Data Analytics 01
Preview & Review

Jakey BLUE
jakeyblue@ntu.edu.tw

Blended Learning Format

- Offline Video Learning on COOL
  - anytime/anywhere you want
- Physical Discussion Session
  - R402, Xinsheng Lecture Building (新生大樓)
  - 11~12h, every working Monday
- Office Hour Session
  - Your TAs: Zoey Chao (趙珮君); Landon (王懷葳)
  - RDV location/time to be determined

Planning

Date	Topics
02/20	Preview & Review
02/27	Regression Analysis
03/06	Regression Analysis
03/13	Multivariate Statistical Inference
03/20	Dimension Reduction
03/27	Partial Least Squares Regression
04/03	Big Data Infrastructure
04/10	Mid-term Exam
04/17	Supervised ML Algorithms
04/24	Supervised ML Algorithms
05/01	Unsupervised ML Algorithms
05/08	Unsupervised ML Algorithms
05/15	Machine Learning Techniques
05/22	Deep Neural Nets
05/15	Deep Neural Nets
06/05	Challenge Presentation Day

Prerequisites

- Fundamental Calculus
- Linear Algebra
- Programming: R or Python
  - R (4.x, RStudio, *.rmd and *.html)
  - Python (3.x, Jupyter Notebook, *.ipynb and *.pdf)
  - RStudio Cloud or Google Colab can be used for coding assignments
- Probability & Statistics
  - will be reexamined with Homework#1
- Understanding the following 6 questions

#0

Data science is about pythoning/coding with machine/deep learning packages?

#1 What is the gradient of $f(x)$ at x_0?

A. $\dfrac{f(x)}{x_0}$

B. $f'(x_0)$

C. $f(x_0) - f(0)$

D. $\dfrac{f(x-x_0)}{x_0}$

#2 What is true for $\mathbf{X}_{n\times p}$?

A. The column space of $\mathbf{X}$ is in n-dimensional space.

B. The column space of $\mathbf{X}$ is in p-dimensional space.

C. If $n \gg p$, the column space of $\mathbf{X}$ is more likely consisted of n bases.

D. If $n \gg p$, the column space of $\mathbf{X}$ is more likely consisted of p bases.

#3 Given $\mathbf{X}_{n\times p}$ with n samples and p variables, How to know if the variables are dependent?

A. $\mathbf{tr}(\mathbf{X}) = 0$

B. $\mathbf{rank}(\mathbf{X}) = p$

C. $\mathbf{X}^T\mathbf{X}$ is positive definite

D. $\left|\frac{1}{n-1}\mathbf{X}^T\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\mathbf{X}\right| = 0$

#4 What is true about the # of unknown parameters?

A. Normal Dist. → 3

B. Poisson Dist. → 2

C. Exponential Dist. → 1

D. Uniform Dist. → 1

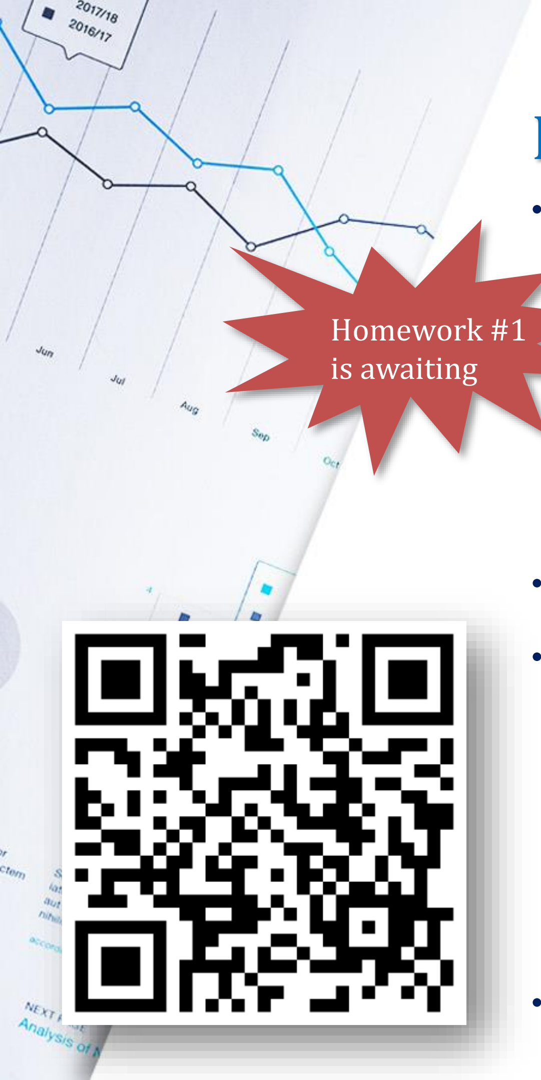
#5 What is true for $f(x) = x^2 - x - 1$?

A. It has the maximal value at $x = 0.5$

B. It has the minimal value at $x = \frac{1-\sqrt{5}}{2}$

C. It has the maximal value at $x = \frac{1+\sqrt{5}}{2}$

D. It has the minimal value at $x = 0.5$



Homework #1 is awaiting

Evaluation

- Homework: 25% (≤ 10 times)
  - writing exercises (pdf only), coding exercise in R (*.rmd and *.html), in Python (*.ipynb and *.pdf)
  - unless specified, each writing exercise costs 10 points.
  - code grading policy, each coding assignment costs 15 points.
    - fulfill basic requirements: 15pts
    - result presentation: 2pts
    - discussion & remarks: 3pts
  - late penalty: 10% off per day and no later than 7 days of delay.
  - plagiarism leads to 0’s for both copies.

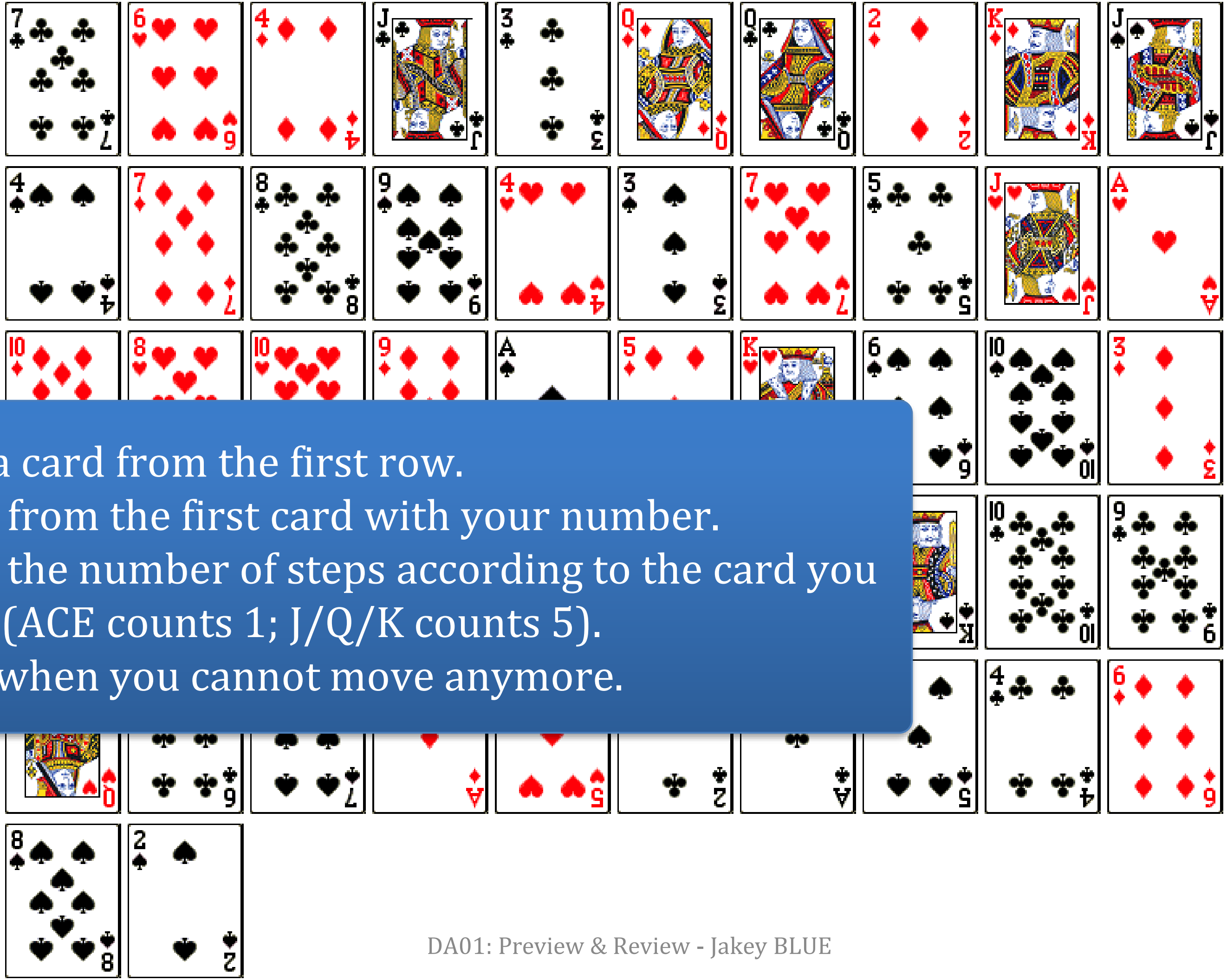
- Mid-term (writing exam): 35% **(past exams will NOT be provided)**

- Team Challenge: 37% = 20% (Ranking) + 12% (Presentation) + 5% (Report)
  - 2 or 3 in one team (> 3 ➔ project score is discounted)
  - team with mixed nationality, project score is promoted
  - ranking is weekly announced
  - oral presentation
    - describe the work breakdown at the beginning
    - each team member presents
    - peer review, **100% presence in the presentation session for everyone (obligatory)**
  - slides uploaded to COOL as the final report, revise it if necessary

- Participation/Typo Hunting: (3%)
  - Each typo found in the slides is graded 0.1 point directly to the final score.

