Statistical Analysis (II)

Raghuram Bommaraju

Assistant Professor of Marketing

Indian School of Business

My Background

ISB

cmi CHENNAI MATHEMATICAL INSTITUTE

STATISTICAL INDIAN INSTITUTE
भिन्नेष्वैक्यस्य दर्शनम्
UNITY IN DIVERSITY

1949 EXCELLENCE & INTEGRITY
XLRI
Xavier School of Management
For the greater good

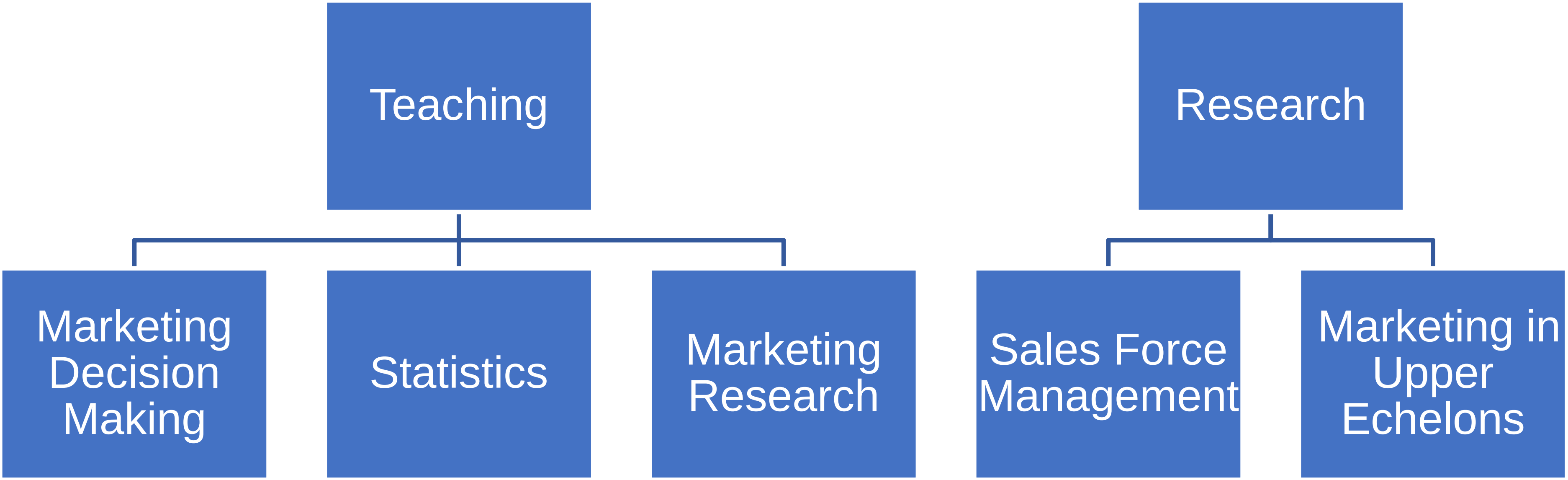
UH

Industries

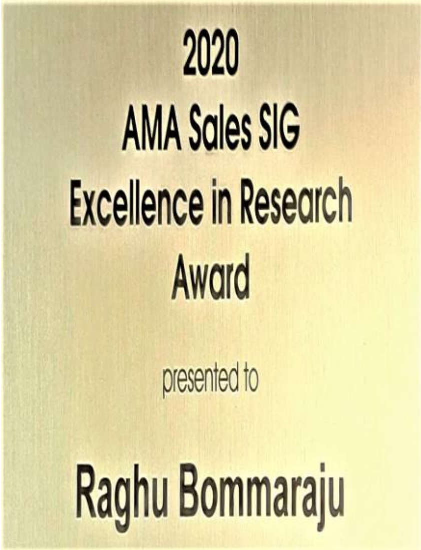
In God we trust, all others must bring data.

Functional Areas

Teaching and Research



Awards and Honors



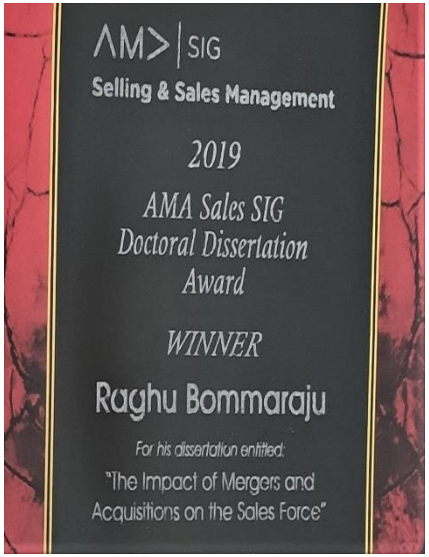
2020

AMA Sales SIG

Excellence in Research Award

presented to

Raghu Bommaraju



AM> | SIG

Selling & Sales Management

2019

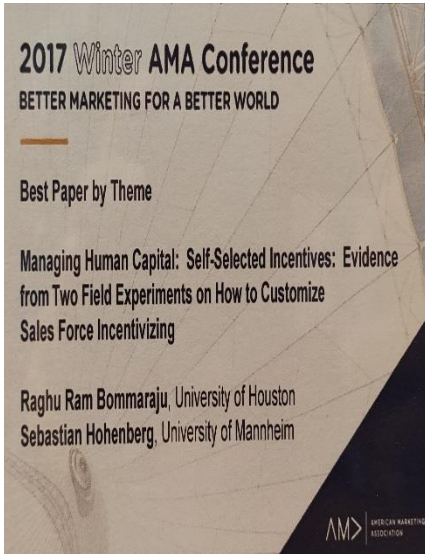
AMA Sales SIG Doctoral Dissertation Award

WINNER

Raghu Bommaraju

For his dissertation entitled:

"The Impact of Mergers and Acquisitions on the Sales Force"



2017 Winter AMA Conference

BETTER MARKETING FOR A BETTER WORLD

Best Paper by Theme

Managing Human Capital: Self-Selected Incentives: Evidence from Two Field Experiments on How to Customize Sales Force Incentivizing

