Probability Review

ECE/CS 498 DS U/G

Probability and Hypothesis Testing;

Lecture 4

Professor Ravi K. Iyer

Electrical and Computer Engineering

University of Illinois

Announcements

- HW1 is due on Jan 30 (Wed) by 23:59 hrs. Please upload your ipynb notebook and PDF of the slides via Compass2G
- No homework will be released this week
- Mini-Project 1 Released
  - Checkpoint 1 is due on Feb 3 (Sun) by 23:59 hrs
- Next lecture (Jan 30): 1st In-class Activity (ICA1)
  - Open notes (laptops, course notes, etc.)
  - Feel free to randomly form discussion groups of at most 3 people
  - Coverage:
    - Conditional probabilities and their applications
    - Hypothesis testing (p-value in particular)

Announcements (Cont.)

- Graduate Students Only :Timeline of Final Project (4-credit students only) [*Tentative*]:
  - Week 11: Propose two ideas by the Friday after Spring break (Mar 29)
  - Week 12: Meet with professor and TAs to discuss about project proposals and select one of the proposed idea as the final project to work on
  - Week 14: Submit Midterm Progress Report
  - Week 17 (final exam week): Deliver project presentation (out of class) and turn in final report

- Options for project ideas:
  - Extend existing mini-projects
  - Work on new data (from Kaggle Competitions, data and topics will be selected by TAs, TBA)

Timeline

Lectures

- **Last Wednesday lecture:**
  - Working with Python
  - Mini Project on Avs - summary
- **Today's lecture:**
  - Hypothesis Testing- review
- **Next lecture(s):**
  - In Class activity
  - Bayesian Networks

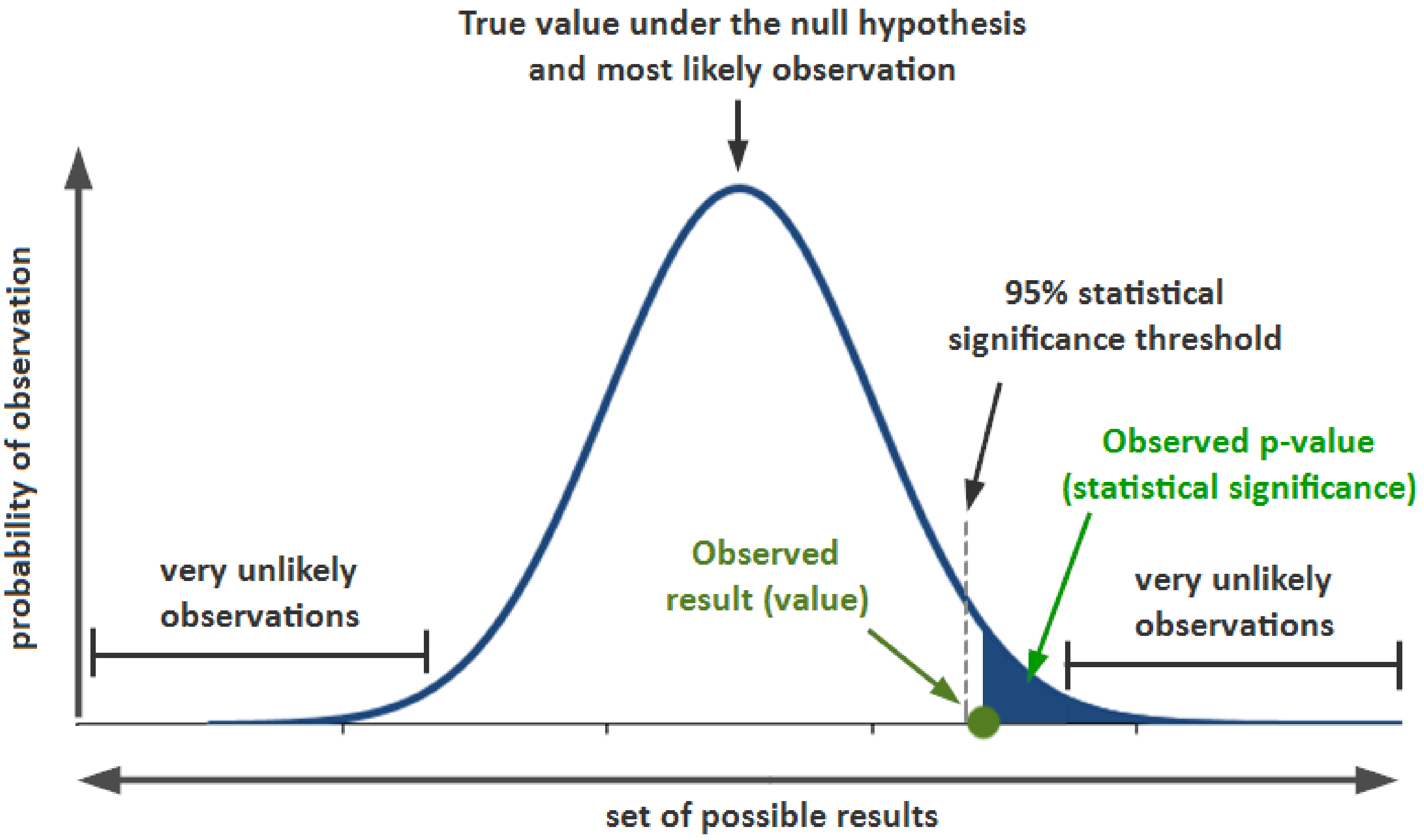
Hypothesis Testing (Review)