A Probable Magic

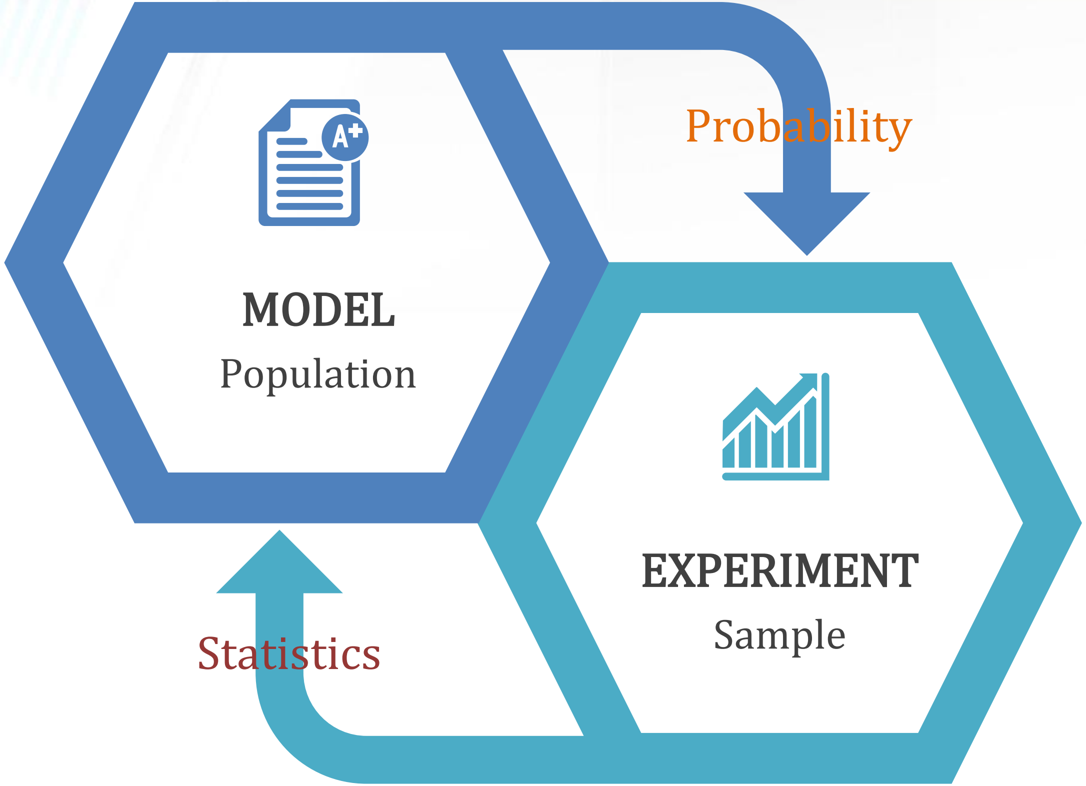
ALL ROADS LEAD TO ROMA

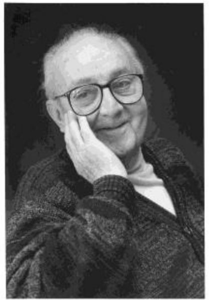


1. Pick a card from the first row.
2. Walk from the first card with your number.
3. Walk the number of steps according to the card you stop. (ACE counts 1; J/Q/K counts 5).
4. Stop when you cannot move anymore.

ARE THERE THEORIES?

The Relation between Probability & Statistics





George E. P. Box (1919-2013)

ALL MODELS ARE WRONG, BUT SOME ARE USEFUL.

Revisit Probability & Statistics

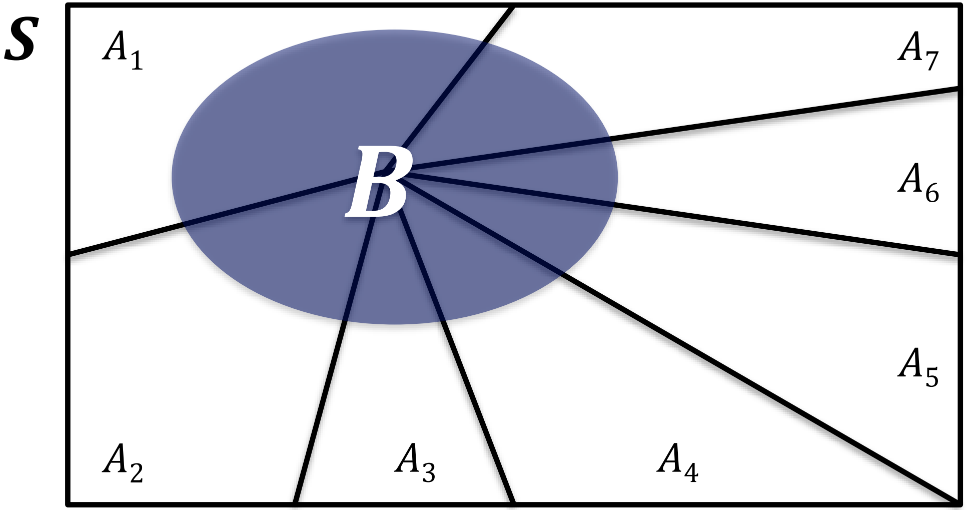
- Probability
  - Law of Total Probability ➔ Bayes’ Theorem
  - Random Distributions ➔ Central Limit Theorem (CLT)

- Statistics
  - Descriptive Statistics
  - Statistical Inference

Law of Total Probability

- Let $A_1, \dots, A_n$ be mutually exclusive and exhaustive events. Then for any other event B,

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$



Thomas Bayes
1701-1761

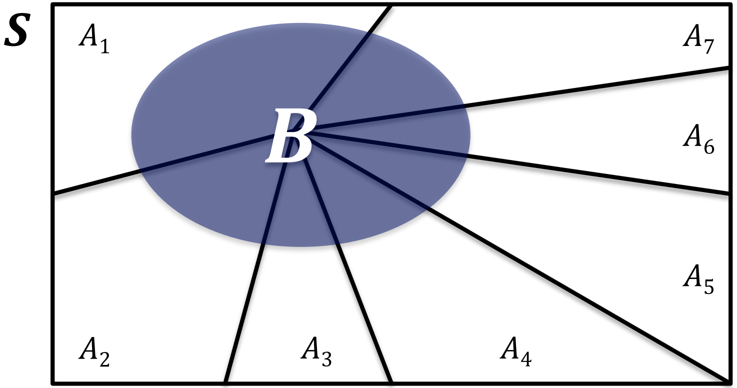
Bayes' Theorem

- Let $A_1, \ldots, A_n$ be a collection of n mutually exclusive and exhaustive events with $P(A_i) > 0$ for $i = 1, \ldots, n$. Then for any other event B with $P(B) > 0$

$$\boxed{P(A_k|B)} = \frac{P(A_k \cap B)}{P(B)} = \frac{\boxed{P(B|A_k)P(A_k)}}{\sum_{i=1}^{n} \boxed{P(B|A_i)P(A_i)}}$$

posterior knowledge

prior knowledge



The Monty Hall Problem

You have one chance to change, will you?

- One door with big prize, the other two with goats.
- The host knows exactly by heart where is the prize.
- You pick door A, the host then open one of the two unchosen doors with goat.

Monty Hall Problem (Given that you choose door A)

- A_p: door A has the prize → $P(A_p) = \frac{1}{3} = P(B_p) = P(C_p)$
- B_g: host opens door B with a goat → $P(B_g) = \frac{1}{2}$, HOW?
  - Case 1: door A with prize → $P(B_g|A_p) = \frac{1}{2}$
  - Case 2: door C with prize →$P(B_g|C_p) = 1$
  - $P(B_g) = {}^1/_2 \times {}^1/_3 + 1 \times {}^1/_3 = {}^1/_2$

- The probability of door A with the prize while the host opens door B with a goat → $P(A_p|B_g) = ?$
  - $P(A_p|B_g) = \frac{P(A_p \cap B_g)}{P(B_g)} = P(B_g|A_p) \times \frac{P(A_p)}{P(B_g)} = \frac{1}{3}$
  - $P(C_p|B_g) = \frac{P(C_p \cap B_g)}{P(B_g)} = P(B_g|C_p) \times \frac{P(C_p)}{P(B_g)} = \frac{2}{3}$

Monty Hall Problem (Given that you choose door A)
A Contingency Table View

	Door B Opened	Door C Opened	Sum
Door A with Prize	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$
Door B with Prize	0	$\frac{1}{3}$	$\frac{1}{3}$
Door C with Prize	$\frac{1}{3}$	0	$\frac{1}{3}$
Sum	$\frac{1}{2}$	$\frac{1}{2}$	1

Random Variable (RV, R.V., r.v.)

- For a given sample space of some experiments, a random variable is any rule that associates a number with each outcome in the sample space. A random variable is always denoted by a capital letter (e.g. X, Y, etc.).

- Types:
  - discrete → if the set of possible values is discrete
  - continuous → if the set of possible values is an entire interval of numbers

Important Distributions & Their Moments

- Discrete
  - Bernoulli
  - Geometric
  - Binominal
  - Poisson
- Continuous
  - Exponential
  - Uniform
  - (Standard) Normal
  - t
  - χ^2
  - F

$$E[X] = \mu_x = \sum_{x \in D} x \cdot p(x) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

$$V[X] = \sigma_x^2 = \sum_{x \in D} (x - \mu)^2 \cdot p(x) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx$$

When 2 Random Variables Meet: Joint Distribution

$Y = X_1 + X_2$

- $E[Y] = ?,\ V[Y] = ?$

$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, X_i \overset{i.i.d.}{\sim} (\mu, \sigma^2)$

- $E[\overline{X}] = ?,\ V[\overline{X}] = ?$

- Convolution of 2 Random Variables: $f_{X+Y}(a) = \int_{-\infty}^{\infty} f_x(a-y) f_y(y) dy$

Central Limit Theorem (CLT)

- If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2. Let $\overline{X}$ denote the sample mean, the limiting form of the distribution of

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

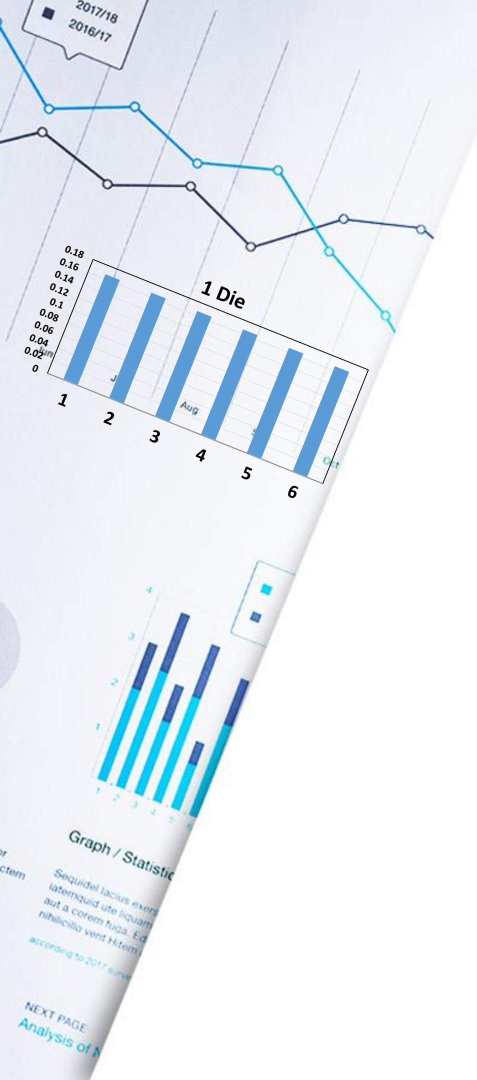
as $n \to \infty$, is the standard normal distribution.

