Announcement: Mid-term Exam

- Scope: All materials provided by Week 7
- Preparation: 완충 휴대폰 + Laptop
  - 휴대폰: Webex 접속 → 본인화면 찍기
    - 컴퓨터화면, 키보드, 본인의 모습이 나타나도록 촬영
  - Laptop: 이러닝 사이트 접속 → 문제 풀기 & 제출
    - 8주차의 퀴즈 클릭 → 문제 별 답안 작성 → 제출
- NOTICE
  - Closed book
  - No cheating → Zero tolerance
  - 최종 제출 후 조기 퇴실 (접속종료) 가능 (재 접속 불가)
- 자세한 내용은 공지사항 참조

Announcement: Intermediate Presentation

- It's graded (0-100)
- Prepare **10 mins** presentation including:
  - Brief description of your research idea/design
    - Motivation + model + dataset + …
  - What you have done & What you will do more
  - (extra credit) share your troubleshooting during research!
- Q&A (**5 mins**)
- NOTES
  - Your progress should be **reasonable** for you to finish off by this semester
  - Attendance will be checked
  - Presentation order will be announced

Week 7: Statistics 2 (Quantitative Analysis Method)

Instructor: Daejin Choi (djchoi@inu.ac.kr)

INCHEON NATIONAL UNIVERSITY

* Most slides are used from www.munmund.net & www.deeplearningbook.org/lecture_slides.html with small updates

Goals & Contents

- Goal: Understanding Statistical Methods (Quant. Analysis)

- Contents
  - Terminologies
  - Central Tendency
  - Probability
  - Distribution & Information Theory
  - Computing Association between Distributions
  - Hyphothesis Testing

Association between Distributions

Pearson Correlation Coeff.

- A measure of the strength of the linear relation between two variables x and y

두 변수간의 상관 관계

- 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation

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

```
>>> from scipy import stats
>>> rvs1 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs6 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> stats.pearsonr(rvs1,rvs6)
(-0.028991426987572115, 0.5177734067731351)
```

↳ correlation p-value

Spearman's correlation

- Pearson's correlation assumes "normality of the distributions" being compared

- The Spearman correlation is a nonparametic measure of the monotonicity of the relationship between two datasets

```
>>> from scipy import stats
>>> rvs1 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs6 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> stats.spearmanr(rvs1,rvs6)
(-0.045163188652754614, 0.3135186410886480 2)
```

Kendall's Tau

- A measure of the correspondence between two *rankings.* → 랭킹 계산

- Close to 1 if strong agreement, -1 indicates strong disagreement
  → 랭크 거꾸로

```
>>> from scipy import stats
>>> rvs1 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs6 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> stats.kendalltau(rvs1,rvs6)
(-0.03191983967935 87, 0.2860220143222 6193)
```

한계점: 한 쪽의 element가 무조건 다른 쪽의 element에 속해있어야 함. size도 동일

- What if two sets have different elements?
  - Or, what if two sets have different #elements?

Hyphothesis Testing

Hyphothesis Testing

- Hyphothesis: A statement about some characteristic of a variable or a collection of variables. (Agresti & Finlay, 1997)

- When a hyphothesis relates to **characteristics of a population**, such as population parameters, one can use statistical methods with sample data to test its validity

- Example: Are female and male Twitter users distinct in terms of visit counts for the sites?

Hyphothesis Testing (cont'd)

- A way of statistically testing a hyphothesis by comparing the data to values predicted by the hyphothesis

- Data that fall from the predicted values provide evidence against the hyphothesis

- All significance tests have five elements: **Assumptions, hyphotheses, test statistic, p-value, and conclusion**

Hyphothesis Testing: Hyphothesis

- Considering two hyphotheses about the value of a population parameter
  - Null hyphothesis: A hyphothesis to directly test = H_0
  - Alternative hyphothesis: a hyphothesis contradicts the null hyphothesis

H_1

$H_0 \leftrightarrow H_1$

Hyphothesis Testing: Test Statistic

- Our goal is to test
  - If null hyphothesis is rejected, or
  - If alternative hyphothesis is accepted

- Compute *p*-value under assumption that H_0 is true. (i.e., we give the benefit of the doubt to the null hyphothesis)
  - Analyze how likely the observed data would be if that hyphothesis were true

H_0 : A와 B는 같다
H_1 : A와 B는 같지않다

- Example: t-test
  - Measures difference of means when comparing two distributions

Hyphothesis Testing: p-value

- The probability when H_0 is true, of a test statistic value at least as contradictory to H_0 as the value actually observed. **The smaller the p-value, the more strongly contradict H_0** (Agresti & Finlay, 1997)

p-value 얼마나 믿을 수 있는지

- p-value is often defined with respect to a chosen confidence level (95%, 99%, … ←→ 0.1, 0.05, 0.01, …)

p-value ↓ → H_0 기각 확률 ↑ = 유의미한 차이가 있다.

