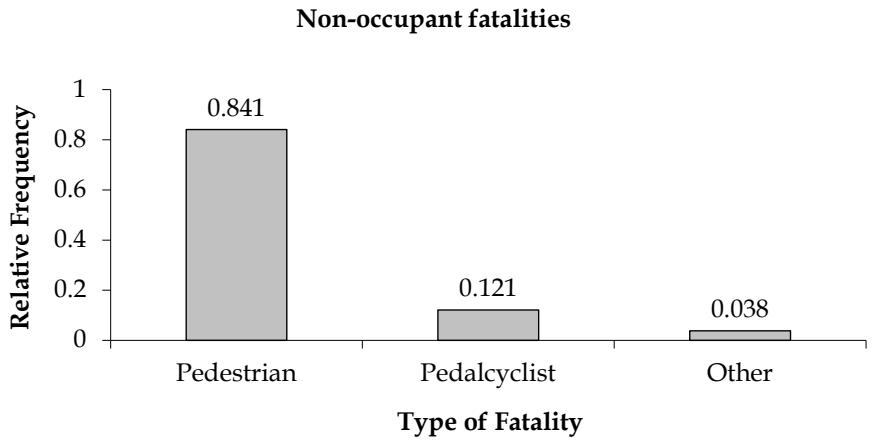
Chapter 2 – Displaying and Describing Categorical Data

Section 2.1

1. Automobile fatalities.

Subcompact and Mini	0.1128
Compact	0.3163
Intermediate	0.3380
Full	0.2193
Unknown	0.0137

2. Non-occupant fatalities.



3. Movie genres.

a) 2008 **b)** 1996 **c)** 2006 **d)** 2012

4. Marriage in decline.

a) People Living Together Without Being Married (ii)

b) Gay/Lesbian Couples Raising Children (iv)

c) Unmarried Couples Raising Children (iii)

d) Single Women Having Children (i)

Section 2.2

5. Movies again.

a) 170/348 ≈ 48.9% of these films were rated R.

b) 41/348 ≈ 11.8% of these films were R-rated comedies.

c) 41/170 ≈ 24.1% of the R-rated films were comedies.

d) 41/90 ≈ 45.6% of the comedies were R-rated.

6. Labor force.

a) 14,824/237,828 ≈ 6.2% of the population was unemployed.

b) 8858/237,828 ≈ 3.7% of the population was unemployed and between 25 and 54.

c) 12,699/21,047 ≈ 60.3% of those 20 to 24 years old were employed.

d) 4378/139,063 ≈ 3.1% of employed people were between 16 and 19.

Chapter Exercises

7. Graphs in the news. Answers will vary.

8. Graphs in the news II. Answers will vary.

9. Tables in the news. Answers will vary.

10. Tables in the news II. Answers will vary.

11. Movie genres.

a) A pie chart seems appropriate from the movie genre data. Each movie has only one genre, and the 193 movies constitute a "whole".

b) "Other" is the least common genre. It has the smallest region in the chart.

12. Movie ratings.

a) A pie chart seems appropriate for the movie rating data. Each movie has only one rating, and the 20 movies constitute a "whole". The percentages of each rating are different enough that the pie chart is easy to read.

b) The most common rating is PG-13. It has the largest region on the chart.

13. Genres, again.

a) SciFi/Fantasy has a higher bar than Action/Adventure, so it is the more common genre.

b) This is easier to see on the bar chart. The percentages are so close that the difference is nearly indistinguishable in the pie chart.

14. Ratings, again.

a) The least common rating was G. It has the shortest bar.

b) The bar chart does not support this claim. These data are for a single year only. We have no idea if the percentages of G and PG-13 movies changed from year to year.

15. Magnet Schools.

There were 1755 qualified applicants for the Houston Independent School District's magnet schools program. 53% were accepted, 17% were wait-listed, and the other 30% were turned away for lack of space.

16. Magnet schools again.

There were 1755 qualified applicants for the Houston Independent School District's magnet schools program. 29.5% were Black or Hispanic, 16.6% were Asian, and 53.9% were white.