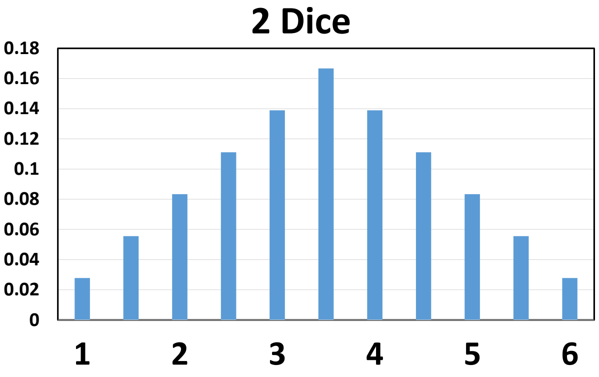
1	2	3	4	5	6
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

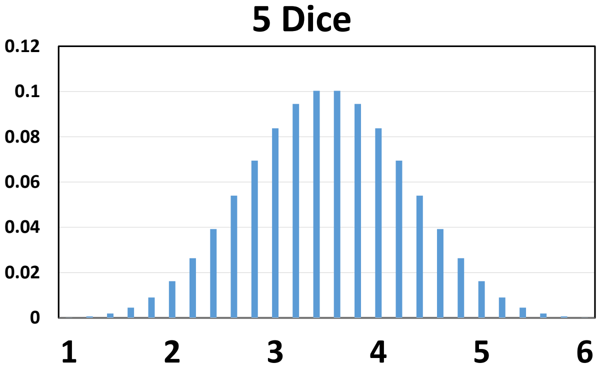
	1	2	3	4	5	6
1	1	1.5	2	2.5	3	3.5
2	1.5	2	2.5	3	3.5	4
3	2	2.5	3	3.5	4	4.5
4	2.5	3	3.5	4	4.5	5
5	3	3.5	4	4.5	5	5.5
6	3.5	4	4.5	5	5.5	6

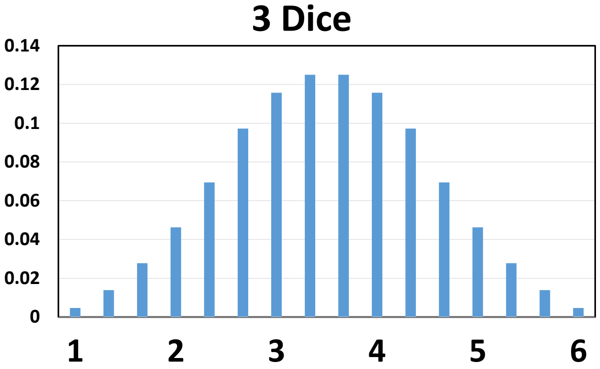
1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6
$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

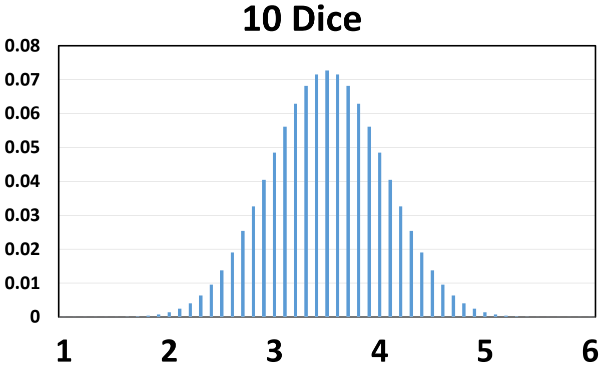
The Averages of Rolling *n* Dices



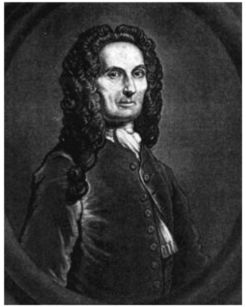








CLT is also known as de Moivre–Laplace Theorem



Abraham de Moivre
(1667-1754)

Pierre-Simon marquis de Laplace
(1749-1827)

- discovered by French.
- regarded as First Principle in Probability (unofficially)
- is the fundamental of mathematical statistics
- key to the parametric estimation, hypothesis testing

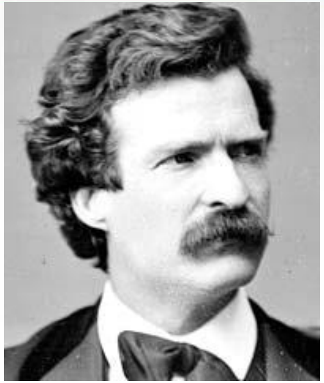
https://www.youtube.com/watch?v=6YDHBFVIvIs

“Probability” in Summary

- Fundamental Probability Concepts
  - Event Relations: Union; Intersection; Complement
  - Event Types: Mutually Exclusive; Exhaustive; Independent
  - Event Operations: Permutation; Combination
  - Law of Total Probability
- Bayes’ Theorem
- Random Variable and Its Distribution
  - Mean (Expected Value) and Variance: $V[X] = E[X^2] - (E[X])^2$
  - Discrete Distribution: Bernoulli; Geometric; Binominal; Poisson
  - Continuous Distribution: Uniform; Normal; Standard Normal; Exponential; χ^2; Gamma;
- Relationships of Multiple R.V.
  - Joint Distribution: $p(X = x\ , Y = y)$; $f(X = x, Y = y)$
  - Marginal Distribution: $p_X(x) = \sum_y p(x, y)$; $f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$
  - Dependence of 2 R.V.: covariance σ_{XY}; correlation ρ_{XY}
  - Convolutional Distribution: $f_{X+Y}(a) = \int_{-\infty}^{\infty} f_x(a - y)f_y(y)dy$
- Central Limit Theorem

There are three kinds of lies:

LIES, DAMN LIES, STATISTICS



Mark Twain

Benjamin Disraeli

The Prestige

	Made	Attempt	%
2P	10.8	21.2	51%
3P	0.5	1.7	29%
FG	11.3	22.9	49%

	Made	Attempt	%
2P	4.1	7.9	52%
3P	1.8	4.7	38%
FG	5.9	12.6	47%

Simpson's Paradox

To describe the data (sample)

SUMMARIZATION
VISUALIZATION

Random Sampling and Summarization

- To measure the central tendency: sample mean
  - SAMPLE Mean of a set of numbers $x_1, x_2, \dots, x_n$ is given by

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- To measure the dispersion: sample variance
  - SAMPLE Variance of the set $x_1, x_2, \dots, x_n$ of numerical observations, denoted by s^2 is given by

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Degrees of Freedom (ν, d.f., DoF)

- Why it is $n - 1$ in the denominator of sample variance formula?
  - The number of independent pieces of information.
  - The number of values free to vary in calculation of a statistic.
- $\bar{x} = 10, x_1 = 5, x_2 = 15$, can you calculate the sample variance s^2?

Sample Relationship

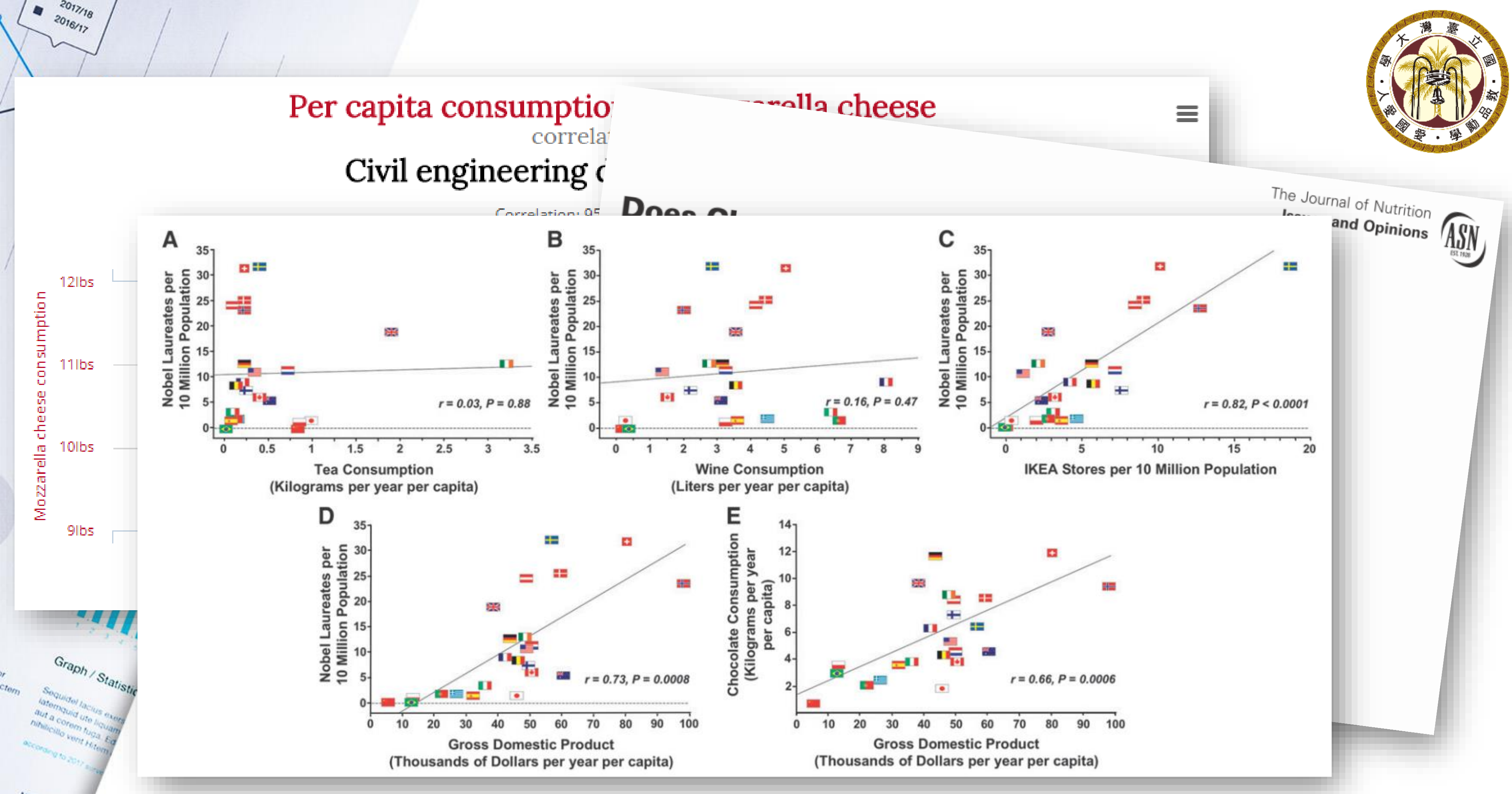
- Sample Covariance

$$\mathrm{cov}(x,y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

- Sample Correlation: (Pearson’s correlation coefficient)

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- Correlation doesn’t imply Causality



Per capita consumptio… …rella cheese

correla…

Civil engineering d…

Correlation: 95…

Does Ch…

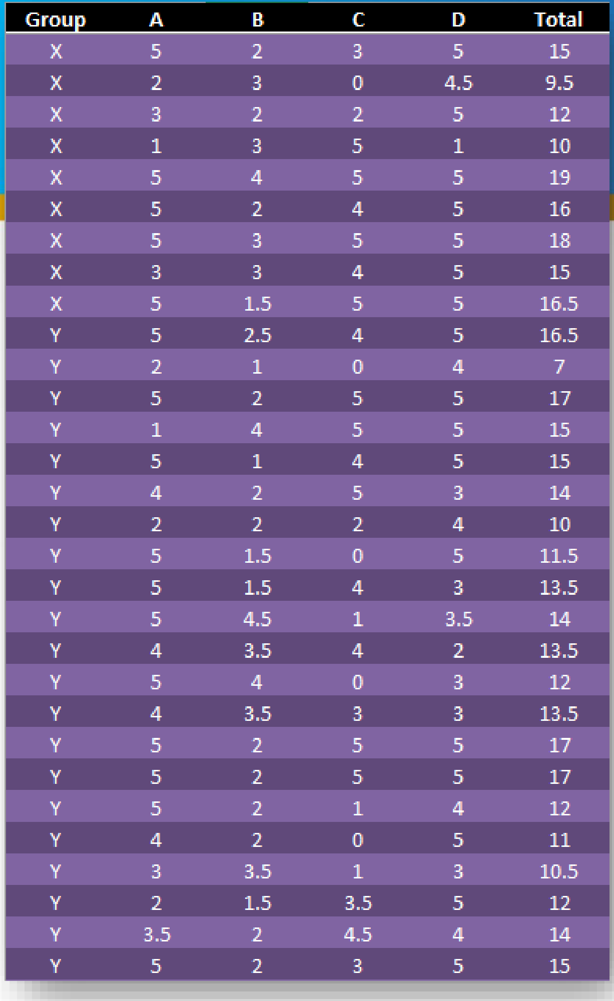
The Journal of Nutrition

Iss… and Opinions

Grades of Probability

- 72 students are tested.
  - 4 exercises $\{A, B, C, D\}$
  - Courses are done in 3 groups $\{X, Y, Z\}$
  - 5 points/question

- Questions to ask?
  - averages of the class, groups?
  - variations of the class, groups?
  - correlations among the exercises?
  - behavior among 3 groups?