Raghu Ram Bommaraju, University of Houston

Sebastian Hohenberg, University of Mannheim

AM> AMERICAN MARKETING ASSOCIATION

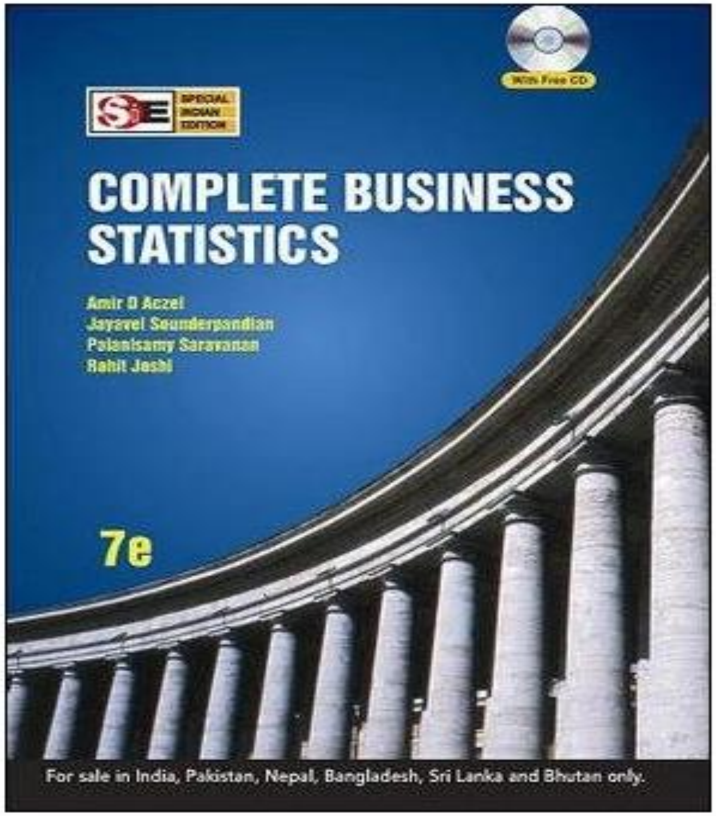
2nd Most Productive Author Among My Peers

(who finished PhD in 2017)

Outline of the Course

- Session 1 – Two Sample Comparisons
- Session 2 – Case Study
- Session 3 – Regression
- Session 4 – Regression
- Session 5 – Regression

Book for the Course


With Free CD
S E SPECIAL INDIAN EDITION
COMPLETE BUSINESS STATISTICS
Amir D Aczel
Jayavel Sounderpandian
Palanisamy Saravanan
Rohit Joshi
7e
For sale in India, Pakistan, Nepal, Bangladesh, Sri Lanka and Bhutan only.

- Textbook is much more exhaustive than what we will cover in five sessions

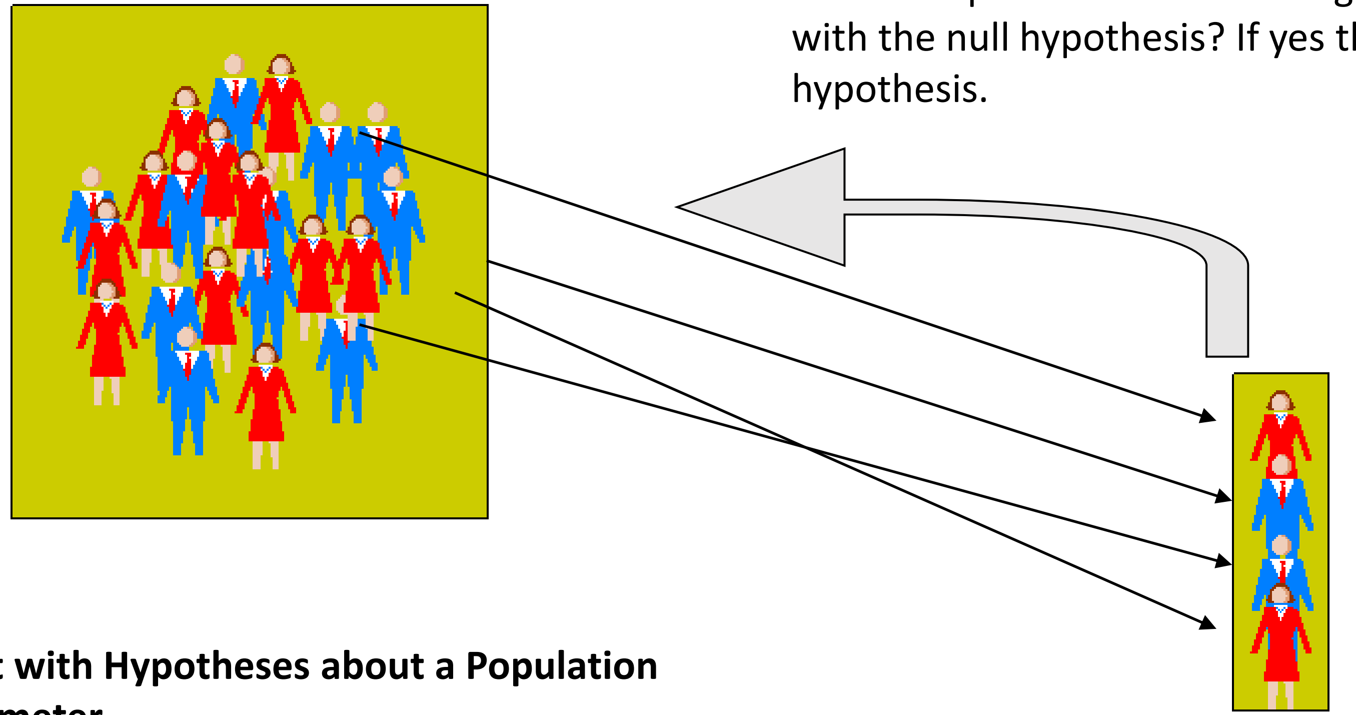
- The best use of the book is as a reference, go to specific sections (<u>given in the course syllabus</u>) of chapter where you need more clarity

- First solve the exercises from the textbook before thinking of more practice problems

Topics for the Session

- Recap of Stat 1
- Comparison Between Two Groups (10 mins Break at 80 minutes)
- ANOVA
- Chi-square Test (if time permits)

Hypothesis Testing Process



1. **Start with Hypotheses about a Population Parameter**
   Parameter could be mean, proportion or something else.