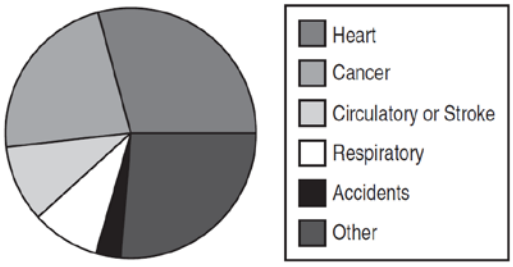
17. Causes of death 2011.

a) Yes. We can add because these categories do not overlap.

(Each person is assigned only one cause of death.)

b) 26.3%

c) Either a bar chart or pie chart with "other" added would be appropriate. A pie chart is shown.

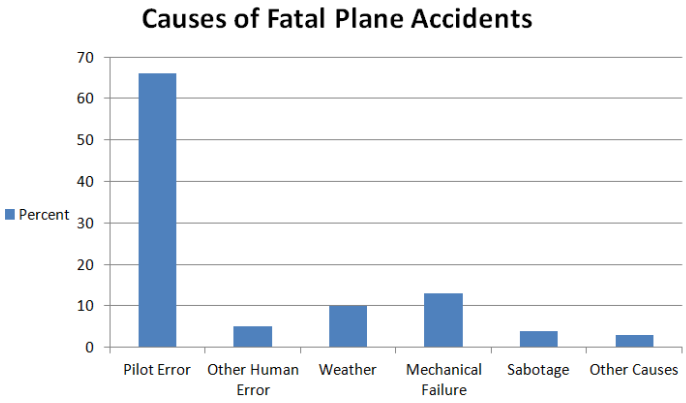


18. Plane crashes.

a) As long as each plane crash had only one cause, it would be reasonable to assume that weather or mechanical failures were the causes of about 23% of crashes.

b) It is likely that the numbers in the table add up to 101% due to rounding.

c) A relative frequency bar chart is a good choice. A pie chart would also be a good display, as long as each plane crash has only one cause.



19. Oil spills as of 2013.

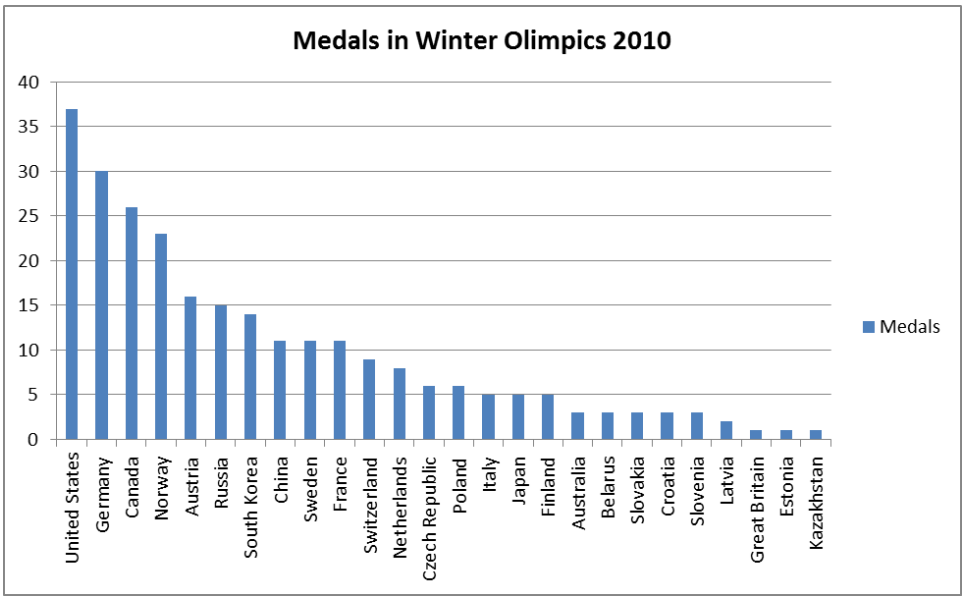
a) Grounding, accounting for approximately 150 spills, is the most frequent cause of oil spillage for these 459 spills. A substantial number of spills, approximately 140, were caused by Collision. Less prevalent causes of oil spillage in descending order of frequency were Hull or equipment failures, Fire & Explosions, and Other/Unknown causes.

b) A pie chart is an appropriate display of the data, since there is only a single cause attributed to each spill, and all spills are represented in some category.

c) There were more spills due to Grounding than Collisions. This is much easier to see on the bar chart.