Group	A	B	C	D	Total
X	5	2	3	5	15
X	2	3	0	4.5	9.5
X	3	2	2	5	12
X	1	3	5	1	10
X	5	4	5	5	19
X	5	2	4	5	16
X	5	3	5	5	18
X	3	3	4	5	15
X	5	1.5	5	5	16.5
Y	5	2.5	4	5	16.5
Y	2	1	0	4	7
Y	5	2	5	5	17
Y	1	4	5	5	15
Y	5	1	4	5	15
Y	4	2	5	3	14
Y	2	2	2	4	10
Y	5	1.5	0	5	11.5
Y	5	1.5	4	3	13.5
Y	5	4.5	1	3.5	14
Y	4	3.5	4	2	13.5
Y	5	4	0	3	12
Y	4	3.5	3	3	13.5
Y	5	2	5	5	17
Y	5	2	5	5	17
Y	5	2	1	4	12
Y	4	2	0	5	11
Y	3	3.5	1	3	10.5
Y	2	1.5	3.5	5	12
Y	3.5	2	4.5	4	14
Y	5	2	3	5	15

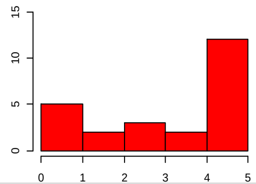
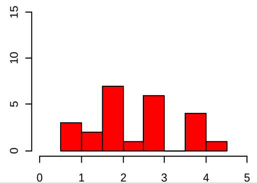
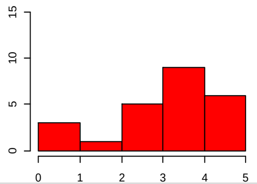
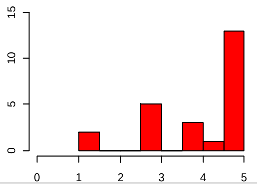
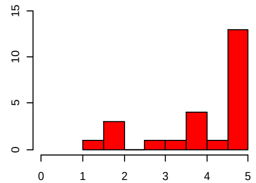
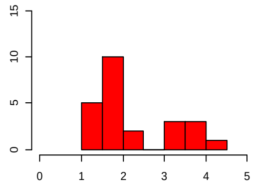
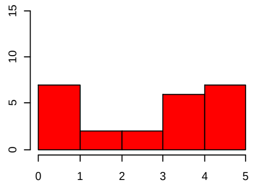
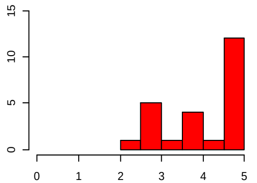
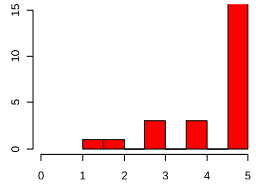
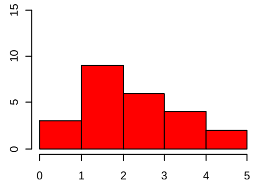
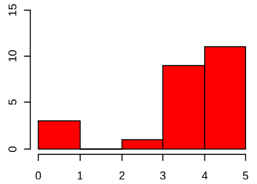
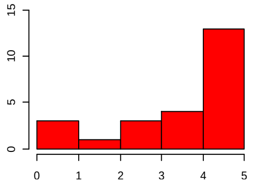
MetaData

$\bar{x}/s$	*A*	*B*	*C*	*D*	Average / Stdev.
Group X	3.5 / 1.8	2.5 / 1.1	3.5 / 1.4	4.1 / 1.3	13.6 / 3.3
Group Y	4.1 / 1.2	2.4 / 1.0	3.0 / 1.9	4.2 / 0.9	13.7 / 2.7
Group Z	4.3 / 1.1	2.5 / 1.2	3.9 / 1.6	3.9 / 1.5	14.6 / 3.8
Average / Stdev.	4.0 / 1.4	2.5 / 1.1	3.4 / 1.7	4.1 / 1.3	14.0 / 3.3

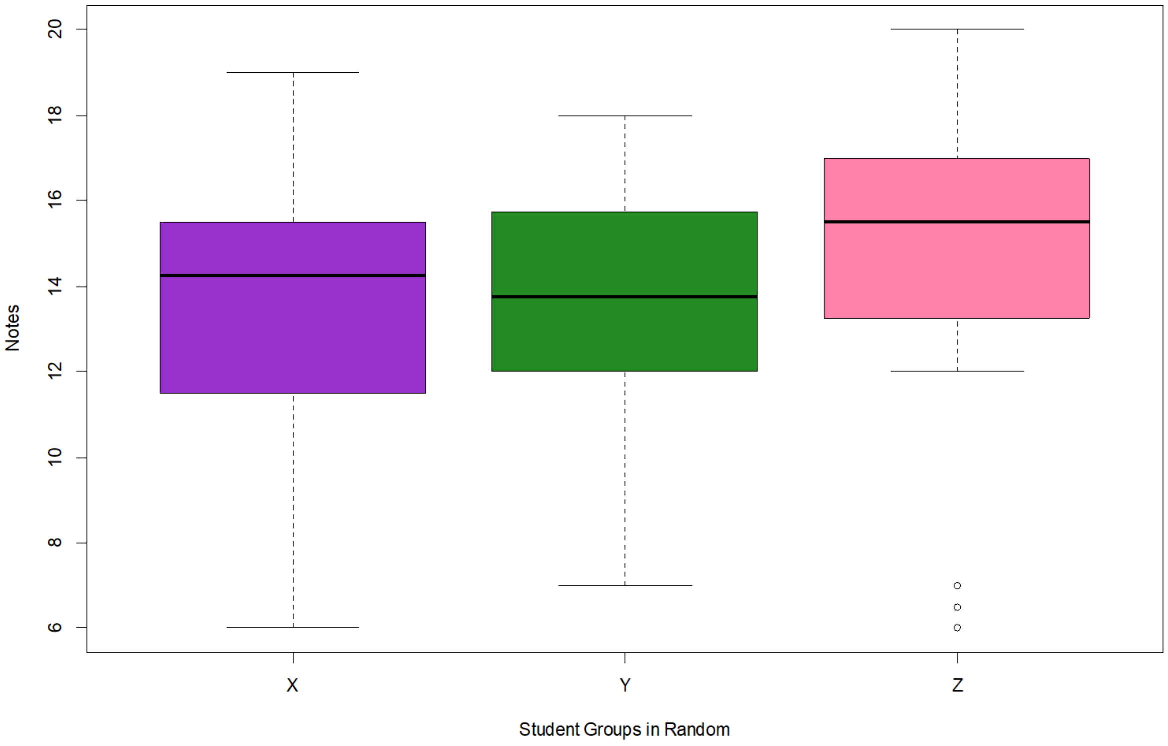
Sample Correlation Matrix

Correlation Coefficient (r)	A	B	C	D
A	1	-0.03	0.20	0.20
B	-0.03	1	0.14	0.01
C	0.20	0.14	1	0.24
D	0.20	0.01	0.24	1

