# CH 6: Two Way Tables
# Plus a few review Questions

## Chapter 6

- Marginal Distributions
- Conditional Distributions
- Simpsons Paradox

## Review Questions (Ch.3-6)

# CH. 6 Two-Way Tables

- **Use Two-Way** Tables to look at relationships between 2 or more **categorical** variables

- **Should be able to**
  - ➢ Interpret data table & use to calculate percentages
  - ➢ understand the differences between marginal distributions and conditional distributions
  - ➢ know what Simpson Paradox is and when it may be a factor

# Two Way Tables(contingency or frequency tables)

- table of counts or proportions classifying cases on two categorical variables A and B
- each cell is a combination of A and B
- a table with r rows and c columns is an rxc table. The body of the table will contain rxc cells

## Example: 2x2 table

| | **Males** | **Females** |
|---|---|---|
| **Science** | 30 | 40 |
| **Other Faculty** | 20 | 60 |

- convert counts to percentages to do comparisons
- can look at marginal distributions and/or conditional distributions

## • Marginal distributions

- ➢ are distributions of each single categorical variable.
- ➢ Row totals & column totals as a percentage of the grand total in a two way table give marginal distributions of the two individual variables

## • Conditional distributions

➢ help describe relationships amongst categorical variables and can look at;

- **conditional distribution of the row variable** given the column variable(table entries as % of column totals),
- **conditional distribution of the column variable** given the row variable(table entries as % of row totals)

- if one variable is considered explanatory and the other response, then interest will be in the conditional distribution of the response variable given the explanatory variable

- use bar graphs to visually compare/present data(bar graphs shown in class for examples not included in notes)

- **SIMSPONS PARADOX** occurs when overall conclusion doesn't hold within some subgroups (extreme form of the fact that observed associations can be misleading)-will show an example in class

# Example-Allergy Treatment Data

A survey was conducted to evaluate the effectiveness of two allergy remedies(capsule and shot) in a small community. The treatments were provided free of charge in the spring of 2006.

Some received the capsule, others the shot, the rest received nothing.

A random sample of 1,000 local inhabitants in the fall of 2006 yielded the results summarized in the following table:

# Two-Way Table- Allergy Treatment Data

| | Severe Symptoms | Mild or No Symptoms |
|---|---|---|
| No Treatment | 44 | 306 |
| Allergy Capsule | 19 | 131 |
| Allergy Shot | 37 | 463 |

**3x2 table with 6 cells**

# Example: Allergy Treatment Data

**Is there a relationship between treatments and symptoms?**

- Look at each variable separately first.
- Calculate row and column totals

| | Severe Symptoms | Mild or No Symptoms | **Row Totals** |
|---|---|---|---|
| No Treatment | 44 | 306 | **350** |
| Allergy Capsule | 19 | 131 | **150** |
| Allergy Shot | 37 | 463 | **500** |
| **Column Totals** | **100** | **900** | **1,000** |

## (i) Marginal distribution of the variable treatment

- **Row totals as a percentage of grand total gives the *marginal distribution* of the variable *treatment***

| | None | Capsule | Shot | Total |
|---|---|---|---|---|
| % | 350/1000=**35%** | 150/1000=**15%** | 500/1000=**50%** | 100% |

## (ii) Marginal distribution of the variable Symptoms

- **Column totals as a percentage of grand total gives the *marginal distribution* of the variable *Symptoms***

| | Severe | Mild/none | Total |
|---|---|---|---|
| % | 100/1000=**10%** | 900/1000=**90%** | 100% |

## (iii)Conditional Distribution of Treatments given Symptoms (calculate column percentages)

| | Severe Symptoms | Mild or No Symptoms |
|---|---|---|
| **No treatment** | 44/100= **44%** | 306/900=**34%** |
| **Allergy Capsule** | 19/100= **19%** | 131/900=**15%** |
| **Allergy Shot** | 37/100= **37%** | 463/900=**51%** |
| **Totals** | 100% | 100% |

## (iv) Conditional Distribution of Symptoms given Treatments (calculate row percentages)

| | Severe Symptoms | Mild/No Symptoms | Totals |
|---|---|---|---|
| **No Treatment** | 44/350=**13%** | 306/350=**87%** | 100% |
| **Allergy Capsule** | 19/150=**13%** | 131/150=**87%** | 100% |
| **Allergy Shot** | 37/500= **7%** | 463/500=**93%** | 100% |

# Which conditional distribution is most meaningful?

- If one variable is considered explanatory and the other response, then;
  - examine the conditional distribution of the response variable given the explanatory variable
  - in example above, treatment was the explanatory variable. To evaluate the difference between treatments, we look at the conditional distribution of the symptoms given treatments

# Simpson's Paradox

- **an association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group**
- the **reversal is called Simpson's Paradox** (occurs when overall conclusion doesn't hold within some subgroups)
- Simpsons paradox is an example of the effect of lurking variables on an observed association