2. **Collect Sample Information**
   Collect information from a randomly chosen sample and calculate the appropriate sample statistic.

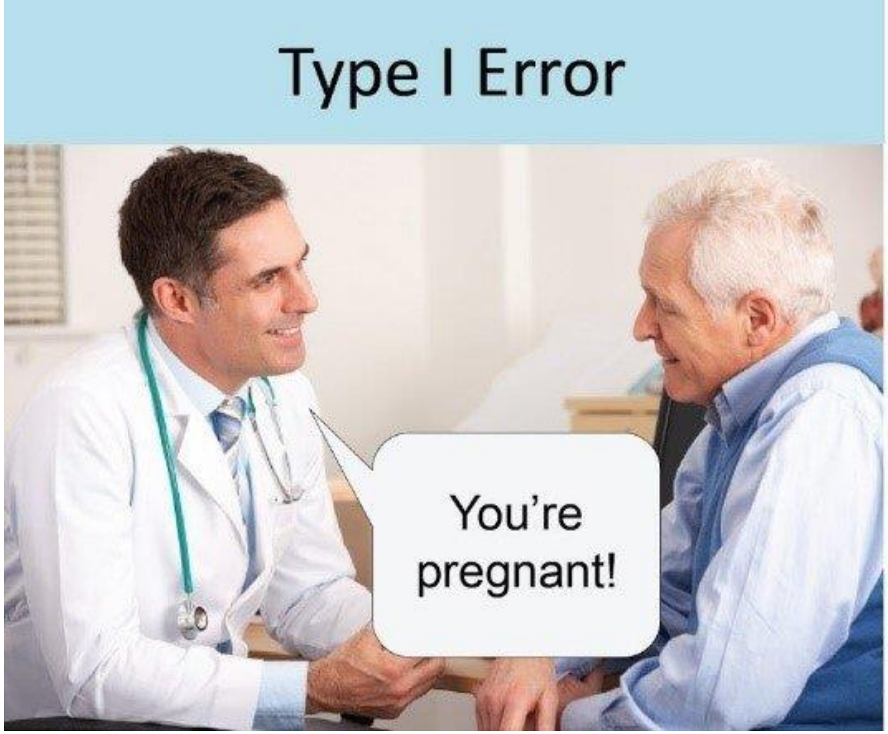
3. **Reject/Do Not Reject Hypothesis**
   Is the sample information strongly inconsistent with the null hypothesis? If yes then the reject hypothesis.

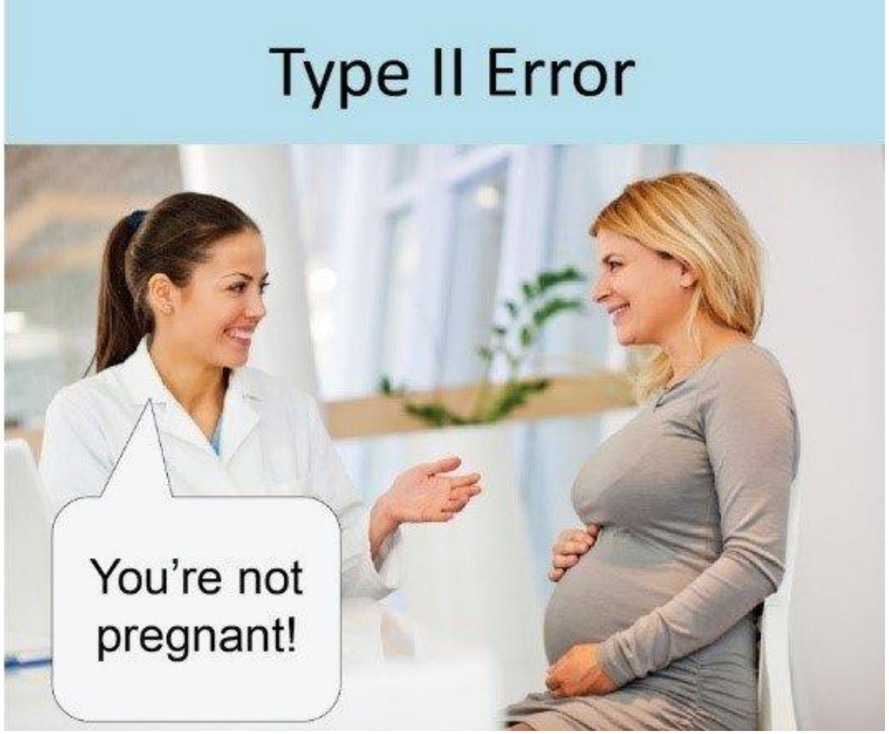
You Can Make Two Types of Errors

Decision / Reality	Do not reject H_0	Reject H_0
H_0 is true	Correct decision	Type I error
H_A is true	Type II error	Correct decision

- Probability of committing a Type-I error is the same as p-value
- α-value can be interpreted as the acceptable probability of making a Type-I error (also called significance level)