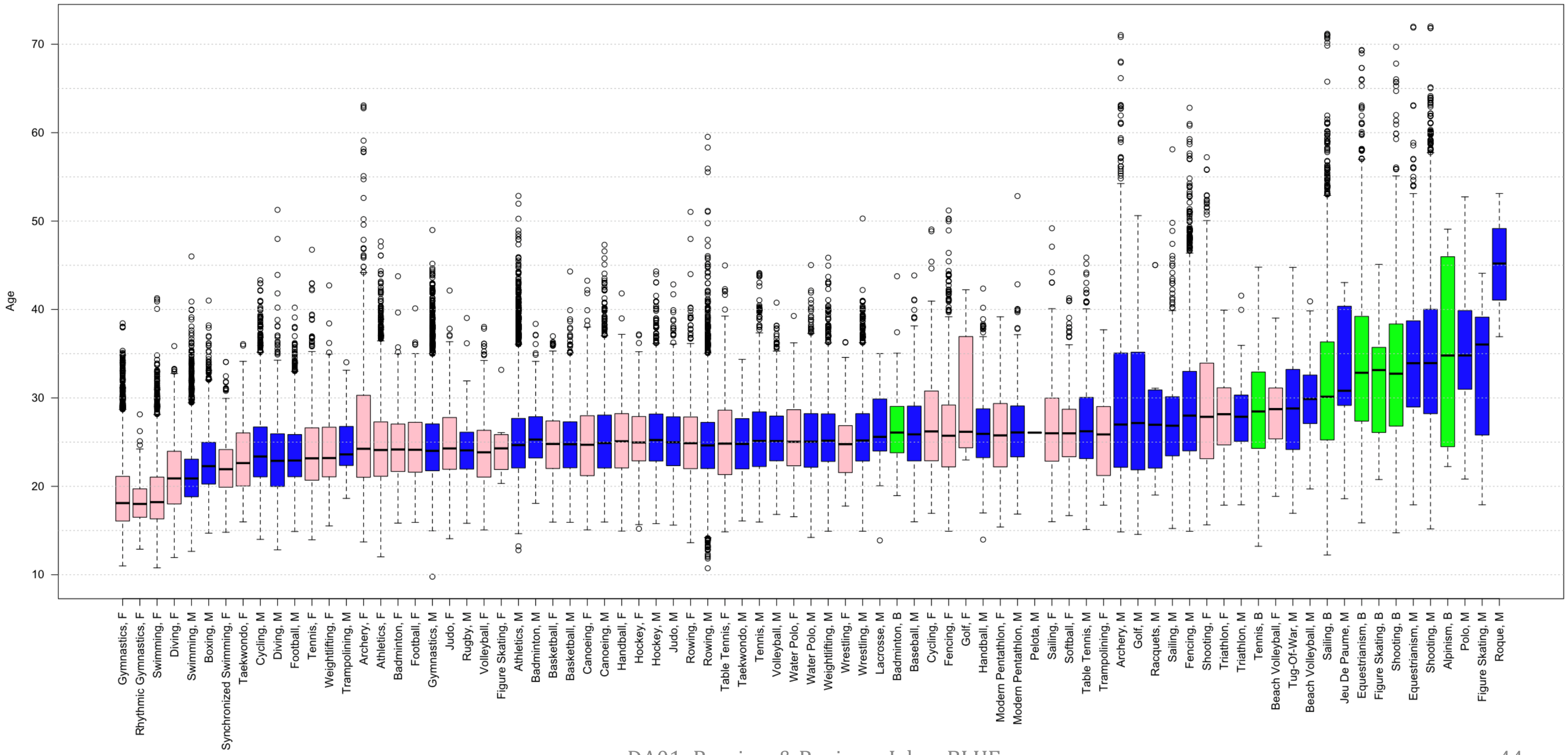
Grade Distributions by Group×Exercise

	A	*B*	*C*	*D*
Group X				
Group Y				
Group Z				

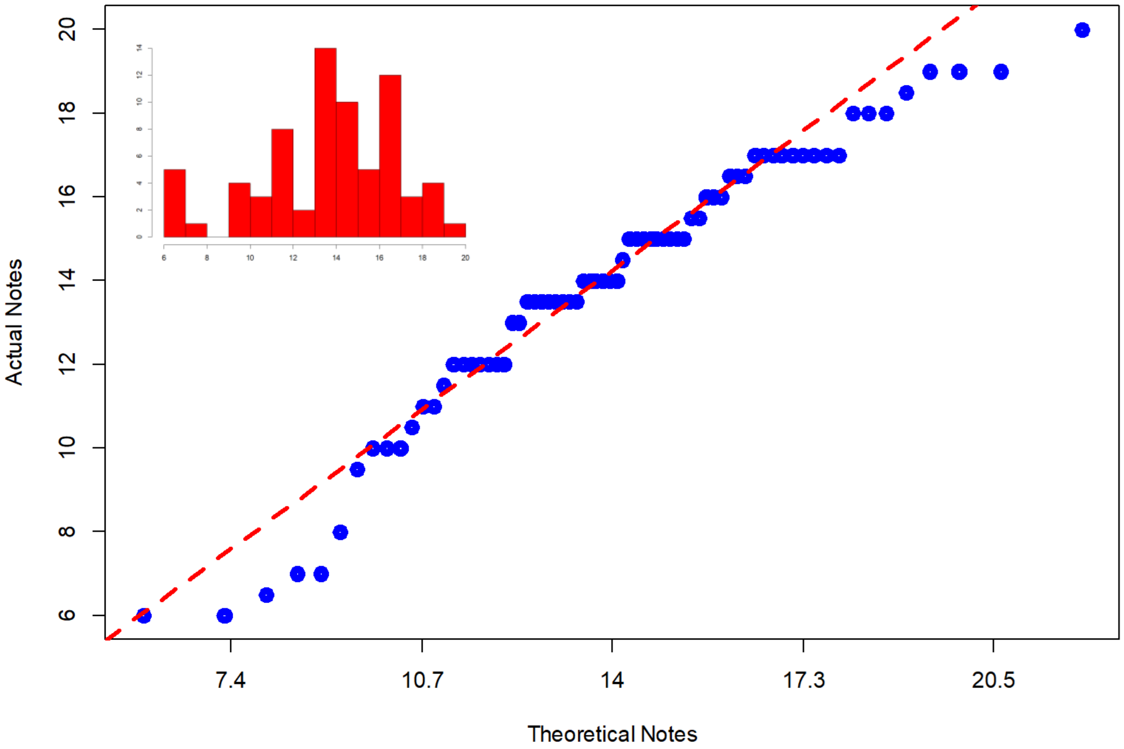
Comparing the Grades among 3 Groups



Age distribution of Olympic Athletes by Sport and Gender: All-time
Female = Pink, Male = Blue, Both = Green



Are the grades “Normally” distributed? Q-Q Plot



Descriptive Statistics in Summary

- Sample Summarization
  - Sample Mean (Average); Sample Variance
  - Degrees of Freedom
  - Sample Covariance; Sample Correlation

- Data Visualization
  - Meta Data; Correlation Matrix
  - Histogram
  - BoxPlot
  - Probability Plot (Q-Q Plot)

Statistical Inference

Jakey BLUE

Statistic(s)

- A statistic is any function of the random variables constituting one or more samples, provided that the function does not depend on any unknown parameter values
  - for examples: sample mean, sample variance
- Sample data:
  - A **sample** = A set of sample observations $[x_1, x_2, \dots, x_i, \dots, x_n]$ and **sample size** $= n$
  - A sample **observation** = A piece of data vector $\boldsymbol{x}_{\boldsymbol{i}} = [x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{im}]$

What Can We INFER?

- **Point Estimate**
  - To estimate the parameters of the probability models with sample data.
  - To evaluate how good the estimators are.

- **Hypothesis Testing**
  - To check/test whether the model parameter(s) has changed.
  - To evaluate how good the tests are (two types errors?).

- **Modeling of Statistics for Performance Evaluation**
  - Model sample observations as “random variables”.
  - Statistic is then a function of random variables and is also a random variable.

It is actually more than one point.

POINT ESTIMATE

What is a Point Estimate?

- A point estimate of a parameter θ is a single number that can be regarded as the most plausible value of θ.

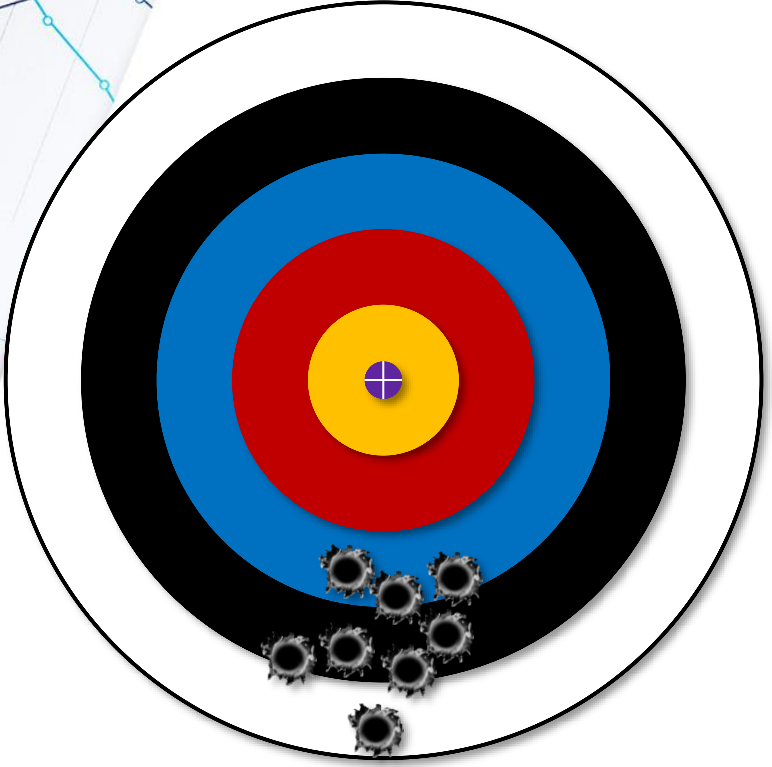
- A point estimate is obtained by selecting a suitable statistic and computing its value from the given sample data.

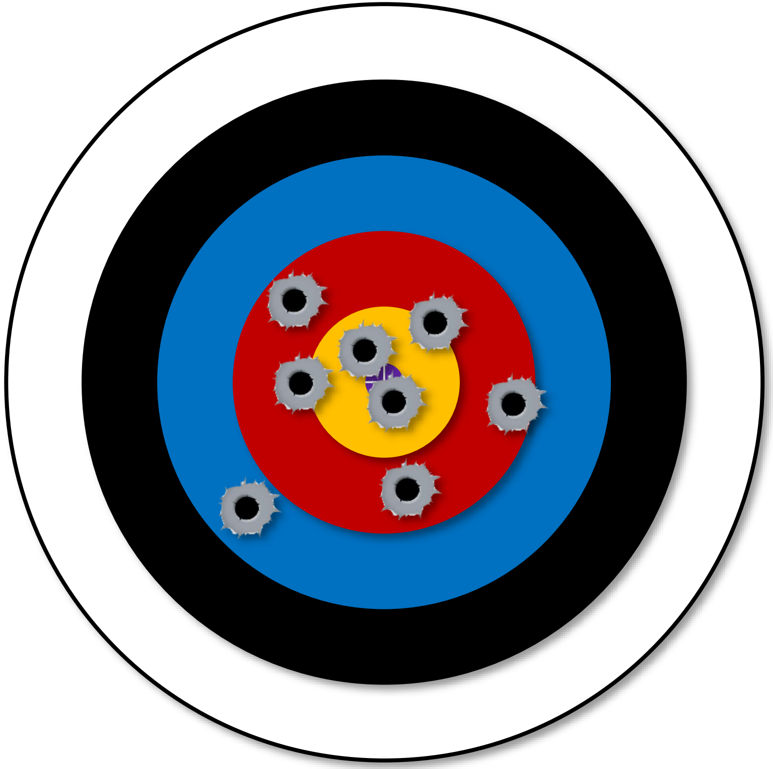
- The selected statistic is called the point estimator of $\theta \leftarrow \hat{\theta}$.

- A point estimator is itself a random variable with a distribution.

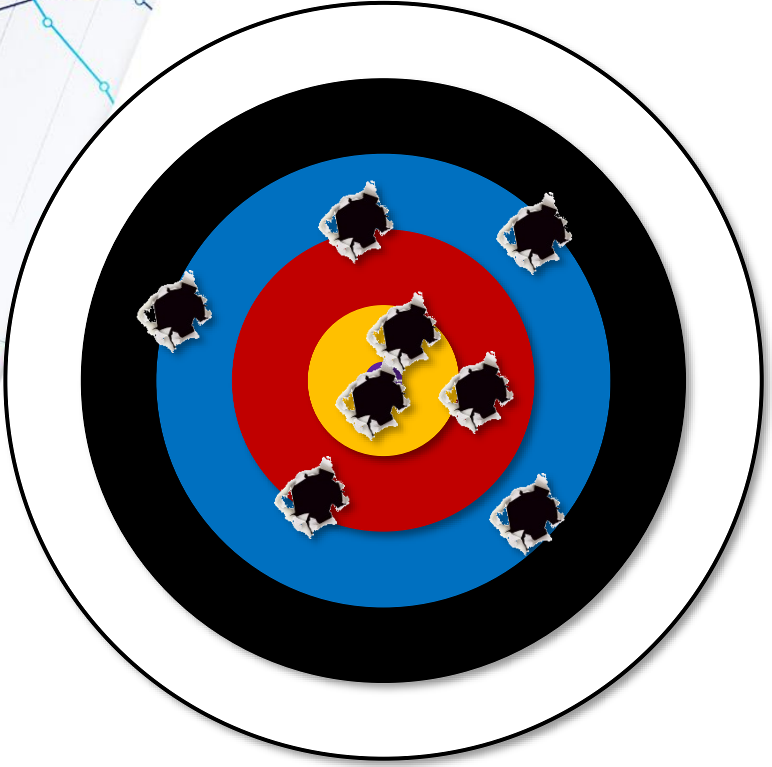
$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

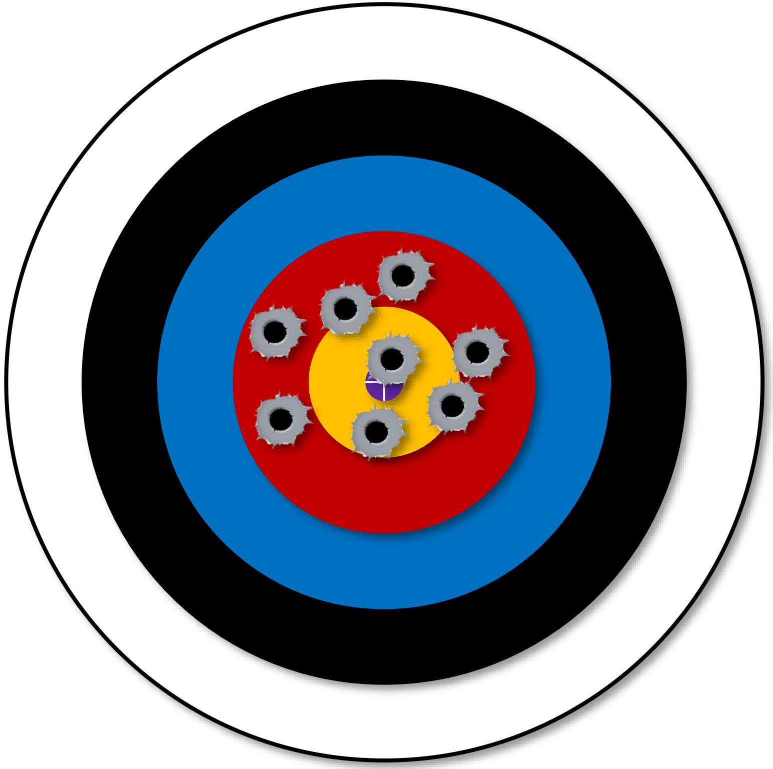
Which one is more ACCURATE?