## Example: Airline Flight Delays

| | Alaska Airlines(AA) | America West(AW) |
|---|---|---|
| **Flights On Time** | 3,274 | 6,438 |
| **Flights Delayed** | 501 | 787 |

# Example: Airline Flight Delays

| | Alaska Airlines(AA) | America West(AW) | Total |
|---|---|---|---|
| **Flights On Time** | 3,274 | 6,438 | **9,712** |
| **Flights Delayed** | 501 | 787 | **1,288** |
| **Total** | **3,775** | **7,225** | **11,000** |

## Conditional Distribution-Arrival time, given airlines

| | AA | AW |
|---|---|---|
| % on time | 3274/3775= **86.7%** | 6438/7225 = **89.1%** |
| % delayed | 501/3775 = **13.3%** | 787/7225 = **10.9%** |

- **AW looks better overall (lower % flights delayed)**

# Data by airport(landing)

| | Alaska | Airlines | | America | West |
|---|---|---|---|---|---|
| | On Time | Delayed | | On Time | Delayed |
| **Los Angelas** | 497 | 62 | | 694 | 117 |
| **Phoenix** | 221 | 12 | | 4,840 | 415 |
| **San Diego** | 212 | 20 | | 383 | 65 |
| **San Francisco** | 503 | 102 | | 320 | 129 |
| **Seattle** | 1,841 | 305 | | 201 | 61 |
| **Totals** | **3,274** | **501** | | **6,438** | **787** |

# Delayed by airport(landing)

| | Alaska Airlines | America West |
|---|---|---|
| **Los Angelas** | 62/(497+62) = **11.1%** | 117/(694+117)=**14.4%** |
| **Phoenix** | 12/(221+12) = **5.2%** | 415/(4840+415)= **7.9%** |
| **San Diego** | **8.6%** | **14.5%** |
| **San Francisco** | **16.9%** | **28.7%** |
| **Seattle** | **14.2%** | **23.3%** |

- America West(AW) does worse at every 1 of the five airports, yet does better overall(Simpsons Paradox)
  - ❖ WHY?

# Review- Sample M.C. questions –Ch 6, 3-5

**1)** The association between the weather office forecast and actual weather is summarized in the following table:

| | Rain | No rain | Total |
|---|---|---|---|
| **Forecast: rain** | 33 | 78 | 111 |
| **Forecast: no rain** | 7 | 382 | 389 |
| Total | 40 | 460 | 500 |

What percentage of the times was the weather office wrong?

(A) 15.6% (B) 8.0% (C) 17.0% (D) 22.2%

**2)** Applicants looking for a job at a restaurant chain may apply to be a server or kitchen worker. The table below summarizes the numbers of male and female applicants hired for the jobs they applied for.

| | Server | | | Kitchen worker | |
|---|---|---|---|---|---|
| | Male | Female | | Male | Female |
| Not hired | 80 | 120 | Not hired | 30 | 15 |
| Hired | 20 | 50 | Hired | 80 | 25 |

The proportion of female applicants for a job as kitchen worker that were hired is

(A) 0.100 (B) 0.375 (C) 0.400 (D) 0.357 (E) 0.625

**3)** A popular graduate school accepts only the top 2.5% of all applicants based on their graduating average grades

Using the 68-95-99.7% rule you are told that the interval containing the middle 99.7% of the average grades for applicants is 80 to 98.

What is the cut off average grade for admittance to this graduate school?

(A) 83 (B) 89 (C) 92 (D) 95

**4)** You are given that the heights of children in kindergarten can be modeled as N(38.2, 1.8) . Measurements are in inches.

At least how tall are the tallest 10% of kindergarten children?

(A) 41.7 inches

(B) 40.5 inches

(C) 44.7 inches

(D) 35.9 inches

**5)** The length x of nails in a large shipment received by a carpenter are approximately normally distributed with mean 2 inches and standard deviation 0.1 inches.

The carpenter cannot use a nail shorter than 1.75 inches or longer than 2.25 inches. What percentage of the shipment of nails will the carpenter be able to use?

(A) 98.76% (B) 99.38% (C) 99.70% (D) 95.00%

**6)** The least squares regression line is;

**A)** the line that always makes the square of the correlation of the data as large as possible

**B)** the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible

**C)** the line that always splits the data in half, with exactly half of the points above the line and half below the line

**D)** the line that satisfies all of the above

**7)** You are given that

$$\bar{x} = 4.14 \quad s_x = 2.14 \quad \bar{y} = 2.73 \quad s_y = 0.43$$

and the least squares regression line is

$$\hat{y} = 1.9 - 0.2x$$

Based on this information, what is the correlation between $x$ and $y$?

(A) 0.995 (B) -0.995 (C) -0.07

(D) 0.04 (E) -0.04