- In many practical problems we need to make decisions on the basis of limited information contained in a sample.
- To arrive at a decision we often make assumptions or a guess (an assertion) about the nature of the underlying population or situation. Such an assertion which may or may not be valid is called a **statistical hypothesis**.
- Procedures that enable us to decide whether to reject or accept hypotheses, based on the available information are called **statistical tests**.
- For Example
  - a system engineer/administrator may want to know if any one of the memory manufacturers is better than others
  - A Clinician may want to determine the efficacy of new drug vs the current drug in treating a given disease

Important Terms

- **Population** ≡ all possible values of a variable
- **Sample** ≡ a selection from the population
- **Statistical inference** ≡ generalizing a result from a sample to a population with calculated degree of certainty
- Two forms of statistical inference
  - **Hypothesis testing**
  - **Estimation**
- **Parameter** ≡ a characteristic of population, e.g., population mean μ
- **Statistic** ≡ calculated from data in the sample, e.g., sample mean ($\bar{x}$)

Hypothesis Testing (Intuition)



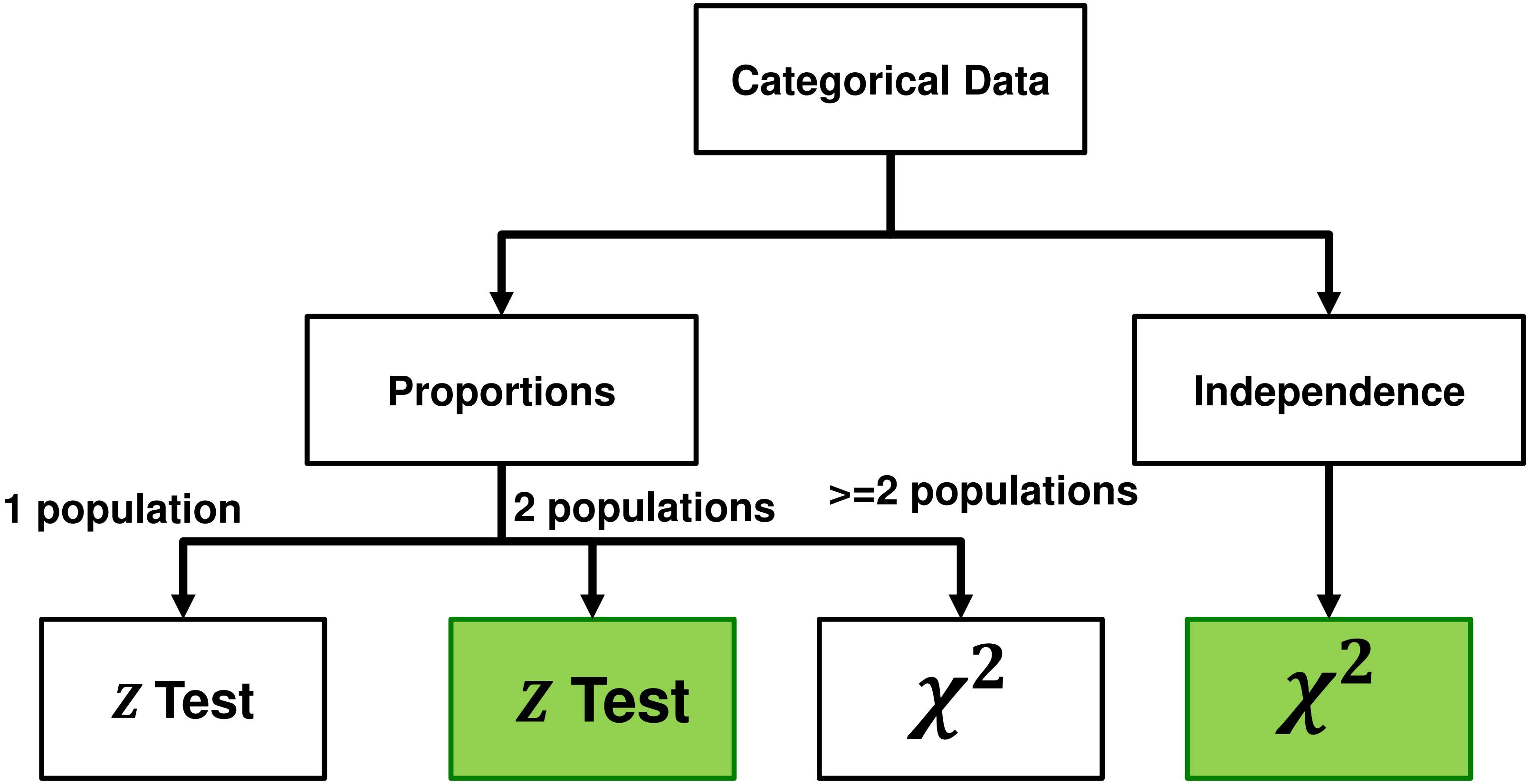
Categorical Data

1. Quantitative Random Variables yield responses that can be put In Categories. Example: Logic (True, False), Weather (Sunny, Rainy)

2. Measurement or Count: Reflects #s in a Category

3. Data can be collected as continuous but recoded to categorical data. Example – electrical signal (On or Off)

Hypothesis Tests for Categorical Data



Hypothesis Testing Steps

- Steps
  - Null and alternative hypotheses
  - Test statistic
  - P-value and interpretation
  - Significance level

Step1: Null and Alternative Hypotheses

- Convert the research question to null and alternative hypotheses
- The **null hypothesis (H_0)** is a claim of “no difference in the population”
- The **alternative hypothesis (H_a)** claims “H_0 is false”
- Collect data and seek evidence against H_0 as a way of bolstering H_a (deduction)

Step 2: Z Test for Difference in Two Proportions

- Assumptions
  - Populations Are Independent
  - Populations Follow Binomial Distribution
  - Normal Approximation Can Be Used for large samples (All Expected Counts > 5)
- Z-Test Statistic for Two Proportions

$$Z \cong \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}.(1-\hat{p}).\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Sample Distribution for Differences

Difference between means:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Difference between proportions:

$$\bar{p}_1 - \bar{p}_2 \cong N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

$$\cong N\left(0, \sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right), \text{under } \mathrm{H}_0 \text{:} \, p_1 = p_2, and\ \bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Step3: P-value