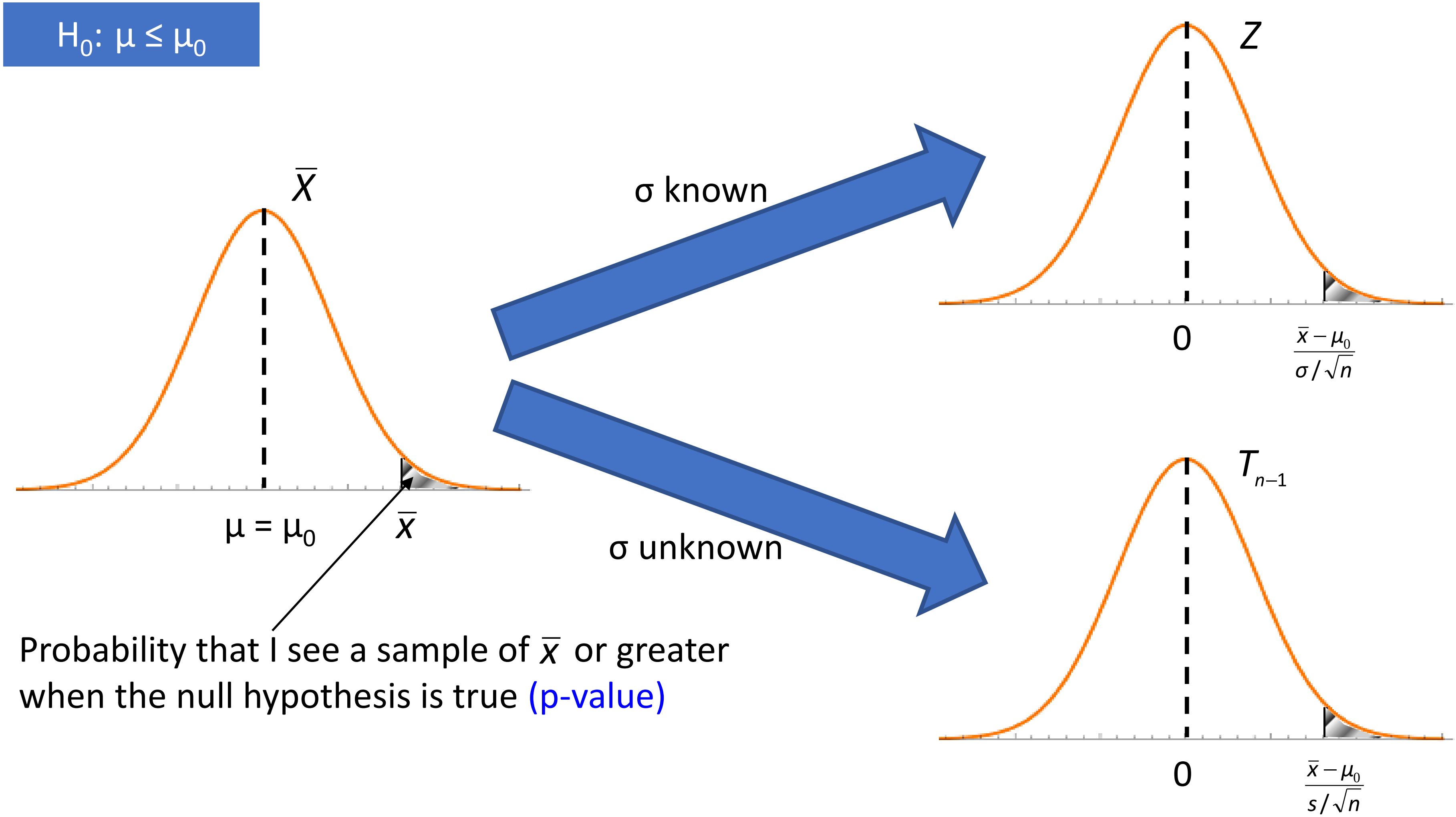
Example - Type I and Type II Errors





Calculating the Probability of Type-I Error

H_0: $\mu \leq \mu_0$


$\bar{X}$
$\mu = \mu_0$
$\bar{x}$
σ known
σ unknown
Z
0
$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
T_{n-1}
0
$\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$

Probability that I see a sample of $\bar{x}$ or greater when the null hypothesis is true (p-value)

Statistical Vs. Practical Significance

- There could be a difference between a statistically significant result and a practically important one
- Large sample sizes often give statistically significant results, even if it has low economic value
- Example?

Comparison Between Two Groups

Managerial Decisions

Are female managers paid less in our company than male managers?

Are salespeople in one zone more productive than others?

Did average household income increase after liberalization in 1991?

Does one of our manufacturing plants have better quality than the other?

Does Angioplasty yield better outcomes than bypass surgery?

Did the child welfare scheme increase the number of school going children?

Learning Objectives

- How to compare means of two populations using paired observations?

- When and how to compare two populations means using independent samples?

- How to test for differences in two population proportions?

Example: Paired Sample t-test

- A nutrition expert would like to assess the effect of organized diet programs on the weight of the participants.

- She randomly chooses 36 participants of the Atkins diet program and measures their weight (in kg) just before enrolling in the program and immediately after the completion of the program.

- Based on this evidence, is the Atkins diet program effective in reducing weight?

Before	After	Before	After
130	123	130	127
123	124	139	132
132	134	120	110
150	152	138	140
146	143	141	136
153	143	120	118
137	133	153	154
140	137	126	125
148	152	148	143
158	149	141	135
144	132	137	135
160	155	159	152
146	142	152	148
146	142	140	138
153	155	140	134
137	130	151	147
138	130	141	144
125	124	139	128

Transforming into a Single Variable

- It is natural and also feasible to take before and after measurements on the same subjects → Paired test

- Let W be the change in weight of a randomly chosen participant after the diet program
  - Mean: μ_W
  - Standard Deviation: σ_W

- "The diet program is effective" ←→ "Average change in weight is negative"

- H_0: $\mu \geq 0$ and H_A: $\mu < 0$

- Test statistic $t = (\mu_w - \mu)/ (\sigma_w /\sqrt{n})$
  with (n-1) degrees of freedom

Example: Independent Sample t-test

- A health chain can informally mention conventional low-calorie diet for free or can recommend Atkins diet by paying $200,000 *licensing fee.*

- The firm has determined that it is worth paying the licensing fee if they can gain enough additional members, which is possible if Atkins diet reduces average weight by 2 pounds or more compared to the conventional low-calorie diet.

- The firm collects weight loss data from two simple random samples of people, one of whom goes through Atkins diet and the other through the conventional diet for 6 months.

