AP Stats: Lunchtime Review #4
Categorical Data

Categorical Data: 5.2, 5.3

Classifying data by

Criteria 2		Criteria 1		
		Criteria 1A	Criteria 1B	
	Criteria 2A			
	Criteria 2B			

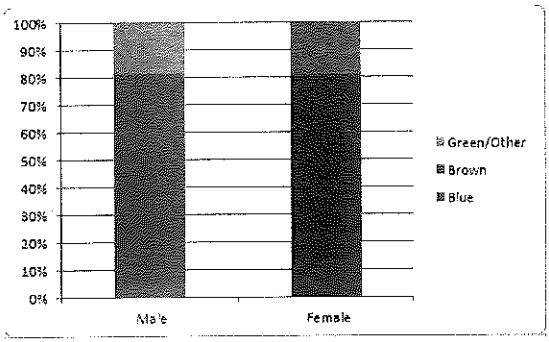
- Joint frequency is the frequency with which

- Marginal Frequency is the frequency of each

- Marginal distributions are the proportions of

- Conditional relative frequencies give the proportion of

- Segmented bar graphs display



Thus Dr. Fixit doe)d condition (88.2% versus Dr. Patch's 87.(patients in poor condition (71.4% versus)r. Fixit has a lower overall patient survival rate (76% versus Dr. Patch's 80%)! How can this be?

This problem is an example *of Simpson's paradox*, where a comparison can be reversed when more than one group is combined to form a single group. The effect of another variable, sometimes called a *lurking variable*, is masked when the groups are combined. In this particular example, closer scrutiny reveals that Dr. Fixit operates on many more patients in poor condition than Dr. Patch, and these patients in poor condition are precisely the ones with lower survival rates. Thus even though Dr. Fixit does better with all patients, his overall rating is lower. Our original table hid the effect of the lurking variable related to the condition of the patients.

Questions on Topic Five: Exploring Categorical Data

Multiple-Choice Questions

Directions: The questions or incomplete statements that follow are each followed by five suggested answers or completions. Choose the response that best answers the question or completes the statement.

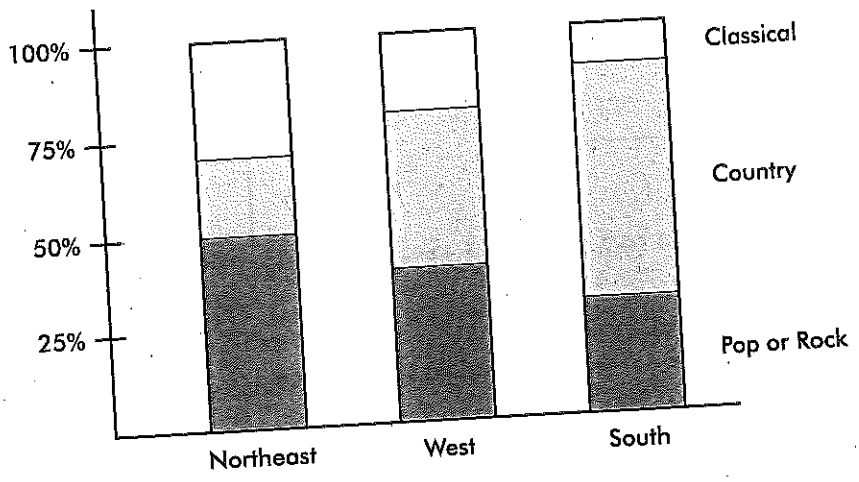
Questions 1–5 are based on the following: To study the relationship between party affiliation and support for a balanced budget amendment, 500 registered voters were surveyed with the following results:

	For	Against	No opinion
Democrat	50	150	50
Republican	125	50	25
Independent	15	10	25

1. What percentage of those surveyed were Democrats?
   (A) 10%
   (B) 20%
   (C) 30%
   (D) 40%
   (E) 50%

2. What percentage of those surveyed were for the amendment and were Republicans?
   (A) 25%
   (B) 38%
   (C) 40%
   (D) 62.5%
   (E) 65.8%
3. What percentage of Independents had no opinion?
   (A) 5%
   (B) 10%
   (C) 20%
   (D) 25%
   (E) 50%
4. What percentage of those against the amendment were Democrats?
   (A) 30%
   (B) 42%
   (C) 50%
   (D) 60%
   (E) 71.4%
5. Voters of which affiliation were most likely to have no opinion about the amendment?
   (A) Democrat
   (B) Republican
   (C) Independent
   (D) Republican and Independent, equally
   (E) Democrat, Republican, and Independent, equally