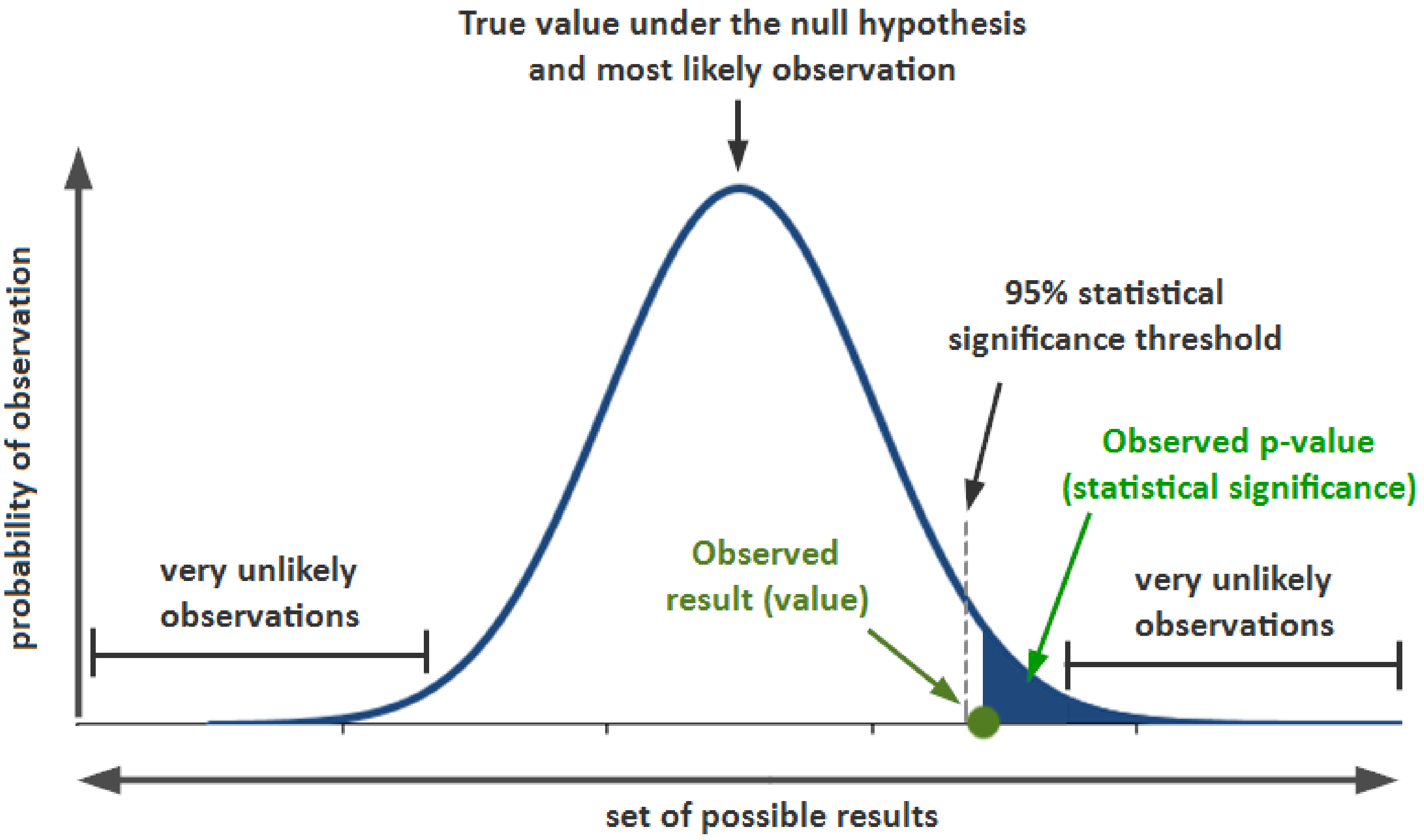
- The *P*-value answers the question: What is the probability of the observed test statistic or one more extreme **when *H_0 is true***?
- This corresponds to the AUC in the tail of the Standard Normal distribution beyond $z_{\text{stat.}}$
- Convert *z* statistics to *P*-value :

For H_a: $p_1 > p_2 \Rightarrow P = \Pr(Z > z_{\text{stat}})$ = right-tail beyond z_{stat}