	Number	Mean	Std Dev
Atkins	33	15.42	14.37
Conventional	30	7.00	12.36

Hypotheses: Difference Between Means

- We wish to test hypotheses of the following form (where μ_1 and μ_2 are the means of the two populations and D_0 is the least acceptable difference)

$$H_0 : \mu_1 - \mu_2 \leq D_0$$

$$H_A : \mu_1 - \mu_2 > D_0$$

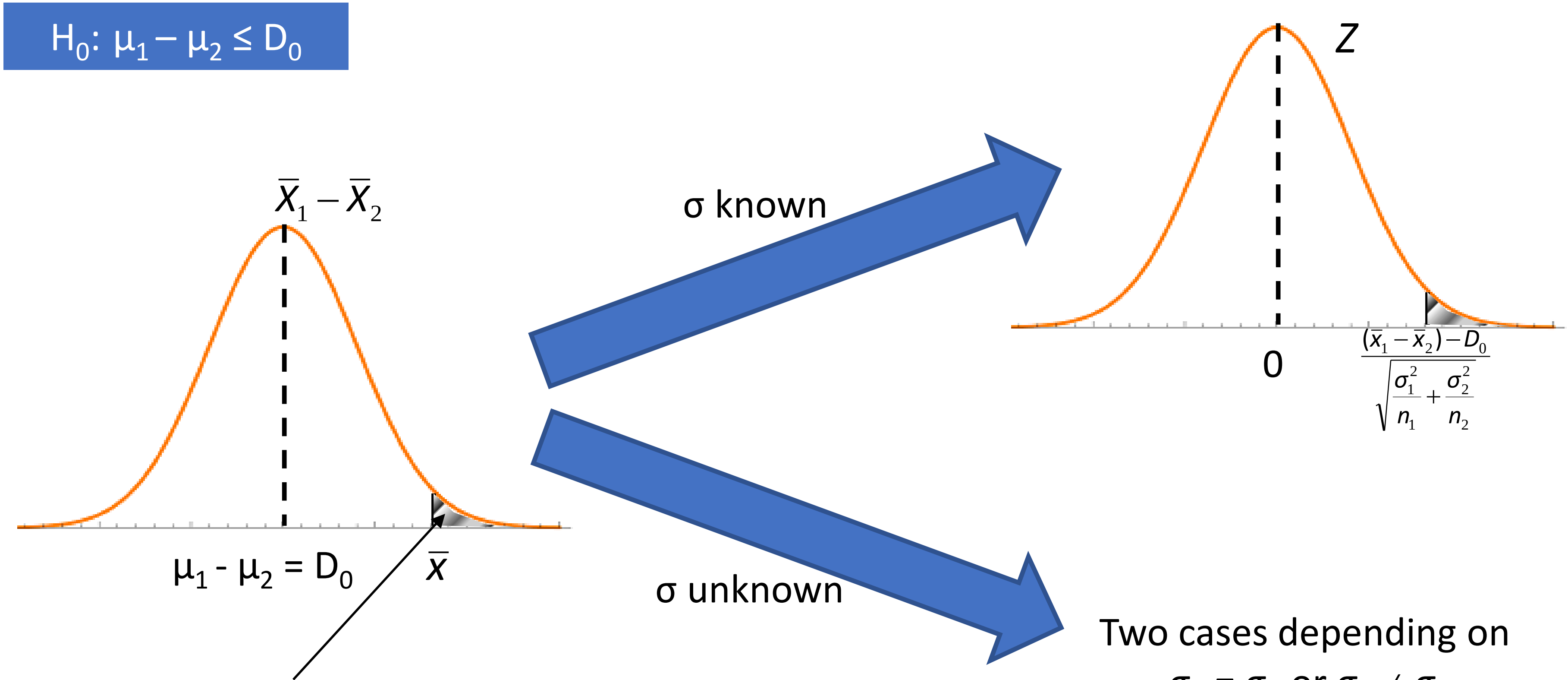
- We will use $\overline{X}_1 - \overline{X}_2$ to make statements about $\mu_1 - \mu_2$

Independence

- Who is in a sample does not influence who else is in that sample
- Who is in a sample does not influence who is in the other sample

Calculating the Probability of Type-I Error

$H_0: \mu_1 - \mu_2 \leq D_0$



$\bar{X}_1 - \bar{X}_2$
$\mu_1 - \mu_2 = D_0$
$\bar{x}$
σ known
σ unknown
Z
0
$\dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

Two cases depending on
$\sigma_1 = \sigma_2$ or $\sigma_1 \neq \sigma_2$

Probability that I see a sample of $\bar{x}$ or greater when the null hypothesis is true (p-value)

With Population Standard Deviations Unknown

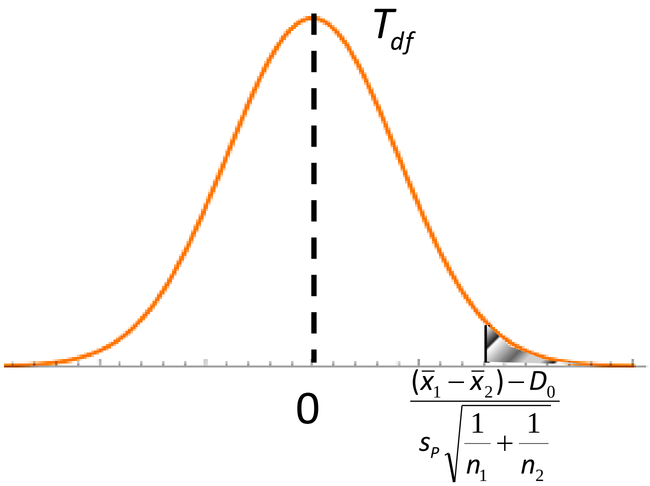
If we believe $\sigma_1 = \sigma_2$

- Calculate the "pooled" sample standard deviation

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

and degrees of freedom

$$df = n_1 + n_2 - 2$$



T_{df}
0
$\dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_P\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

