0, 0.0)

-------------------------------------------------------------------------------------------------------

>>> import pandas

>>> import numpy as np

>>> df = pandas.DataFrame(np.random.randn(10,4))

>>> df

0 1 2 3

0 0.569728 -1.579015 -1.051091 0.953708

1 -0.725900 0.567823 -0.564611 1.681534

2 1.030777 -0.736870 0.224325 1.096261

3 1.088989 0.098579 -0.706706 0.083662

>>> df.corr()

0 1 2 3

0 1.000000 -0.633255 0.105366 0.278315

1 -0.633255 1.000000 0.096879 -0.092352

2 0.105366 0.096879 1.000000 -0.128076

3 0.278315 -0.092352 -0.128076 1.000000

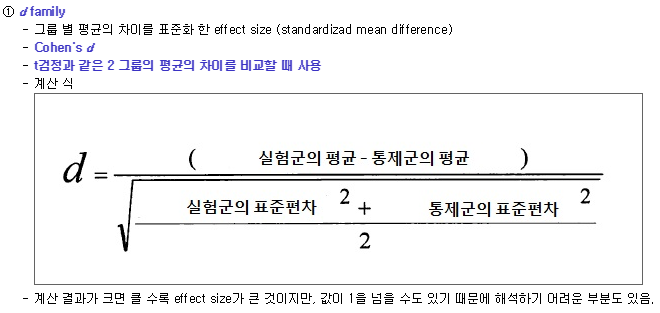
-------------------------------------------------------------------------------------------------------

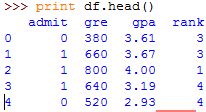
* **Foundations of Inferential Statistics**

1. Whenever a statistical inference is made, a sample is first extracted (or is considered

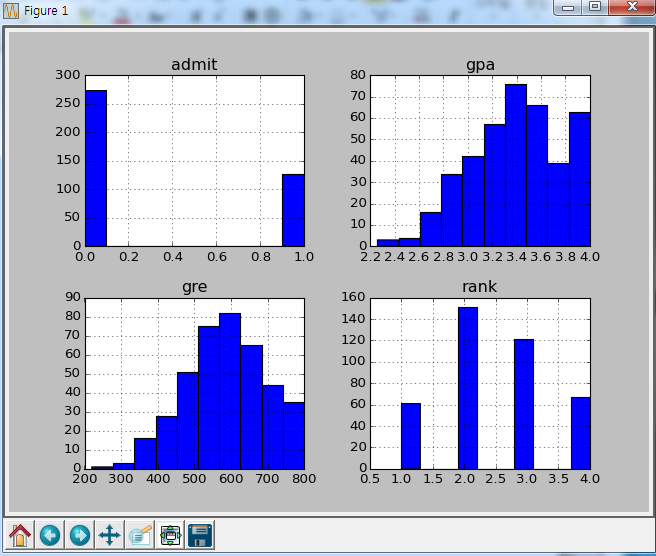
to have come from) a larger group called the population.

1. This educated guess as to the population’s numerical characteristic is the statistical inference.
2. If measurements could be obtained on all people (or objects) contained in the population, statistical inference would be unnecessary.
3. it is sometimes too costly (in dollars or time) to measure every member of the population
4. The principal’s guess about the average intelligence is based on a sample of students taken from the population made up of all students in the high school. In this example, the principal is sampling from a tangible population
5. In this case, the principal creates an abstract population to fit an existing sample.
6. Figure5.1
7. Excerpt 5.1-5.2
8. When researchers engage in inferential statistics, they must deal with four questions before they can make their educated guess, or inference, that extends from the sample to the population:
   1. 1. What is/are the relevant population(s)?
   2. 2. How will a sample be extracted from each population of interest, presuming the population(s) is/are tangible in nature?
   3. 3. What characteristic of the sample people, animals, or objects will serve as the target of the measurement process?
   4. 4. What will be the study’s statistical focus?
9. Table 5.1
10. With a simple random sample, The key feature of this kind of sample is an equal opportunity for each member of the population to be included in the sample
11. Excerpt5.3
12. In this excerpt, you see the term sampling frame. Generally speaking, a sampling frame is simply a list that enumerates the things—people, animals, objects, or whatever—in the population.
13. To reduce the possibility that the sample might turn out to be unrepresentative of the population, researchers sometimes select a stratified random sample
14. To do this, the population must first be subdivided into two or more parts based on the knowledge of how each member of the population stands relative to one or more stratifying variables
15. if a researcher knows that the population contains 60 percent males and 40 percent females, a random sample stratified on gender should contain six males for every four females.
16. EXCERPT 5.4
17. researchers make the size of the sample associated with one or more of the strata larger than that strata’s proportionate slice of the population(Oversampling)
18. Three Reasons
19. anticipated difficulty in getting people in certain strata to participate in the study,
20. a desire to make comparisons between strata (in which case there are advantages to having equal strata sizes in the sample, even if those strata differ in size in the population),
21. a need to update old strata sizes, when using archival data, because of recent changes in the characteristics of the population
22. Excerpt 5.5
23. a systematic sample, is created when the researcher goes through an ordered list of members of the population and selects,
24. Excerpt 5.6
25. When this technique is used to extract a sample from a population, the researcher first develops a list of the clusters in the population.
26. The clusters might be households, schools, litters, car dealerships, or any other groupings of the things that make up the population.
27. Next, a sample of these clusters is randomly selected.
28. Finally, data are collected from each person, animal, or thing that is in each of the clusters that has been randomly selected
29. Excerpt 5.7