- The t-test has important assumptions that must be satisfied in order for the associated p-value to be valid
  - The samples are independent
  - Each sample is from a normally distributed population

Example of T-test

- Calculating t-test for the means of TWO INDEPENDENT samples of scores

```
>>> from scipy import stats

>>> rvs1 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs2 = stats.norm.rvs(loc=5,scale=10,size=500)

>>> stats.ttest_ind(rvs1,rvs2)
(0.26833823296239279, 0.78849443369564776)
>>> rvs4 = stats.norm.rvs(loc=5, scale=20, size=100)
>>> stats.ttest_ind(rvs1,  rvs4, equal_var = False)
(-0.69712570584654099, 0.48716927725402048)
```

Alter. hyphothesis is NOT reliable

유의미한 차이 X

i.e., mean is statistically same

U-test as an alternative way

- Limitation of t-test
  - Should be **normal** distribution
  - Not work when *n* is small
  - Not work when the values are order (or rank)

Not considering population

- The **Mann-Whitney U test** is a nonparametic test of the null hyphothesis that two populations are the same against an alternative hyphothesis 평균 고려 X, Normal distribution X
  - Good when a particular population tends to have larger values than the other

```
>>> from scipy import stats
>>> rvs2 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs5 = stats.norm.rvs(loc=50,scale=10,size=200)
>>> stats.ranksums(rvs2,rvs5)
(-17.268624879732251, 8.1050738020911939e-67)
```

= 차이가 있다.

↳ 매우 유의하다

NOTE: Type error

- Type Error: "INCORRECTLY" determining Null hyphothesis

1종 오류

- **Type Error 1**: Determines as "rejects" even though H_0 is true = H_0는 참이지만 틀렸다고 여생
  - When two "same" distribution is different sampled

- **Type Error 2**: Determines as "accepts" even though H_0 is false

H_0는 거짓이지만 맞았다고 예측

ANOVA as a Generalization of T-test

- Doing multiple two-sample t-tests would result in an increased chance of committing a statistical type I error

- ANOVA provides a statistical test of whether or not the means of several **groups** are equal
  - Generalizes the t-test to more than two

- Note: Similar to t-test, ANOVA assumes three key items
  - The samples are **independent**
  - Each sample is from a **normally distributed** population
  - **Homoscedasticity**: the standard deviations of the groups are all equal
    - How we can measure this? → F-test!

$$F = \frac{\text{variance between treatments}}{\text{variance within treatments}}$$

ANOVA Testing

- 1-way ANOVA for three independent samples:

```
>>> from scipy import stats
>>> rvs1 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs2 = stats.norm.rvs(loc=5,scale=10,size=200)
>>> rvs3 = stats.norm.rvs(loc=5,scale=10,size=240)
>>> [f_value, p_value] = stats.f_oneway(rvs1, rvs2, rvs3)
(1.6144352794299781, 0.19955607421980850)
```

↳ p-value > 0.05

⇒ H_0 가각 못함

= 차이 없다.

Kruskal Wallis H test

- If any of three assumptions of ANOVA are not true, it is recommended to use the **Kruskal Wallis H test**, which is a non-parametic version of ANOVA

- The Kruskal-Wallis H-test tests the null hyphothesis that the population median of all of the groups are equal

- The test works on 2 or more independent samples, which may have different sizes

```
>>> from scipy import stats
>>> rvs1 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs2 = stats.norm.rvs(loc=5,scale=10,size=200)
>>> rvs4 = stats.norm.rvs(loc=5,scale=20,size=240)
>>> stats.kruskal(rvs1,rvs2,rvs4)
(2.9637587853571858, 0.2272102695486149 2)
```

Kolmogorov-Smirnov Statistic

- A two-sided test for the null hypthothesis that 2 independent samples are drawn from the same continuous distribution

```
>>> from scipy import stats
>>> rvs1 = stats.norm.rvs(size=200, loc=0., scale=1)
>>> rvs2 = stats.norm.rvs(size=200, loc=0.5, scale=1.5)
>>> stats.ks_2samp(rvs1, rvs2)
(0.20833333333333337, 4.667497551580698e-005)
```

Thank you!

Instructor: Daejin Choi (djchoi@inu.ac.kr)