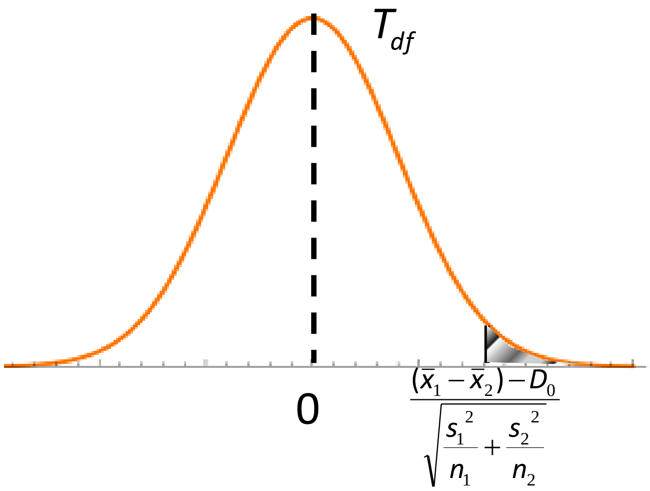
If we believe $\sigma_1 \neq \sigma_2$

- Calculate the standard error

$$se(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and degrees of freedom

$$df = \left\lfloor \frac{\left(s_1^{\ 2}/n_1 + s_2^{\ 2}/n_2\right)^2}{\left(s_1^{\ 2}/n_1\right)^2/(n_1 - 1) + \left(s_2^{\ 2}/n_2\right)^2/(n_2 - 1)} \right\rfloor$$



T_{df}
0
$\dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{s_1^{\ 2}}{n_1} + \dfrac{s_2^{\ 2}}{n_2}}}$

Example: Health Chain

- Hypotheses
  - H_o: $\mu_A - \mu_C \leq 2$
  - H_A: $\mu_A - \mu_C > 2$

- Recall

	Number	Mean	Std Dev
Atkins	33	15.42	14.37
Conventional	30	7.00	12.36

$T^* = \frac{(15.42 - 7.00) - 2}{3.369} \approx 1.91$ and df = 60.82 → P-value = 0.0308 < 0.05.

- Can reject the null hypothesis with less than 5% chance of Type I error

Can We Attribute the Difference to Diets?

- Could there be other systematic differences between the two groups?
  - Atkins is adopted by younger individuals, who are more motivated to lose weight
  - Atkins is adopted by individuals from higher socio-economic strata who have easier access to healthy food options

- These factors can be "controlled" for using additional variables in a regression model

- An alternative way is to randomly assign individuals to one or the other diet program and then compare the difference → Field Experiment

Example: Proportion of Dieters Who Lose Weight

- Suppose an alternate metric to measure the performance of the diet program is proportion of participants who have lost more than 10 pounds

	Number	Successful	Proportion
Atkins	33	20	0.606
Conventional	30	15	0.50

- Hypotheses:
  - H_o: $\pi_A - \pi_c = 0$
  - H_A: $\pi_A - \pi_c \neq 0$

- Similar to previous calculation, we can calculate and proceed with hypothesis testing accordingly

$$z = \frac{(p_A - p_C)}{\sqrt{\bar{p}(1-\bar{p})(1/n_A + 1/n_C)}}$$

- Here $\bar{p}$ is the pooled sample proportion given by

$$\frac{p_A n_A + p_C n_C}{n_A + n_C}$$

Summary of Comparison Tests

- The best way to compare the means of two distributions is using paired observations, if it is feasible

- When paired observations are not possible, we use independent samples and formulate hypothesis on the difference of two means

- It is important to ensure that subjects are randomly assigned to the two samples to avoid any confounding errors

- Similar approach can be used to test the difference in proportions between two populations

Analysis of Variance (ANOVA)

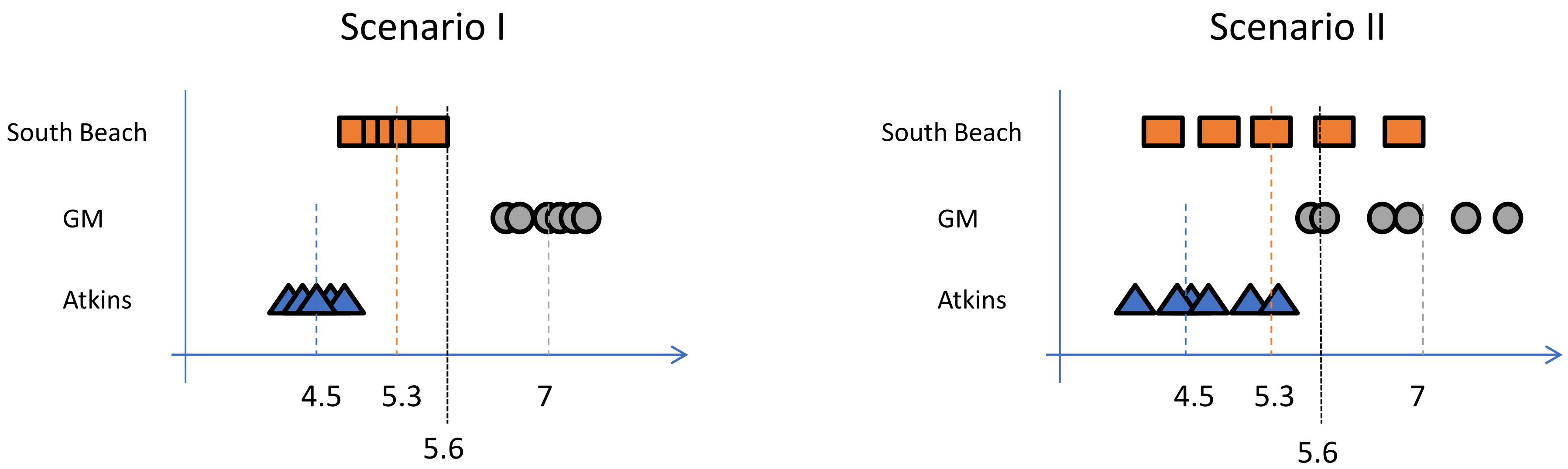
Learning Objectives

- Why is analysis of variance (ANOVA) required to compare means of populations?
- What is the principal of sum of squares?
- How to conduct the ANOVA test?
- What follow-up analysis should be done if ANOVA test is significant?