Questions 6–10 are based on the following: A study of music preferences in three geographic locations resulted in the following segmented bar chart:



6. What percentage of those surveyed from the Northeast prefer country music?
   (A) 20%
   (B) 30%
   (C) 40%
   (D) 50%
   (E) 70%
7. Which of the following is greatest?
   (A) The percentage of those from the Northeast who prefer classical.
   (B) The percentage of those from the West who prefer country.
   (C) The percentage of those from the South who prefer pop or rock.
   (D) The above are all equal.
   (E) It is impossible to determine the answer without knowing the actual numbers of people involved.
8. Which of the following is greatest?
   (A) The number of people in the Northeast who prefer pop or rock.
   (B) The number of people in the West who prefer classical.
   (C) The number of people in the South who prefer country.
   (D) The above are all equal.
   (E) It is impossible to determine the answer without knowing the actual numbers of people involved.
9. All three bars have a height of 100%.
   (A) This is a coincidence.
   (B) This happened because each bar shows a complete distribution.
   (C) This happened because there are three bars each divided into three segments.
   (D) This happened because of the nature of musical patterns.
   (E) None of the above is true.
10. Based on the given segmented bar chart, does there seem to be a relationship between geographic location and music preference?
   (A) Yes, because the corresponding segments of the three bars have different lengths.
   (B) Yes, because the heights of the three bars are identical.
   (C) Yes, because there are three segments and three bars.
   (D) No, because the heights of the three bars are identical.
   (E) No, because summing the corresponding segments for classical, summing the corresponding segments for country, and summing the corresponding segments for pop or rock all give approximately the same total.
11. In the following table, what value for n results in a table showing perfect independence?

20	50
30	n

(A) 10
(B) 40
(C) 60
(D) 75
(E) 100

12. A company employs both men and women in its secretarial and executive positions. In reports filed with the government, the company shows that the percentage of female employees who receive raises is higher than the percentage of male employees who receive raises. A government investigator claims that the percentage of male secretaries who receive raises is higher than the percentage of female secretaries who receive raises, and that the percentage of male executives who receive raises is higher than the percentage of female executives who receive raises. Is this possible?

(A) No, either the company report is wrong or the investigator's claim is wrong.
(B) No, if the company report is correct, then either a greater percentage of female secretaries than of male secretaries receive raises or a greater percentage of female executives than of male executives receive raises.
(C) No, if the investigator is correct, then by summation of the corresponding numbers, the total percentage of male employees who receive raises would have to be greater than the total percentage of female employees who receive raises.
(D) All of the above are true.
(E) It is possible for both the company report to be true and the investigator's claim to be correct.

Answer Key

1. **E**	4. **E**	7. **B**	10. **A**
2. **A**	5. **C**	8. **E**	11. **D**
3. **E**	6. **A**	9. **B**	12. **E**

Answers Explained

1. **(E)** Of the 500 people surveyed, $50 + 150 + 50 = 250$ were Democrats, and $\frac{250}{500} = .5$ or 50%.
2. **(A)** Of the 500 people surveyed, 125 were both for the amendment and were Republicans, and $\frac{125}{500} = .25$ or 25%.
3. **(E)** There were $15 + 10 + 25 = 50$ Independents; 25 of them had no opinion, and $\frac{25}{50} = .5$ or 50%.
4. **(E)** There were $150 + 50 + 10 = 210$ people against the amendment; 150 of them were Democrats, and $\frac{150}{210} = .714$ or 71.4%.
5. **(C)** The percentages of Democrats, Republicans, and Independents with no opinion are 20%, 12.5%, and 50%, respectively.
6. **(A)** In the bar corresponding to the Northeast, the segment corresponding to country music stretches from the 50% level to the 70% level, indicating a length of 20%.