For H_a: $p_1 < p_2 \Rightarrow P = \Pr(Z < z_{\text{stat}})$ = left tail beyond z_{stat}

For H_a: $p_1 \neq p_2 \Rightarrow P = 2$ × one-tailed *P*-value

P-value


True value under the null hypothesis
and most likely observation
probability of observation
95% statistical
significance threshold
Observed p-value
(statistical significance)
very unlikely
observations
Observed
result (value)
very unlikely
observations
set of possible results

Step4: α-Level (significance)

- Let α ≡ probability of erroneously rejecting H_0
- Set α threshold (e.g., let α = .10, **.05**, *or other*)
- Reject H_0 when $P \leq \alpha$
- Retain H_0 when $P > \alpha$
- Example: Set α = .10. Find $P = 0.27 \Rightarrow$ retain H_0
- Example: Set α = .01. Find $P = .001 \Rightarrow$ reject H_0

Z Test for Comparing Drug Efficacy

We are studying the efficacy of two new drugs for treating depression. The first drug was administered to 100 patients out of which 30 patients got cured. The second drug was administered to 200 patients of which 85 got cured. At **.05** level, is the second drug better than the first drug.

Z-test for two proportions setup

Let p_i be the efficacy of drug i.

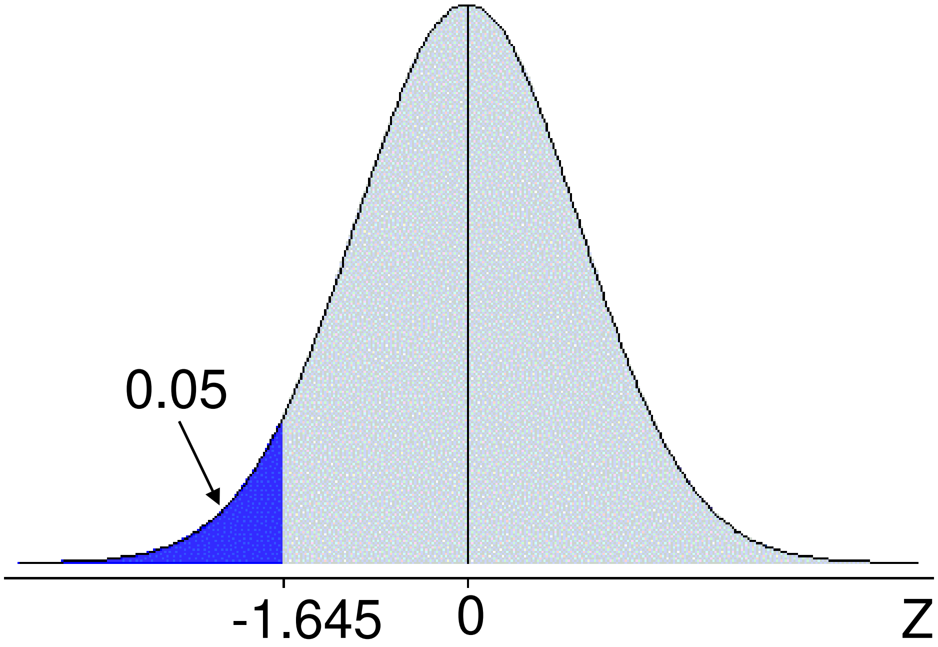
$H0$: $p_1 - p_2 \geq 0$

Ha: $p_1 - p_2 < 0$

$\alpha = 0.05$

$n_1 = 100$, $n_2 = 200$

We want probability in the tail to be less than 0.05, so Z < -1.645



Z-stat calculation

$$\hat{p}_1 = \frac{30}{100} = 0.3$$

$$\hat{p}_2 = \frac{85}{200} = 0.425$$

$$\hat{p} = \frac{30 + 85}{100 + 200} = \frac{115}{300} = 0.383$$

$$Z = \frac{(0.3\ - 0.425) - 0}{\sqrt{(0.383)(1 - 0.383)\left(\frac{1}{100} + \frac{1}{200}\right)}} = \frac{-0.125}{0.0595} = -2.1$$

Z-test for two proportions solution

Let p_i be the efficacy of drug i.