Example: More Weight Reduction Programs

- Suppose the nutrition expert would like to do a comparative evaluation of three diet programs (Atkins, South Beach, GM)
- She randomly assigns equal number of participants to each of these programs from a common pool of volunteers
- Suppose the average weight losses in each of the groups (arms) of the experiments are 4.5 kg, 7 kg, 5.3 kg
- What can she conclude?

Two Kinds of Variation Matter


Scenario I

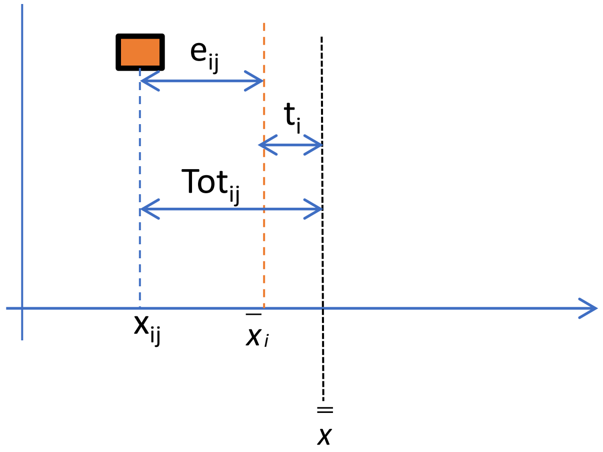
Scenario II

- Not every individual in each program will respond identically to the diet program

- Easier to identify variations across programs if variations within programs are smaller

- Hence the method is called Analysis of Variance (ANOVA)

Formalizing the Intuition Behind Variations



j -> individual
i-> treatment

- It should be obvious that for every observation: $Tot_{ij} = t_i + e_{ij}$

- What is more surprising and useful is:

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i} Tot_{ij}^2 = \sum_{i=1}^{r} n_i t_i^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i} e_{ij}^2$$

Sum of Squares Total (SST)	Sum of Squares Treatment (SSTR)	Sum of Squares Error (SSE)

Statistical Test for Equality of Means

- n subjects equally divided into r groups

- Hypotheses
  - H0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_r$
  - Not all μ_i are equal

- Calculate
  - Mean Square Treatment **MSTR** = SSTR / (r-1)
  - Mean Square Error **MSE** = SSE / (n-r)
  - The ratio of two squares **f** = MSTR/MSE
  - Strength of this evidence **p-value** = $\Pr(F_{(r-1,n-r)} \geq f)$

- Reject the null hypothesis if p-value < α

Example: Weight Reduction Programs

- Suppose 12 participants are allocated to each of the three diet programs
  - r = 3
  - n = 36
- ANOVA Table

	DF	Sum Sq.	Mean Sq.	F value	Pr(>F)
Diet	2	38.89	19.444	3.571	0.0394
Residuals (Error)	33	179.67	5.444		

- If we reject the null hypothesis that all means are equal, probability of you making a mistake is less than 4%
- Can we conclude that GM diet is more effective than Atkins diet?

Further Analysis: Pairwise Differences

- To test whether any two means are different
  - Construct the test statistic $q = \dfrac{\left|\bar{x}_i - \bar{x}_j\right|}{\sqrt{\dfrac{MSE}{n}}}$
  - Calculate the p-value associated with this test statistic: ptukey(q,r,n-r)
  - Reject the null hypothesis that the two means are equal if p-value < α

Summary of ANOVA

- The extent of variation between and within groups determines the strength of evidence against the null hypothesis that means of all groups are equal

- The sum of squared deviations total (around the grand mean) is equal to the sum of squared deviations errors (around respective group means) plus the sum of squared deviations treatment (group means around grand mean)

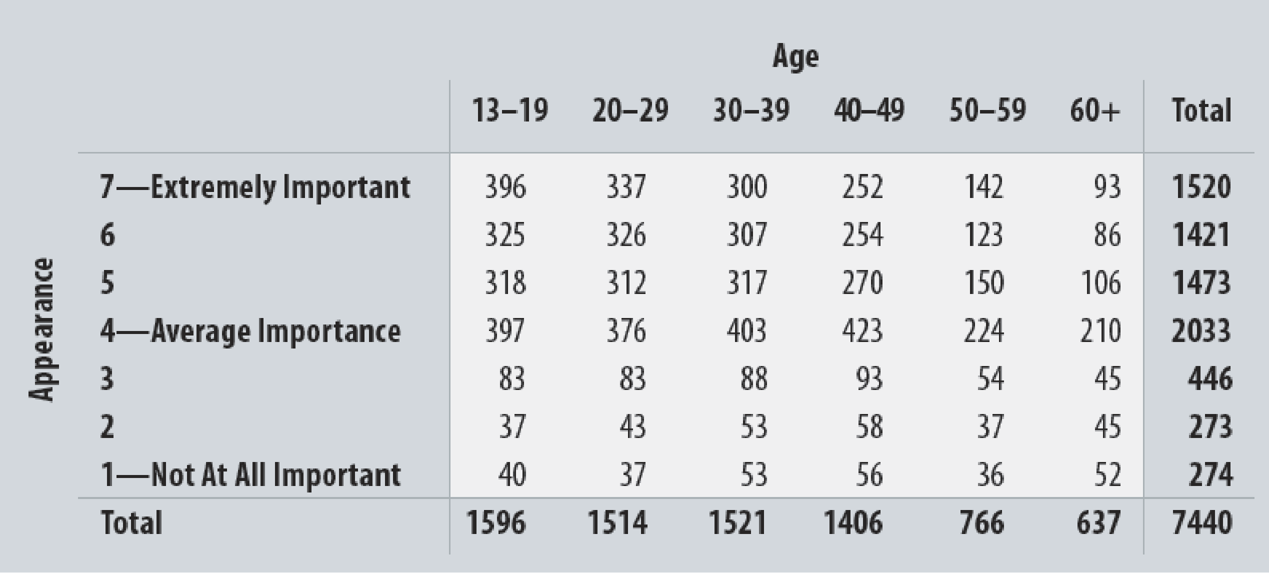
- ANOVA test compares mean squared treatment with mean squared errors. If this ratio is “significantly” greater than 1, we can reject the null hypothesis that the means are equal

- We can conduct a series of Tukey tests for pair-wise comparisons of group means

Chi-Square Test

Chi-Square Test of Independence

The table below shows the importance of personal appearance for several age groups.