7. **(B)** Based on lengths of indicated segments, the percentage from the West who prefer country is the greatest.
8. **(E)** The given bar chart shows percentages, not actual numbers.
9. **(B)** In a complete distribution, the probabilities sum to 1, and the relative frequencies total 100%.
10. **(A)** The different lengths of corresponding segments show that in different geographic regions different percentages of people prefer each of the music categories.
11. **(D)** Relative frequencies must be equal. Either looking at rows gives $\frac{20}{70} = \frac{30}{30+n}$ or looking at columns gives $\frac{20}{50} = \frac{50}{50+n}$. We could also set up a proportion $\frac{n}{30} = \frac{50}{20}$ or $\frac{n}{50} = \frac{30}{20}$. Solving any of these equations gives $n = 75$.
12. **(E)** It is possible for both to be correct, for example, if there were 11 secretaries (10 women, 3 of whom receive raises, and 1 man who receives a raise) and 11 executives (10 men, 1 of whom receives a raise, and 1 woman who does not receive a raise). Then 100% of the male secretaries receive raises while only 30% of the female secretaries do; and 10% of the male executives receive raises while 0% of the female executives do. However, overall 3 out of 11 women receive raises, while only 2 out of 11 men receive raises. This is an example of Simpson's paradox.

Free-Response Questions

Directions: You must show all work and indicate the methods you use. You will be graded on the correctness of your methods and on the accuracy of your final answers.

Two Open-Ended Questions

1. Suppose that in a telephone survey of 800 registered voters, the data are cross-classified both by gender of respondent and by respondent's opinion on an environmental bond issue.

	Bond issue	
	For	Against
Men	450	150
Women	160	40

Analyze this table looking both at marginal distributions and conditional distributions.

2. Following are the results of a study (*New England Journal of Medicine*, Vol. 319, No. 17, p. 1108, 1988) to determine whether taking an aspirin every other day reduces the risk of a heart attack in men:

	Aspirin	Placebo
Fatal attack	5	18
Nonfatal attack	99	171
No heart attack	10,933	10,845

Analyze this table by looking both at marginal distributions and conditional distributions.

Answers Explained

1. First we find the row and column totals:

	Bond issue: For	Bond issue: Against	
Men	450	150	600
Women	160	40	200
	610	190	800

We calculate the following marginal distributions:

Of the 800 people interviewed, $\frac{600}{800}$ or 75% are men and $\frac{200}{800}$ or 25% are women.

Of the 800 people interviewed, $\frac{610}{800}$ or 76.25% are for the bond issue and $\frac{190}{800}$ or 23.75% are against the bond issue.

Looking at the body of the table, we calculate the following conditional distributions:

Of the 600 men interviewed, $\frac{450}{600}$ or 75% are for the bond issue and $\frac{150}{600}$ or 25% are against it.

Of the 200 women interviewed, $\frac{160}{200}$ or 80% are for the bond issue and $\frac{40}{200}$ or 20% are against it.

Of the 610 people for the bond issue, $\frac{450}{610}$ or 73.77% are men and $\frac{160}{610}$ or 26.23% are women.

Of the 190 people against the bond issue, $\frac{150}{190}$ or 78.95% are men and $\frac{40}{190}$ or 21.05% are women.

There doesn't seem to be much of a relationship between the gender of a respondent and the respondent's opinion on the environmental bond issue.

2. First we find the row and column totals:

	Aspirin	Placebo	
Fatal attack	5	18	23
Nonfatal attack	99	171	270
No heart attack	10,933	10,845	21,778
	11,037	11,034	22,071

We calculate the following marginal distributions:

Of the 22,071 men in the study, $\frac{11,037}{22,071}$ or 50% took aspirin, and $\frac{11,034}{22,071}$ or 50% took the placebo.

Of the 22,071 men in the study, $\frac{23}{22,071}$ or .10% had a fatal heart attack, $\frac{270}{22,071}$ or 1.22% had a nonfatal attack, and $\frac{21,778}{22,071}$ or 98.67% had no heart attack.

Looking at the body of the table, we calculate the following conditional distributions:

Of the 11,037 men taking aspirin, $\frac{5}{11,037}$ or .045% had a fatal heart attack, $\frac{99}{11,037}$ or .897% had a nonfatal attack and $\frac{10,933}{11,037}$ or 99.058% had no heart attack.

Of the 11,034 men taking the placebo, $\frac{18}{11,034}$ or .163% had a fatal heart attack, $\frac{171}{11,034}$ or 1.550% had a nonfatal attack, and $\frac{10,845}{11,034}$ or 98.287% had no heart attack.