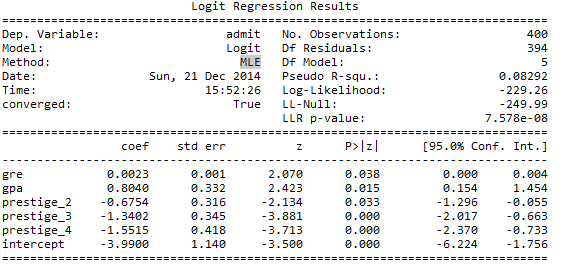
* **Estimation**
  + To understand how **interval estimation works**, you must become familiar with three concepts: sampling errors, standard errors, and confidence intervals
  + Sampling errors can be eliminated by selecting samples randomly from their appropriate populations.
  + The distribution of sample statistics alluded to in the preceding paragraph is called a **sampling distribution**, and the standard deviation of the values that make up such a distribution is called **a standard error (SE)**.
* **Hypothesis Testing (연구가설은 연구의문에서 나오는 결과이며, 이때 귀무가설은 연구가설에 반하는 결과를 얘기함, 즉 귀무가설은 Nullify -> Reject하려는 의도가 있음** 
  + When engaged in hypothesis testing, a researcher begins by stating a **null hypothesis**. If there is just one population involved in the study, the null hypothesis is a pinpoint statement as to the unknown quantitative value of the parameter in the population of interest.
  + In this case, the researcher **rejects** The other option available to the researcher is to refrain from asserting that is probably false. In this case, a **fail-to-reject** decision is made.
* **Effect Size, Power, Cis, and Bonferroni**
  + **연구에 베스트한 결과는 통계적 유의미한 차이도 있지만 Effect Size값도 높은 것이 가장 좋은 결과라고 말할 수 있음 Type I error발생때문에 베타 값(검정력)값을 올리면 오히려 Type II에러가 잘 발생할 수 있는 결과를 초래할 수 있음**
  + **연구를 Strict(엄격한)하게 연구를 진행하다 보면 위에서 말한 것과 같이 Type II에러가 발생할 수 있음 Bonferroni 사용하는 이유는 반복 검증하게 되면 1종 오류가 발생할 수도 있기 때문에 researchers sometimes use the Bonferroni technique to adjust their level of significance**
  + **예를 들면 Using the Bonferroni approach to control for type I error across the 10 correlations, a p value of less than .005 (.05/10 = .005) was specified for significance. (Excerpt 9.14)**
  + In Excerpts 8.1 and 8.2, this critically important distinction between **statistical significance and practical significance** is discussed
  + Researchers who are sensitive to the distinction between statistical significance and practical significance often add a seventh step to the basic version of hypothesis testing by estimating the study’s **effect size**
  + [**http://ultradorosy.blog.me/50191590744**](http://ultradorosy.blog.me/50191590744) **)**
  + 
* **Logistic Regression**
  1. gpa, gre, Rank or prestige of an applicant’s undergraduate alma mater 세 가지의 예측 변수를 사용해서 UCLA 대학원 합격 여부를 판단할 수 있는 회귀식을 만들고자 함
  2. 회귀분석이라는 것은 인과 관계를 보기 위해 사용되는 통계 방법으로 원인인 변수들을 예측 변수로 두었고, 대학원 승인 여부를 종속 변수로 사용함
  3. UCLA’s Logit Regression in R 데이터를 사용하여 본 과제의 결과를 수행함
  4. 사용할 예측변수(Predictor variables)
     1. Gpa
     2. Gre score
     3. Rank or prestige of an applicant’s undergraduate alma mater
  5. 데이터를 확인한 결과 총 400개의 데이터가 4열의 형태로 구성되어짐