		Age						
		13–19	20–29	30–39	40–49	50–59	60+	Total
Appearance	7—Extremely Important	396	337	300	252	142	93	**1520**
	6	325	326	307	254	123	86	**1421**
	5	318	312	317	270	150	106	**1473**
	4—Average Importance	397	376	403	423	224	210	**2033**
	3	83	83	88	93	54	45	**446**
	2	37	43	53	58	37	45	**273**
	1—Not At All Important	40	37	53	56	36	52	**274**
	Total	**1596**	**1514**	**1521**	**1406**	**766**	**637**	**7440**

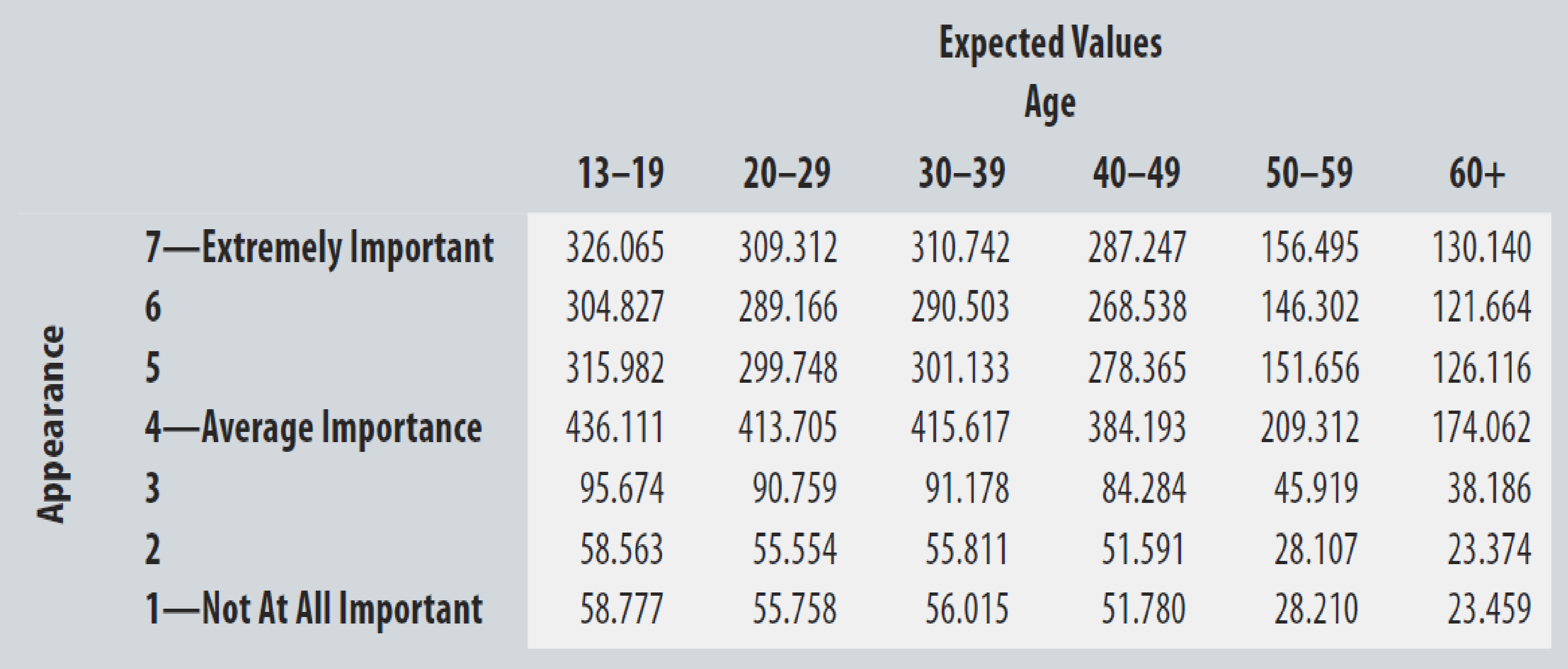
Are *Age and Appearance* independent, or is there a relationship?

Chi-Square Test of Independence – Expected Values

Expected refers to the values we'd expect to see if the null hypothesis is true

Null: The variables are independent. No relationship exists

Exp_{ij} = (sum of row i * sum of column j) / Total sum



		Expected Values					
		Age					
		13–19	20–29	30–39	40–49	50–59	60+
Appearance	7—Extremely Important	326.065	309.312	310.742	287.247	156.495	130.140
	6	304.827	289.166	290.503	268.538	146.302	121.664
	5	315.982	299.748	301.133	278.365	151.656	126.116
	4—Average Importance	436.111	413.705	415.617	384.193	209.312	174.062
	3	95.674	90.759	91.178	84.284	45.919	38.186
	2	58.563	55.554	55.811	51.591	28.107	23.374
	1—Not At All Important	58.777	55.758	56.015	51.780	28.210	23.459

Chi-Square Test of Independence

To run the test, we use a chi-square model with
(7 – 1)(6 – 1) = 30 degrees of freedom:

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp} = 170.7762$$

$$\text{P-value} = P(\chi^2_{30} > 170.7762) < 0.001$$

With the very low P-value, we reject the null hypothesis and conclude that attitudes on personal appearance are not independent of *Age*.

Example : Automobile Manufacturers

Consumer Reports uses surveys to measure reliability in automobiles. Annually they release survey results about problems that consumers have had with vehicles in the past 12 months and the origin of manufacturer. Is consumer satisfaction related to country of origin?

State the hypotheses.

Given the P-value = 0.231, state your conclusion.

$$H_0 : \text{Rate of problems is independent of manufacturer's origin}$$
$$H_A : \text{Rate of problems is not independent of manufacturer's origin}$$

Fail to reject the null hypothesis. There is not enough evidence to conclude there is an association between vehicle problems and origin of vehicle.

Summary

- **Chi-square method is appropriate when both the variables are counts.**

- **Don’t say that one variable “depends” on the other just because they’re not independent.**