Math 243 **Sections 3.2 and 2.3** **Categorical Data and Contingency Tables**

Overview

- Frequency and Relative Frequency Tables
- Pie Charts and Bar Charts
- Contingency Tables
- Marginal, Joint and Conditional Probabilities
- Independence Test

Displaying a Single Categorical Variable – Frequency and Relative Frequency Tables

Activity 1. A frequency table uses category names for each row and records the total count of each value. A relative frequency table gives the percentage in each category.

a. Using our class data, create a frequency table and relative frequency table with "Award" as the categorical variable.

Award	Frequency (Count)	Relative Frequency (%)
Academy Award	6	$\frac{6}{23} = .26$ or 26%
Olympic Gold	1	$\frac{1}{23} = .04$ 4%
Nobel Prize	13	$\frac{13}{23} = .57$ 57%
None	3	$\frac{3}{23} = .13$ 13%

$\frac{3}{23}$

Bar Charts

b. Create the above frequency table in Excel and create a bar chart and a relative frequency bar chart from the data. Add axis labels and a title. Note the difference in the vertical scale. Also note that there should be space between the bars for categorical data. (Histograms for numerical data should not have spaces between the bars)

✓

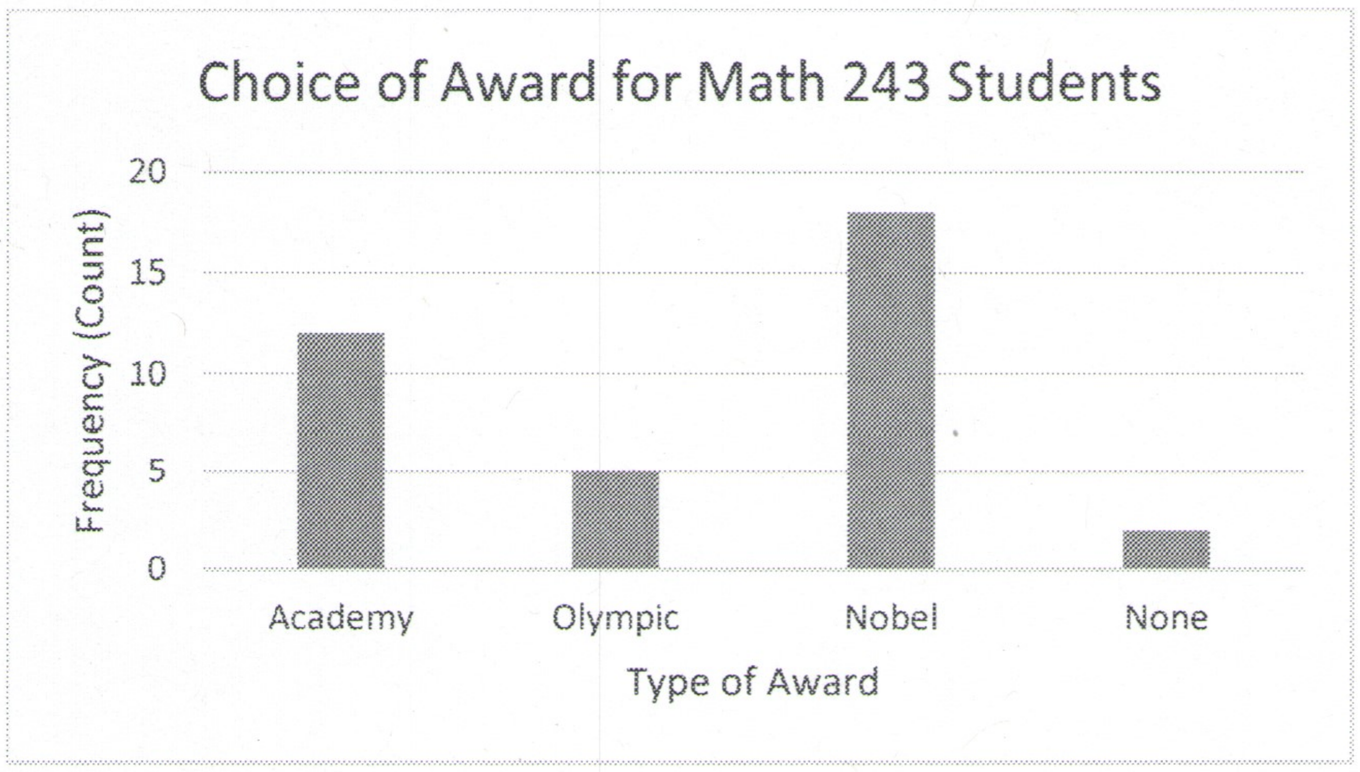
Pie Charts

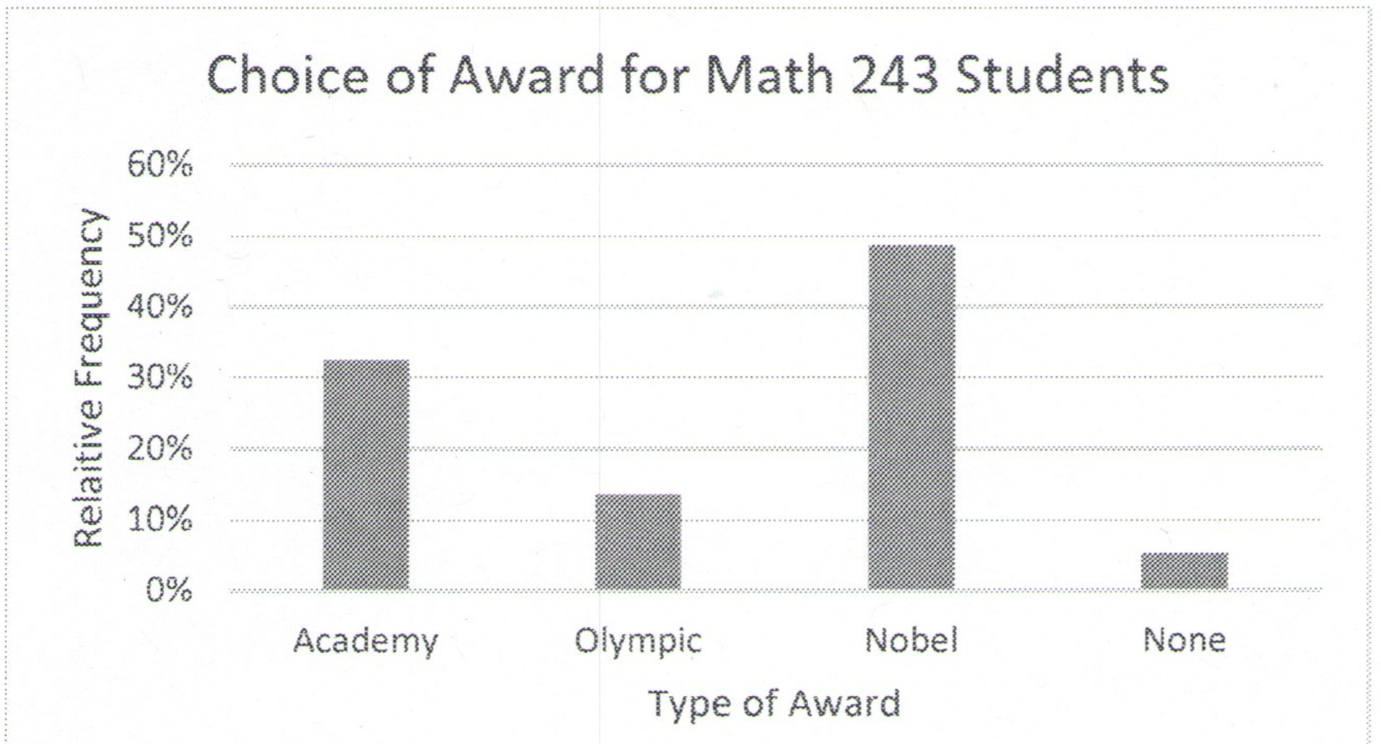
c. Use Excel to create a pie chart from the data.