Which one is more PRECISE?





How can we say if an estimate is good?

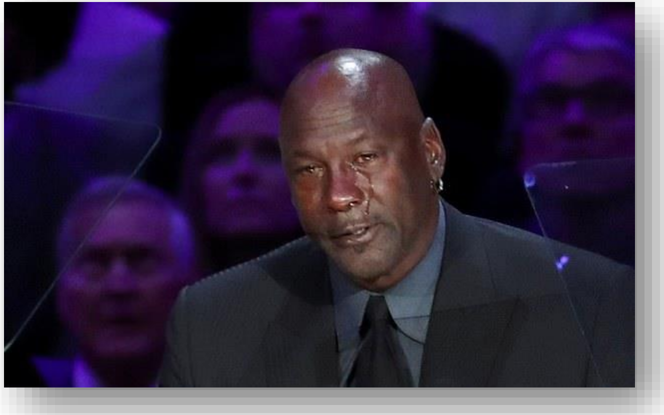
- On target? → Unbiased
- Very sure? → Minimum variance
- Minimum Variance Unbiased Estimator (MVUE)
  - Among all the unbiased estimators, the one with the minimum variance.
- Example: sample mean is a MVUE for normally distributed populations.
- However, sometimes a biased estimator is preferable to the MVUE. Why?

Different Point Estimates of **Mean**

- Point estimates: $\bar{X}, \tilde{X}, \bar{X}_e, \bar{X}_{tr(m)}$
  - $\bar{X}$ is the arithmetic average called sample mean.
  - $\tilde{X}$ is the median that is the center observation of the entire sample.
  - $\bar{X}_e$ is the extreme mean (an average of two extreme observations).
  - $\bar{X}_{tr(m)}$ is a trimmed mean that trims $m\%$ of observations from each end of the sample.

Is the Arithmetic Average $\overline{X}$ the Best?

- In 1998, the University of North Carolina at Chapel Hill made a statistics census on the income of its graduates.
  - Graduates from the Department of Cultural Geography earns most not only in NCSU but also among whole US.



Common Methods of Point Estimate

- **Moment Estimator**
  - Raw Moments: $m_k = E\left[X^k\right], k = 1, 2, \ldots, \infty$
  - Central Moments: $E\left[(X - \mu_X)^k\right]$
  - Let $M_X(t) = E[e^{tX}] = \begin{cases} \sum_x p(x)e^{tx} \\ \int_{-\infty}^{\infty} f(x)e^{tx}dx \end{cases}$ be the Moment Generating Function.
    - $m_k = M_X^{(k)}(0)$, e.g., $m_1 = M_X'(0) = \mu,\ m_2 = M_X''(0) = E[X^2]$

Maximum Likelihood Estimate (MLE)

- Let $X_1, X_2, \ldots, X_n$ are independent random sample observations from a population following an identical probability model with likelihood function $P(X)$ or $f(X)$.
  - The joint likelihood for $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ is:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X = x_1)P(X = x_2) \cdots P(X = x_n)$$
$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2) \cdots f(x_n)$$

- MLE is the estimate of a parameter that maximizes the joint likelihood function is maximized.

MLE for μ of Normal Distribution

- Let X follow a (μ, σ^2) normal distribution and μ is unknown.
  - We take a sample of n observed values $x_1, x_2, \dots, x_n$.
  - the joint likelihood function:

$$f(x_1, \dots, x_n|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)} = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\sigma^n} e^{\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right)}$$

- Maximizing $f(x_1, \dots, x_n)$ is equivalent to maximizing log $f(x_1, \dots, x_n)$. Take the derivative of log$f(x_1, \dots, x_n)$ and set it to zero:

$$\frac{d}{d\mu} \log f(x_1, \dots, x_n|\mu) = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})}{\sigma^2} = 0 \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

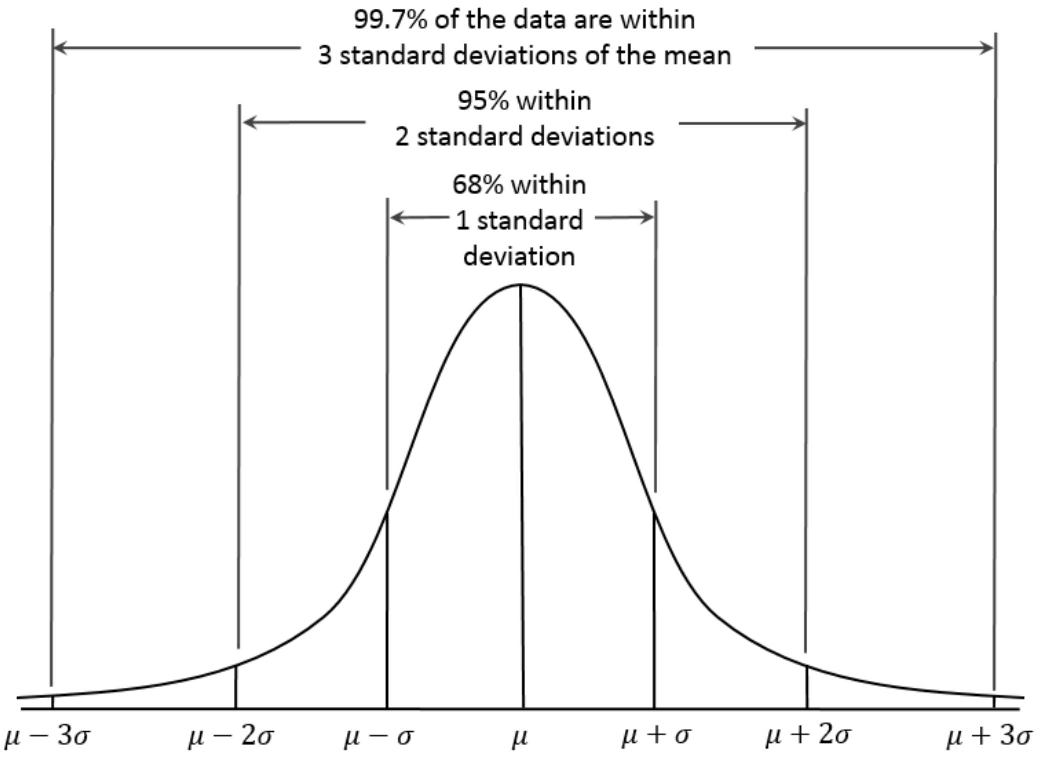
An extension to Unsupervised Learning, explained later.

MLE → EM (Expectation Maximization)

Confidence Region/Interval

POINT ESTIMATE HAS A RANGE.

THE Very Much Useful “Interval”

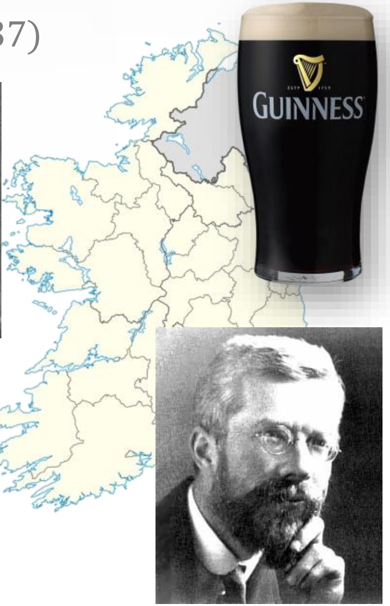


Student *t* Distribution

$$f_v(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\frac{1}{\left(\frac{(1+x^2)}{\nu}\right)^{\frac{\nu+1}{2}}}$$

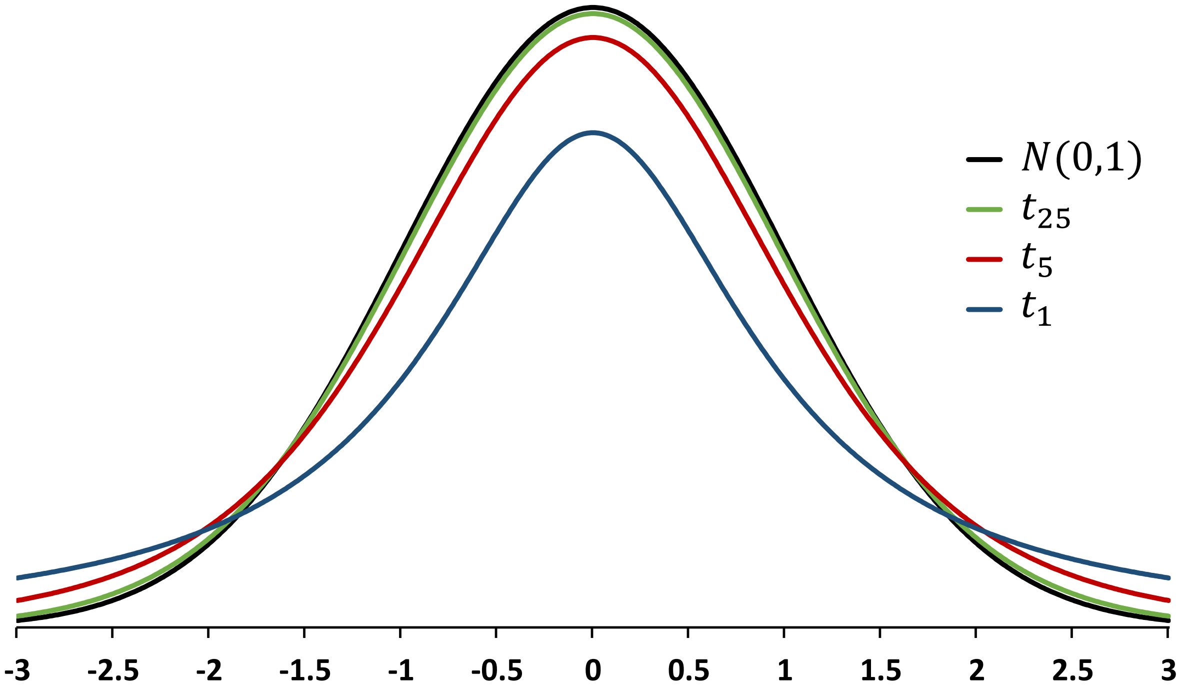
- Bell-shaped and centered at 0 → very similar to Normal.
- $\boldsymbol{\nu}\uparrow$ distribution spread $\downarrow$
- The distribution spreads wider than the normal distribution (heavier tails).
- $\boldsymbol{\nu}\to\infty$, $\boldsymbol{t_\nu}\to$ Standard Normal $\boldsymbol{N(0,1)}$.