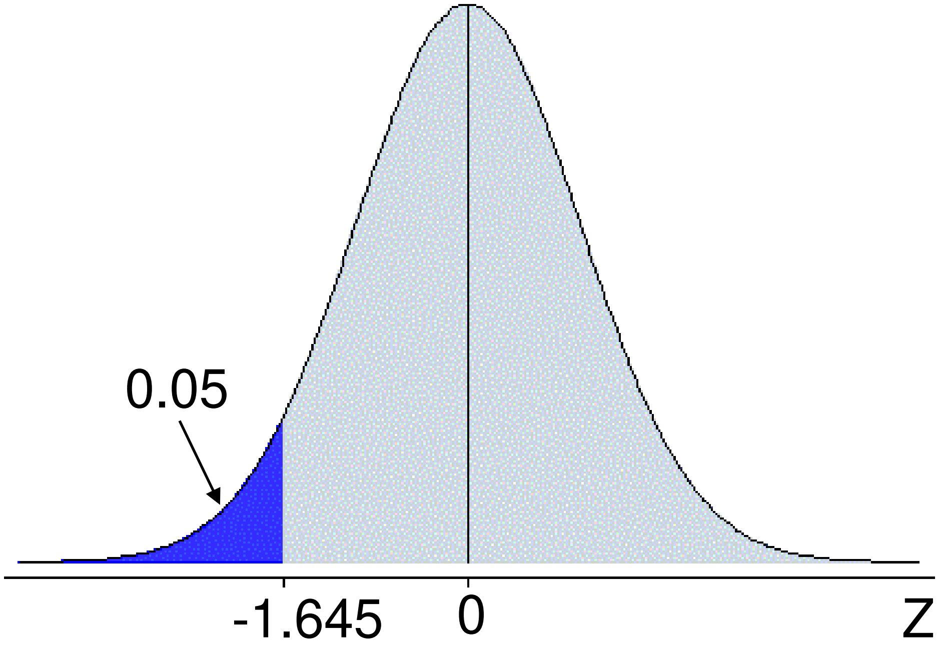
$H0$: $p_1 - p_2 \geq 0$; Ha: $p_1 - p_2 < 0$

$\alpha = 0.05$

$n_1 = 100$, $n_2 = 200$

Test statistic: Z = -2.1

Decision:
Z = -2.1 < -1.645
Reject at $\alpha = 0.05$



Conclusion:
Drug 2 is better than Drug 1

χ^2 Test for Independence

- Test the independence of two categorical variables
- Assumptions
  - The sampling method is simple random sampling.
  - The variables under study are each categorical.
  - If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5

- Hypothesis
  - H0: In the population, the two categorical variables are independent.
    Ha: In the population, two categorical variables are dependent

- Gather data and summarize in the two-way contingency table
  - Represents the observed counts and is called the **Observed Counts Table**

χ^2 Test for Comparing AV performance by weather

An AV manufacturer wants to know whether there is a statistically significant dependency between AV disengagement (first categorical variable) and the weather conditions while driving (second variable). For that, they count the number of miles driven with and without disengagement. A mile is counted as disengaged if there was at least one disengagement within it.

	Sunny	**Rainy**	**Snow**	total
Disengagement	28	59	34	121
No Disengagement	821	465	283	1569
total	849	524	317	1690

Steps in χ^2 Test for Independence

- Steps
  - Null and alternative hypotheses
  - Test statistic
  - Analyze sample data
  - P-value and interpretation
  - Significance level

Hypothesis Formulation

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.
  - H_0: the categorical variables, disengagement and weather, are independent
  - H_a: Categorical variables are not independent.
- Test Statistic: chi-square
- Significance level: 0.05

Analyze Sample Data

	Sunny	**Rainy**	**Snow**	total
Disengagement	28	59	34	121
No Disengagement	821	465	283	1569
total	849	524	317	1690

Contingency Table

Degrees of freedom (DF):