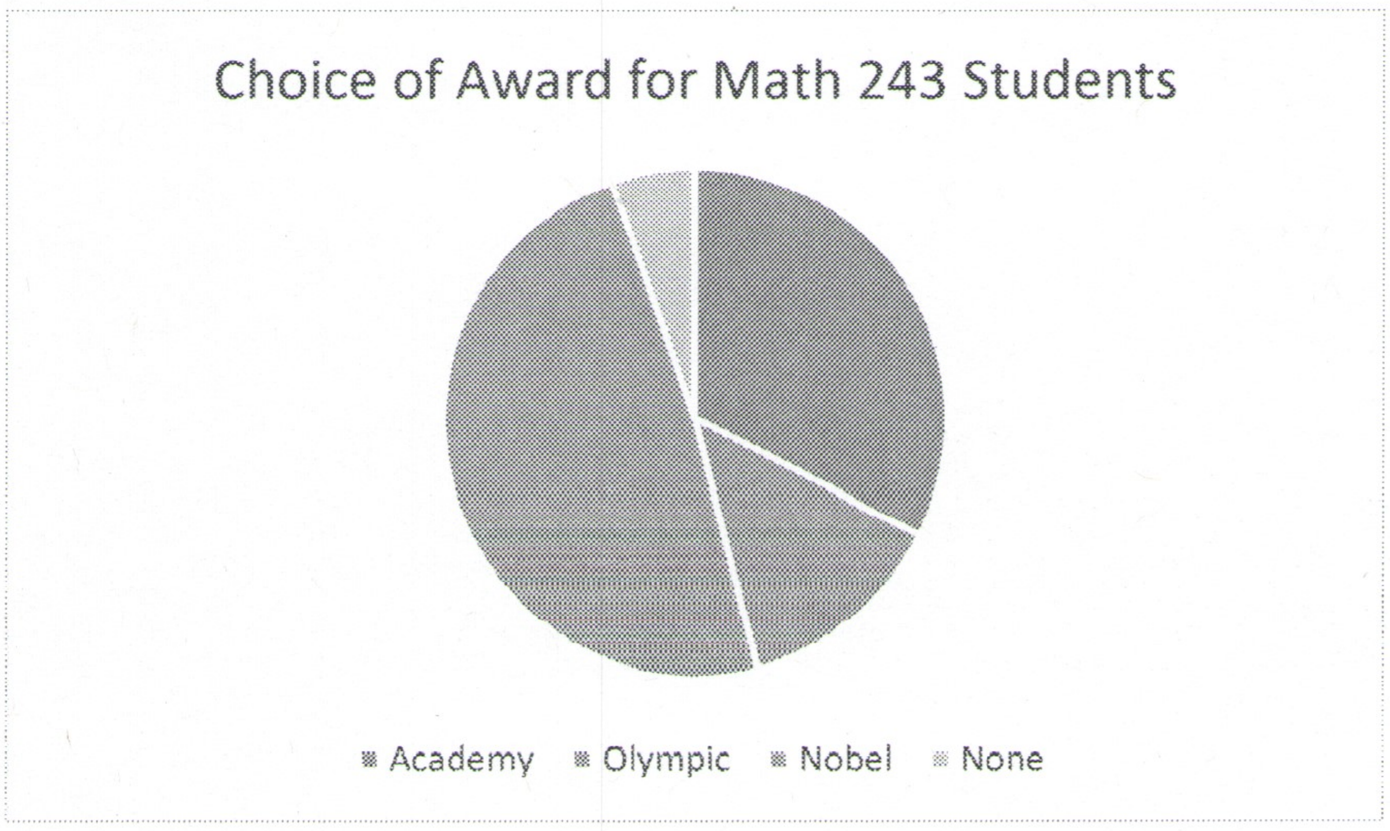
✓

d. You cannot use a pie chart when... we have a categorical variable where people check all that apply (or more than one answer)

Examples from Excel. Label both axes and give your graph a descriptive, meaningful title.







Comparing Two Categorical Variables – Contingency Tables with Counts

Example 1. Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard treatment for cocaine) and placebo. Results of the study are summarized below. Source: *http://www.oswego.edu/~srp/stats/2_way_tbl_1.htm*

Marginal Probabilities (Margins or Totals)

a. If we selected a participant at random, what is the probability that they had a relapse?