Of the 23 men who had a fatal heart attack, $\frac{5}{23}$ or 21.74% had taken aspirin, while $\frac{18}{23}$ or 78.26% had taken the placebo.

Of the 270 men who had a nonfatal heart attack, $\frac{99}{270}$ or 36.67% had taken aspirin, while $\frac{171}{270}$ or 63.33% had taken the placebo.

Of the 21,778 men who had no heart attack, $\frac{10,933}{21,778}$ or 50.2% had taken aspirin, while $\frac{10,845}{21,778}$ or 49.8% had taken the placebo.

It seems clear that in men there is a relationship between taking an aspirin every other day and the risk of having a heart attack.

Two Investigative Tasks

1. The graduate school at the University of California at Berkeley reported that in 1973 they accepted 44% of 8442 male applicants and 35% of 4321 female applicants. Concerned that one of their programs was guilty of gender bias, the graduate school analyzed admissions to the six largest graduate programs and obtained the following results:

Program	Men Accepted	Men Rejected	Women Accepted	Women Rejected
A	511	314	89	19
B	352	208	17	8
C	120	205	202	391
D	137	270	132	243
E	53	138	95	298
F	22	351	24	317

   a. Find the percentage of men and the percentage of women accepted by each program. Comment on any pattern or bias you see.
   b. Find the percentage of men and the percentage of women accepted overall by these six programs. Does this appear to contradict the results from part *a*?
   c. If you worked in the Graduate Admissions Office, what would you say to an inquiring reporter who is investigating gender bias in graduate admissions?

2. Suppose you are a baseball scout and are interested in comparing two relief pitchers, Curver and Knuckler, both of whom play college ball. In particular you are interested in their strikeout ability. A quick glance at their records shows you that each came in to relieve in a similar number of games and that each faced 150 batters during his last season. However, Curver struck out 95 batters, while Knuckler struck out only 85 batters. When you tell their coaches that you're probably going to pick Curver, Knuckler's coach asks you to please look at their separate records against right- and left-handed batters. You do the required calculations and are surprised to find out that Knuckler has a higher strikeout rate against right-handed batters than does Curver, and that Knuckler also has a higher strikeout rate against left-handed batters than does Curver!

   Construct a set of hypothetical data to show how this is possible and explain the apparent paradox. *Hint:* One way of doing this involves having one pitcher face many more right-handed batters than the second, while the second faces many more left-handed batters than the first.

Answers Explained

1. *a.*

Program	Percentage of Men Accepted (%)	Percentage of Women Accepted (%)
A	62	82
B	63	68
C	37	34
D	33	35
E	28	24
F	6	7

There doesn't appear to be any real pattern; however, women seem to be favored in four of the programs, while men seem to be slightly favored in the other two programs.

b. Overall, 1195 out of 2691 male applicants were accepted, for a 44% acceptance rate, while 559 out of 1835 female applicants were accepted, for a 30% acceptance rate. This appears to contradict the results from part *a*.

c. You should tell the reporter that while it is true that the overall acceptance rate for women is 30% compared to the 44% acceptance rate for men, program by program women have either higher acceptance rates or only slightly lower acceptance rates than men. The reason behind this apparent paradox is that most men applied to programs A and B, which are easy to get into and have high acceptance rates. However, most women applied to programs C, D, E, and F, which are much harder to get into and have low acceptance rates.

2. One possible set of data is

Against right-handed batters

	Curver	Knuckler
Strikeout	80	45
Non-strikeout	20	5

Against left-handed batters

	Curver	Knuckler
Strikeout	15	40
Non-strikeout	35	60

Note that Knuckler struck out $\frac{45}{50}$ or 90% of the right-handed batters he faced, while Curver struck out only $\frac{80}{100}$ or 80% of the right-handed batters he faced. Similarly, Knuckler struck out $\frac{40}{100}$ or 40% of the left-handed batters he faced, while Curver struck out only $\frac{15}{50}$ or 30% of the left-handed batters he faced. So Knuckler did better against right-handed and against left-handed batters than Curver did, even though Curver seemed to do better overall. The reason for the apparent paradox is that both pitchers seem to have more trouble against left-handed batters, and Knuckler faced many more of these than Curver did. Curver was able to fatten up his overall strikeout percentage by pitching more often against the easier right-handed batters!