DF = $(r - 1)$*$(c - 1)$

where r and c are levels of categorical variables

Expected frequencies:

$$E_{rc} = \frac{n_r \times n_c}{n}$$

Test Statistic:

$$\chi^2 = \sum \frac{(O_{rc} - E_{rc})^2}{E_{rc}}$$

Analyze Sample Data

DF = (r - 1) * (c - 1) = (2 - 1) * (3 - 1) = 2

$E_{r,c} = (n_r * n_c)/n$

	Sunny	Rainy	Snow
Disengagement	60.8	37.5	22.7
No Disengagement	788.2	486.5	294.3

Expectation Table

$$\chi^2 = \sum \frac{(O_{rc} - E_{rc})^2}{E_{rc}}$$

$$= \frac{(28-60.8)^2}{60.8} + \frac{(59-37.5)^2}{37.5} + \frac{(34-22.7)^2}{22.7} + \frac{(821-788.2)^2}{788.2} + \frac{(524-486.5)^2}{486.5} + \frac{(283-294.3)^2}{294.3}$$

$$= 38.36$$

P-Value

- The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 38.86

$$P(\chi^2 \ > 38.86) = 4.7 \times 10^{-9}.$$

- p-value (4.7×10^{-9}) is less than the significance level (0.05)
  - ➢ Reject the null hypothesis.
  - ➢ Conclude that there is a relationship between AV disengagement and weather condition

Kolmogorov Smirnov Test

- Till now we have seen statistical tests that compare the means and/or proportions of distributions
- Not all distributions are completely summarized by their mean; therefore we need tests to compare distributions
- Suppose we have observations $X_1, X_2 \ldots X_n$ which we think come from a distribution P.
- The Kolmogorov-Smirnov (KS) Test is used for the following hypothesis test:

$$\mathrm{H_0}\text{: the samples come from } P$$
$$\mathrm{H_a}\text{: the samples do not come from } P$$

KS Test: Test statistic

The CDF uniquely characterizes a probability distribution

$$F(x) = P(X < x),$$

where $F(x)$ is the CDF if the samples are generated from the probability distribution P. Let,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[X_i \leq x]$$

where $F_n(x)$ is the empirical CDF that is calculated from the samples, and the indicator function

$$\mathbb{I}[X_i \leq x] = \begin{cases} 1, & X_i \leq x \\ 0, & X_i > x \end{cases}$$

The test statistic is:

$$D_n = \max_x |F_n(x) - F(x)|$$

KS Test: Critical value and Conclusion

At the 95% level, the critical value is approximately given by

$$D_{crit,0.05} = \frac{1.36}{\sqrt{n}}$$

Fail to reject the null hypothesis if,

$$D_n < D_{crit,0.05}$$

Intuition: If the samples are drawn from the probability distribution P, then the CDF and the empirical CDF would be "close" to each other.

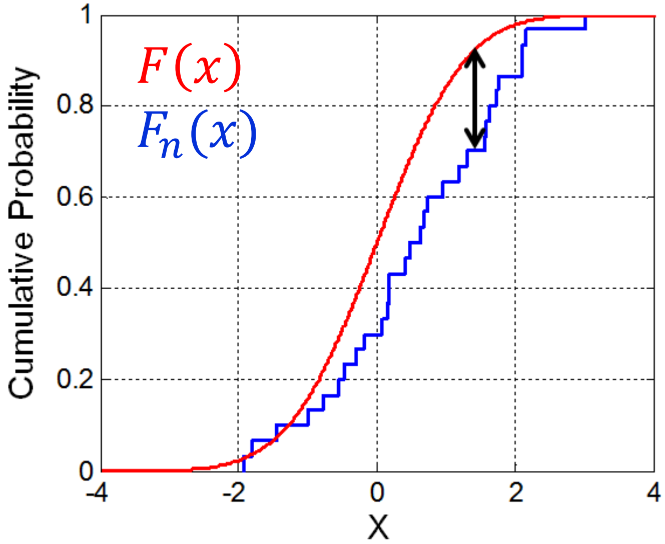


Image Source: https://commons.wikimedia.org/w/index.php?curid=25222928