William Sealy Gosset (1876-1937)



Ronald Aylmer Fisher (1890-1962)

Student t Distribution



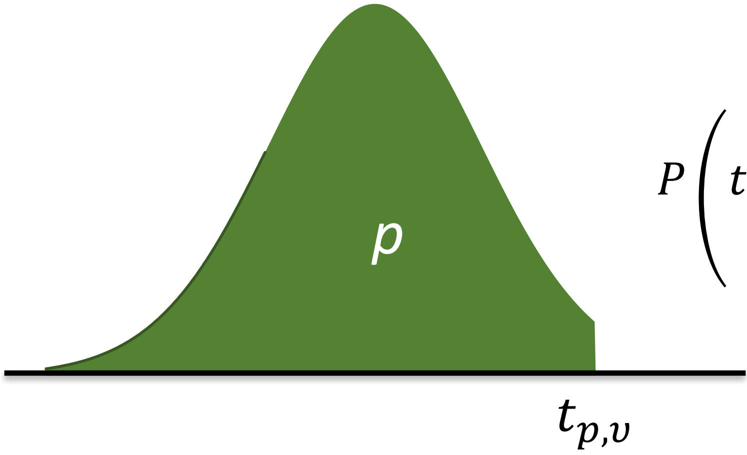
t statistics: C.I. for **Unknown σ**

- $\bar{X}$ is the average of a random sample of size n from a normal distribution with mean μ . Then, the random variable

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

follows a probability distribution called t distribution with $n - 1$ degrees of freedom.

Confidence Interval Using t Statistic



$$P\left(t_{\frac{\alpha}{2},\upsilon} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{1-\frac{\alpha}{2},\upsilon}\right) = 1-\alpha$$

- Then a $100(1-\alpha)\%$ confidence interval for μ is: $\left(\bar{x} - t_{1-\frac{\alpha}{2},\upsilon}\frac{s}{\sqrt{n}}, \bar{x} - t_{\frac{\alpha}{2},\upsilon}\frac{s}{\sqrt{n}}\right)$.

H_0 vs. H_a

HYPOTHESIS TESTING

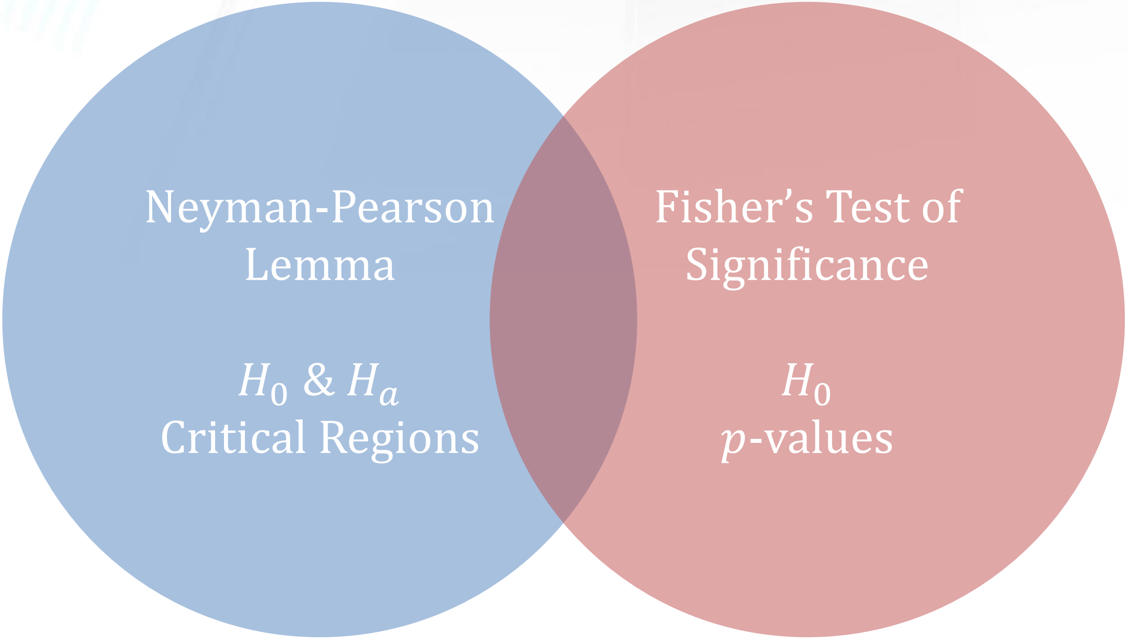
Testing?! Test What?

- Motivation
  - To reject an initial claim and to statistically prove that a scientific effort really makes differences
- Example: medical experiments, pool results, social science experiments
- Initial claim
  - **Null hypothesis $\boldsymbol{H_0}$**
- Claim otherwise
  - Alternative hypothesis H_1 or H_a

The Debates on Hypothesis Testing

Jerzy Neyman
(1894-1981)

Egon Pearson
(1895-1980)



Neyman-Pearson Lemma

H_0 & H_a
Critical Regions

Fisher’s Test of Significance

H_0
p-values

Ronald Fisher
(1890-1962)

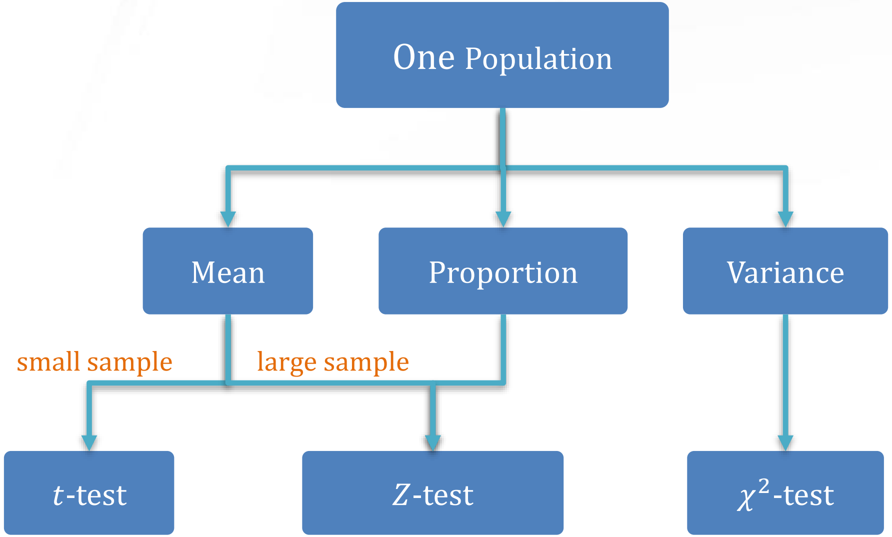
There is a Test, There are Errors!

- Type I error: rejecting the null hypothesis H_0 when it is true
  - Probability of type I error, α.
- Type II error: not rejecting H_0 when H_0 is false
  - Probability of type II error, β.

	H_0 is TRUE	H_0 is FALSE
Reject H_0	Type I Error, α False Positive	BINGO! True Negative
Do Not Reject H_0	BINGO! True Positive	Type II Error, β False Negative

		He is really innocent.	He is guilty.
	H_0: The prisoner is innocent!	H_0 is TRUE	H_0 is FALSE
I think he is guilty.	Reject H_0	INJUSTICE!!	Got YOU!
OK, he is innocent.	Do Not Reject H_0	You are free to go.	At large.

One Population Hypothesis Testing



One Population

Mean

Proportion

Variance

small sample

large sample

t-test

Z-test

χ^2-test

Common Hypothesis Tests

Purpose	Sample Statistics	Critical Region	Condition
population mean (μ)	$\overline{x}$	$\pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$	X is normally distributed and σ is known; or $n \geq 30$
population mean (μ)	$\overline{x}$	$\pm t_{1-\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$	$n < 30$; and/or σ unknown
population proportion (p)	$\hat{p}$	$\pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$n\hat{p}$ & $n(1-\hat{p}) \geq 10$
difference of two population means ($\mu_1 - \mu_2$)	$\overline{x}_1 - \overline{x}_2$	$\pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	X_1, X_2 are normally distributed or $n_1, n_2 \geq 30$; σ_1, σ_2 are known
difference of two population means ($\mu_1 - \mu_2$)	$\overline{x}_1 - \overline{x}_2$	$\pm t_{1-\frac{\alpha}{2}, n_1+n_2-1} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$	$n_1, n_2 \leq 30$; and/or σ_1, σ_2 are unknown
difference of two population proportions ($p_1 - p_2$)	$\hat{p}_1 - \hat{p}_2$	$\pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$n\hat{p}$ & $n(1-\hat{p}) \geq 10$ for the two groups

Proceeding a Hypothesis Testing

1. Define the Null Hypothesis, H_0

2. Find the Test Statistic: a function of the sample data on which the decision (reject H_0 or not) is to be based. Try to think about we actually turn the whole data into a value.

3. Set the Critical Value and reject region based on the distribution of the test statistic under H_0 and the Type I error probability α.

4. H_0 will then be rejected if and only if the observed or computed test statistic values falls in the reject region.

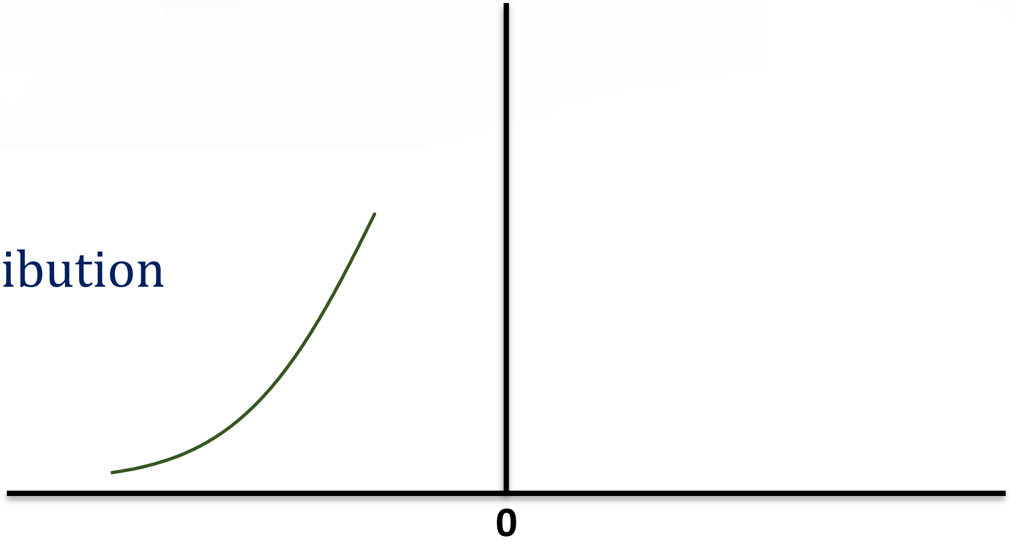
Mean Test while σ UNKNOWN ➔ t Test

- $H_0: \mu = \mu_0$
  $H_1: H_0$ is false, or, $\mu \neq \mu_0$

  Test statistic: t-test $= \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$

- Distribution under H_0: t-Distribution with $\nu = n - 1$.