$\frac{48}{72} = .6667$ or 67%

b. What is the probability that a participant did not have a relapse?

$\frac{24}{72} = .3333$ or 33%

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

These are called marginal probabilities because we use the numbers in the margins. We use the row or column total over the total of all participants.

Joint Probabilities (And)

c. What is the probability that a participant took desipramine and had a relapse?

$\frac{10}{72} = .1389$ or 14%

d. What is the probability that a participant had the placebo and had a relapse?

$\frac{20}{72} = .2778$ or 28%

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

These are called joint probabilities because they are the intersection between two variables. They are "and" probabilities. (Notice we don't multiply like we did for independent events)

Contingency Tables with Proportions

e. Let's rewrite the table in terms of proportions or percentages so that the joint and marginal probabilities can be read right from the table.

	relapse	no relapse	total
desipramine	.14	.19	.33
lithium	.25	.08	.33
placebo	.28	.06	.34
total	.67	.33	1.00

Conditional Probabilities (If or Given)

f. If a person took desipramine, what is the probability that they had a relapse.

P(Relapse|Desipramine) = $\frac{10}{24} = .4167$ or 42%

↑ vertical bar. "given"

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

g. Given that a person had lithium, what is the probability that they had a relapse? Write the probability statement and the answer.

$P(\text{relapse} \mid \text{lithium}) = \frac{18}{24} = .75$ or 75%

h. What is the probability that someone had a relapse if they were in the placebo group? Write the probability statement and the answer.

$P(\text{relapse} \mid \text{placebo}) = \frac{20}{24} = .8333$ or 83%

These are called conditional probabilities because we are given one of the variable values. We only use a single row or column to find the conditional probability.

Conditional Probability Formula: For events A and B,

$$P(B|A) = \frac{P(B \text{ and } A)}{P(A)}$$

$\frac{P(\text{relapse and desipramine})}{P(\text{desipramine})} = \frac{10}{24}$

Note: this is exactly how we calculated the conditional probabilities using the table so you do not need to memorize this formula.

i. Given that a person had a relapse, what is the probability that they were in the placebo group?

$P(\text{placebo} \mid \text{relapse}) = \frac{20}{48} = .4167$ 42%

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

j. What is the probability that a person in the relapse group took lithium?

given $P(\text{lithium} \mid \text{relapse}) = \frac{18}{48} = .375$ or 38%

k. What is the probability that a person took desipramine if they had a relapse?

$P(\text{desipramine} \mid \text{relapse}) = \frac{10}{48} = .2083$ or 21%

Example 2. In a Math 243 class, 82% said they like chocolate, 59% said they like expresso, and 53% like both. We can use these three proportions to complete the rest of the contingency table.

Chocolate and Espresso Likability for a Math 243 Class

	Likes Chocolate	Does not like chocolate	Total
Likes Espresso	0.53	0.06	0.59
Does not like espresso	.29	0.12	0.41
Total	0.82	0.18	1.00

.59 − .53 = .06

.82 − .53 = .29

1.00 − .59 = .41

1.00 − .82 = .18

0.41 − .29 = .12

a. What's the probability that a randomly selected student likes espresso given that they like chocolate? Write a probability statement and find calculate the answer.

$$P(E|C) = \frac{.53}{.82} = .6463$$

b. What's the probability that a randomly selected student likes espresso given that they do not like chocolate? Write a probability statement and find calculate the answer.

$$P(E|C^c) = \frac{.06}{.18} = .3333$$

different, so liking espresso is dependent on liking chocolate for this class

c. What's the overall probability that students like espresso? Write a probability statement and find calculate the answer.

$$P(E) = \frac{.59}{1.00} = .59$$

Independence Test – Is liking chocolate independent of liking expresso?

If the conditional row or column probabilities are equal all the way across (or not significantly different), the variables are independent.

If $P(B|A) = P(B)$, then A and B are independent. This means knowing that event A occurred does not affect the chance of B occurring.

d. We can compare the probabilities that we found in parts a-c. Fill in the numbers and state your conclusion. Are liking chocolate and espresso independent for this class?