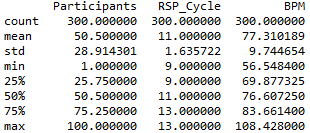
* 1. 해당 데이터의 분포 모양을 확인하기 위하여 화면에 분포 그림을 띄우는 작업을 수행함



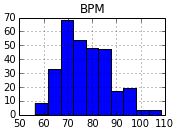
* 1. 지원자들의 모교 순위 같은 경우에는 명목 데이터이므로 Dummy 변수로 만드는 작업을 수행함
  2. 데이터 분석한 결과



* + 1. MLE(최대우도측정) 방법으로 회귀분석을 수행하였으며 Observations를 사용된 데이터 샘플수를 확인할 수 있음 400개데이터가 사용됨
    2. 모두 유효한 변수로 확인할 수 있었으며 각 모델의 계수 값은 gre = 0.0023, gpa = 0.8040, prestige\_2 = -0.06754, prestige\_3=-1.3402, 그리고 prestige\_4 = -1.5515 결과를 나왔고 상수 값은 -3.9900 값을 확인할 수 있었음
* **KruskalWallis Analysis**
* 모수/비모수 검정을 하기 위해선 크게 세가지 조건을 만족해야 함 첫째, 데이터가 정규분포를 따르고 있는가? 즉 정규성 검정을 수행해야 함 둘째, 등분산이 같다는 가정이 되어야 하며, 마지막으로 충분한 데이터 개수가 있어야 함
* 대표적인 모수 검정으로는 Independent t-Test, Paired t-Test, one-way Anova, Two-way Anova를 얘기할 수 있으며, T테스트와 아노바가 나뉘는 큰 이유는 집단의 개수 차이로 T테스트는 두 집단을 비교할 때 사용하는 통계 기법이며, 아노바 같은 경우에는 세 집단 이상을 비교할 때 사용하는 기법을 얘기함
* Independent t-Test와 Paired t-Test는 디자인된 실험이 독립적인 그룹으로 나누어 데이터를 받았으며, Independent t-Test를 피험자들이 Task만 다르고 반복 실험에 참여했으며 Paired t-Test를 사용하여 통계 처리를 해야함
* One-way와 Two-way로 나누어지는 이유는 독립변수 개수의 차이로 독립변수가 한 개인 경우에는 One-way 아노바를 독립변수가 2개이상인 경우에는 Two-way Anova를 사용하여 통계처리를 해야함
* **본 실험은 호흡 주기에 따른 BPM에 변화를 확인한 데이터로 Repeated 실험으로 설계가 되었으며 독립변수는 1개, Level는 세 개, 종속변수는 BPM(분당 맥박수)을 측정함**
* 총 100명이 참여 했으며 호흡 주기 Task가 세 종류로 나누어 진행됨(깊은 호흡, 보통, 짧은 호흡)

****

* BPM 데이터를 뿌려본 결과 다음과 같았으며 정규성 검정을 수행하였는데 정규분포를 따르지 않는다고 통계적 결과를 확인할 수 있었음 유의확률 (p < 0.05)
* Name: BPM, Length: 300, (10.264309696282368, 0.0059038249108728583)

****

* 따라서 비모수 검정을 실시하였으며 파이썬 다음 명령어를 사용하여 세 가지의 호흡주기에 따른 BPM(분당 맥박수)의 차이를 확인하였음
* print stats.mstats.kruskalwallis(New\_Raw\_Data)
* 그 결과 세 가지의 호흡 주기에 따라 p>0.05에 결과를 확인할 수 있었으며 따라서 세 가지의 호흡 주기에 따라 통계적 차이가 없다라는 결과를 확인함