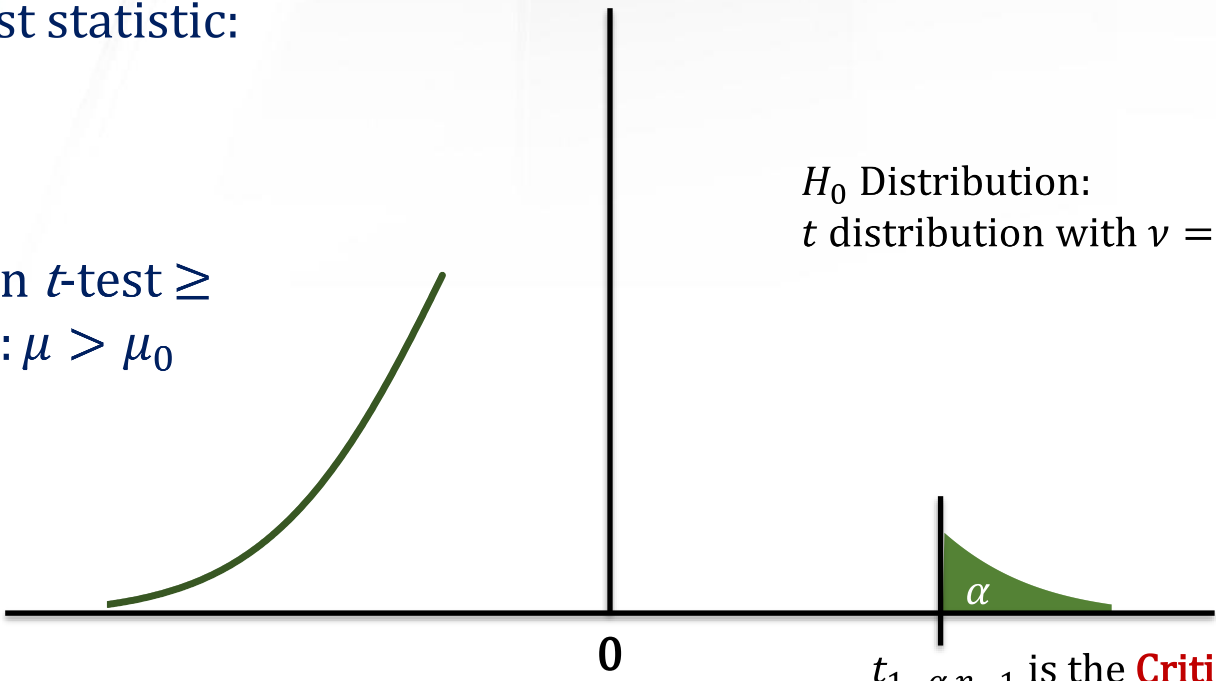
$\mu = \mu_0 \Rightarrow \mathrm{E}(t) = 0$



t Test Criteria to Reject H_0: Reject Region

- $H_0: \mu = \mu_0$, Test statistic:
  $t\text{-test} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

- Reject H_0 when t-test $\geq$ $t_{1-\alpha,n-1} \Rightarrow H_1: \mu > \mu_0$



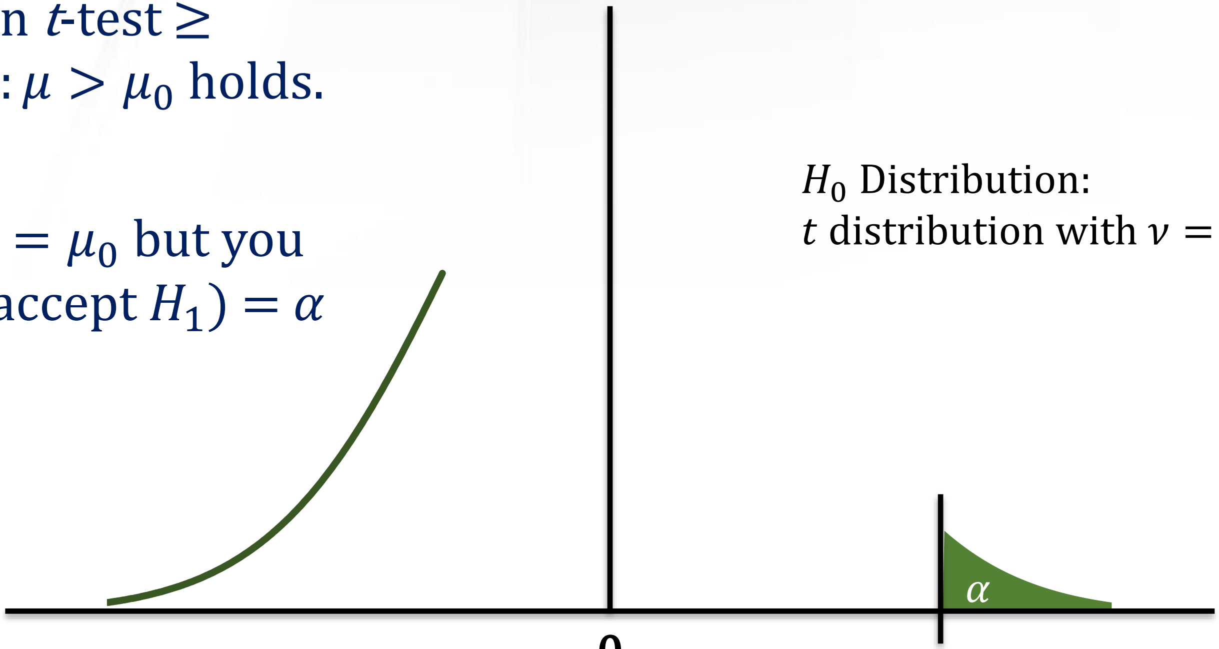
H_0 Distribution:
t distribution with $\nu = n - 1$

$t_{1-\alpha,n-1}$ is the **Critical Value**

t Test Type I Error Probability of Rejecting H_0

- Reject H_0 when t-test $\geq t_{1-\alpha,n-1} \Rightarrow H_1: \mu > \mu_0$ holds.

- Probability ($\mu = \mu_0$ but you reject H_0 and accept H_1) $= \alpha$



H_0 Distribution:
t distribution with $\nu = n - 1$

$t_{1-\alpha,n-1}$ is the **Critical Value**

Meaning of p-value (R. Fisher's Test of Significance)

- p-value defines the probability of getting an "UNEXPECTED/EXTREME SAMPLE" given that H_0 is assumed to be true. (Fisher's Test of Significance)

- Once the p-value of the sample dataset has been calculated, the testing conclusion at a given significance level α can be made by comparing the p-value with α.
  - CAUTION: Fisher's Test of Significance does not define/require any α, which exists only in the Neyman-Pearson Lemma as the Type I error.

Using another distribution for hypothesis testing

χ^2 TEST

χ^2 Test for WHAT?

- Proposed by Karl Pearson (correlation coefficient) in 1900.
- Used to test the population properties other than the parameters.
  - Goodness of Fit (to quantify Q-Q Plot)
    - if the population is following certain distribution
  - Test of Independence
    - if two random variables are independent
  - Test of Homogeneity
    - if two or more than two populations are from the same distribution

Goodness of Fit (GoF)

- Compare the “observed frequency”, O_i, with “expected frequency”, E_i in a sample data set. The Test Statistics is

$$C = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{n-k-1}$$

  - where n is the number of samples and k is the number of unknown parameters in the population distribution.

- Hypothesis Testing

  H_0: The sampling data set is following the distribution

  H_1: H_0 is not true

[Example] Are the tastes of customers the same?

- A company wants to know if its 3 products with different flavors create different preferences. 120 customers are interviewed.

Product	A	B	C
Preferred #	35	42	43

  - What does it mean that the 3 products make no preference?
  - What are the expected frequencies?

[Example] Are the tastes of customers the same?

- Let p_i denote the preference rate of product i.

$$H_0: p_A = p_B = p_C = \frac{1}{3}$$

$$H_1: p_A \neq p_B \neq p_C$$

Product	A	B	C
O_i	35	42	43
E_i	40	40	40

$$C = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = \frac{(35-40)^2 + (42-40)^2 + (43-40)^2}{40} = 0.95$$

$$\chi^2_{v=3-1,\alpha=0.05} = 5.99$$

[Example] Is it a POISSON distribution?

- The times of French people who ever visited Asia is said to follow a Poisson Distribution, is it true?

times been to Asia	0	1	2	≥3
O_i	32	12	6	0
E_i	?	?	?	?

  - How do we calculate the “expected frequency”?
  - Remember $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$?

[Example] Is it a POISSON distribution?

H_0: It is following Poisson distribution

H_1: H_0 is not true

- Firstly we need to estimate $\hat{\lambda} = \frac{1\times12+2\times6}{50} = 0.48$
  - $P_i = P(X = i) = \frac{e^{-\hat{\lambda}}\hat{\lambda}^i}{i!}, i = 0, 1, 2$

times been to Asia	0	1	2	≥3
O_i	32	12	6	0
P_i	0.62	0.30	0.07	0.01
E_i	30.94	14.85	3.56	0.65

$$C = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 2.89 < \chi^2_{v=4-1-1,\alpha=0.05} = 5.99$$

Test of Independence

- The paired random variables can be arranged in a $r \times c$ Contingency Table.

$X \setminus Y$	1	2	…	c	row total
1	O_{11}	O_{12}	…	O_{1c}	R_1
2	O_{21}	O_{22}	…	O_{2c}	R_2
⋮	⋮	⋮	⋱	⋮	⋮
r	O_{r1}	O_{r2}	…	O_{rc}	R_r
column total	C_1	C_2	…	C_c	n

- Hypothesis Testing

  $H_0: X, Y$ are independent;

  $H_1: X, Y$ are dependent

Where to find E_{ij}'s? Find p_{ij} first!

- According to H_0, if X and Y are independent
  $p_{ij} = p_i \times p_j$ where $i = 1, \dots, r$ and $j = 1, \dots, c$.
- What are p_i and p_j?
  p_i are the ratios of row total to n
  p_j are the ratios of column total to n

$X \setminus Y$	1	2	…	c	row p_i
1	p_{11}	p_{12}	…	p_{1c}	$p_1 = R_1/n$
2	p_{21}	p_{22}	…	p_{2c}	$p_2 = R_2/n$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	p_{r1}	p_{r2}	…	p_{rc}	$p_r = R_r/n$
column p_j	$p_1 = C_1/n$	$p_2 = C_2/n$	…	$p_c = C_c/n$	1

Therefore, $E_{ij} = p_{ij} \times n$

- With O_{ij} and E_{ij}, we can again use χ^2 test.

$$\text{Test Statistics: } C = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

[Example] Does the education level impact the supporting rate?

O_{ij} vs. E_{ij}	high school	bachelor degree	master or higher	total
support	25 vs. 32	30 vs. 24	25 vs. 24	80
not support	35 vs. 28	15 vs. 21	20 vs. 21	70
total	60	45	45	150

$$E_{11} = p_{ij} \times n = p_i \times p_j \times 150 = \frac{60}{150}\frac{80}{150} \times 150 = 32$$

$$C = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \quad (\sim\chi^2_{(r-1)(c-1)})$$

$$= \frac{(25-32)^2}{32} + \frac{(30-24)^2}{24} + \frac{(25-24)^2}{24} + \frac{(35-28)^2}{28} + \frac{(15-21)^2}{21} + \frac{(20-21)^2}{21} = 6.58$$

$$\chi^2_{\nu=(2-1)(3-1),\,\alpha=0.05} = \chi^2_{\nu=2,\alpha=0.05} = 5.99$$

“Statistics” in Summary

- Point Estimate
  - Accuracy vs. Precision
  - Unbiased Estimate
  - Maximum Likelihood Estimate (MLE)
  - $(1-\alpha)$ Confidence Interval
  - Student t-Distribution
- Hypothesis Testing
  - H_0 vs. H_1
  - t-Test on the mean level
- χ^2 Test
  - Goodness of Fit
  - Test of Independence
  - Test of Homogeneity