$Is\ P(E|C) = P(E|C^C) = P(E)?$

.6463 ≠ .333 ≠ .59

equal ↔ independent

not equal ↔ not independent dependent

Example 3. Here are the results from a Psychology class where 80% like Chocolate, 40% like Espresso and 32% like both.

Chocolate and Espresso Likability for a Psychology Class

	Likes Chocolate	Does not like chocolate	Total
Likes Espresso	0.32	0.08	0.40
Does not like espresso	0.48	0.12	0.60
Total	0.80	0.20	1.00

a. What's the probability that a randomly selected student likes espresso given that they like chocolate? Write a probability statement and find calculate the answer.

$$P(E|C) = \frac{.32}{.80} = 0.40$$

b. What's the probability that a randomly selected student likes espresso given that they do not like chocolate? Write a probability statement and find calculate the answer.

$$P(E|C^c) = \frac{.08}{.20} = 0.40$$

c. What's the overall probability that students like espresso? Write a probability statement and find calculate the answer

$$P(E) = \frac{.40}{1.00} = .40$$

d. Are the events liking chocolate and liking espresso independent for this class? Fill in the numbers and state your conclusion. Are liking chocolate and espresso independent for this class?

$Is\ P(E|C) = P(E|C^C) = P(E)?$

$.40 = .40 = .40$ independent

Independence Test – What does independence look like visually?

e. Open the Excel Template in D2L and create a clustered bar chart for the math class and the psychology class. How can you tell whether the variables are independent?

Practice 1. Going back to the cocaine addiction treatment experiment, do the treatments and having a relapse appear to be independent of each other? Does one of the treatments seem more effective than the others? Test the rows or columns:

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Desipramine

$\frac{10}{48} \stackrel{?}{=} \frac{14}{24} \stackrel{?}{=} \frac{24}{72}$

.2083 ≠ .5833 ≠ .3333

Lithium

$\frac{18}{48} \stackrel{?}{=} \frac{6}{24} \stackrel{?}{=} \frac{24}{72}$

.375 ≠ .25 ≠ .3333

The probabilities are not the same for each group so the treatments and relapsing are not independent.

Practice 2. How are the smoking habits of students related to their parents' smoking? Here is a contingency table of data from a survey of students in 8 Oregon high schools.

	Two parents smoke	One parent smokes	No parents smoke	Total
Student Smokes	400	416	188	1004
Student does not smoke	1380	1823	1168	4371
Total	1780	2239	1356	5375

Find the following probabilities.

(a) $P(\text{student smokes}) = \frac{1004}{5375} = .1868$

(b) $P(\text{no parents smoke}) = \frac{1356}{5375} = .2523$

(c) $P(\text{at least 1 parent smokes})$?

$= \frac{1780 + 2239}{5375} = \frac{4019}{5375} = .7477$

(d) $P(\text{student smokes and 1 parent smokes})$?

$= \frac{416}{5375} = .0774$

(e) $P(\text{student smokes and no parents smoke})$

$= \frac{188}{5375} = .0350$

(f) $P(\text{student smokes \& at least 1 parent smokes})$

$= \frac{400+416}{5375} = \frac{816}{5375} = .1518$

(g) What is the probability that a student who smokes has no parents that smoke? (given)

$P(\text{no parents smoke} \mid \text{student smokes}) = \frac{188}{1004} = .1873$

(h) What is the probability that if two parents smoke, their child will smoke? (given)

$P(\text{child smokes} \mid \text{2 parents smoke}) = \frac{400}{1780} = .2247$

(i) Do parents smoking and the student smoking appear to be independent of each other?

$\frac{400}{1780} \stackrel{?}{=} \frac{416}{2239} \stackrel{?}{=} \frac{188}{1356} \stackrel{?}{=} \frac{1004}{5375}$

.2247 ≠ .1858 ≠ .1386 ≠ .1868

The probabilities are fairly different so they seem to be dependent.

Practice 3. Using our class data, make a contingency table of award vs. gender identity. Is the type of award desired dependent on gender identity for our class? Show mathematical proof using the independence test and look at the clustered bar chart.