20. Winter Olympics 2010.

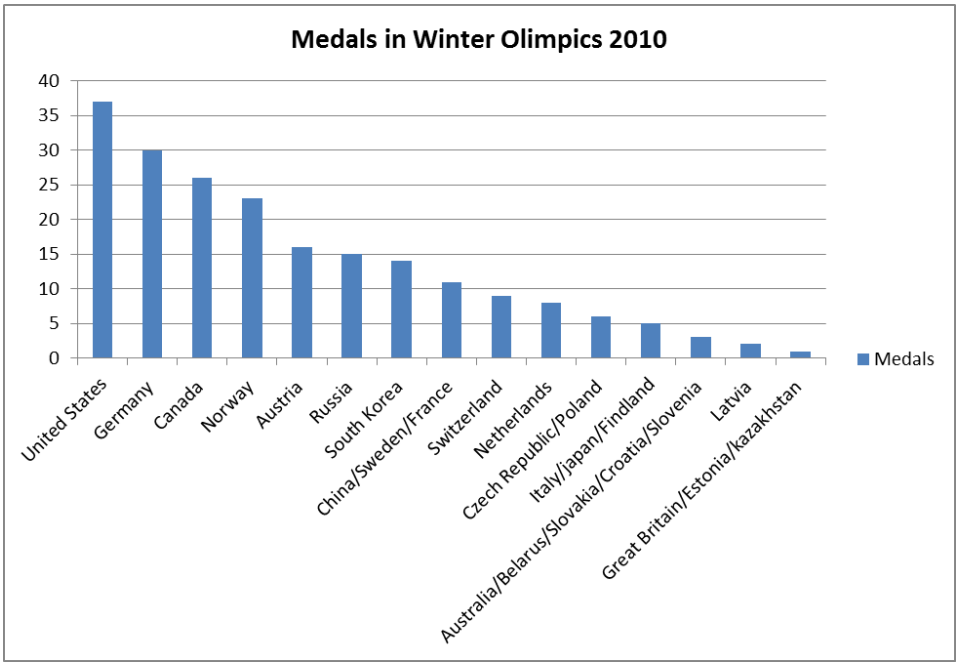
a)



Medals in Winter Olympics 2010

Difficulties: Unit on Y-axis becomes clumsy (as in Figure 1) if we take one unit or two units representing one medal. In Figure 2, 5 units have been taken on Y-axis to represent one meal. Also there are too many categories to display the data.

b) Countries with same number of medals are represented by single bar.



Medals in Winter Olympics 2010

21. Global warming.

Errors in the given pie-diagram are–

i) Showing the pie on a slant violates the area principle and makes it much more difficult to compare fractions of the whole made up of each class—the principal feature that a pie chart ought to show.

ii) In a pie chart, the proportions shown by each slice of the pie must add up to 100% and each individual must fall into only one category. Here the total percentage is 93%.

22. Modalities.

a) The bars have false depth, which can be misleading. This is a bar chart, so the bars should have space between them. Running the labels on the bars from top to bottom and the vertical axis labels from bottom to top is confusing.

b) The percentages sum to 100%. Normally, we would take this as a sign that all of the observations had been correctly accounted for. But in this case, it is extremely unlikely. Each of the respondents was asked to list *three* modalities. For example, it would be possible for 80% of respondents to say they use ice to treat an injury, and 75% to use electric stimulation. The fact that the percentages total greater than 100% is not odd. In fact, in this case, it seems wrong that the percentages add up to 100%, rather than correct.

23. Teen smokers.

According to the Monitoring the Future study, teen smoking brand preferences differ somewhat by region. Although Marlboro is the most popular brand in each region, with about 58% of teen smokers preferring this brand in each region, teen smokers from the South prefer Newports at a higher percentage than teen smokers from the West, 22.5% to approximately 10%, respectively. Camels are more popular in the West, with 9.5% of teen smokers preferring this brand, compared to only 3.3% in the South. Teen smokers in the West are also more likely to have no particular brand than teen smokers in the South. 12.9% of teen smokers in the West have no particular brand, compared to only 6.7% in the South. Both regions have about 9% of teen smokers that prefer one of over 20 other brands.

24. Handguns.

76.4% of handguns involved in Milwaukee buyback programs are small caliber, while only 20.3% of homicides are committed with small caliber handguns. Along the same lines, only 19.3% of buyback handguns are of medium caliber, while 54.7% of homicides involve medium caliber handguns. A similar disparity is seen in large caliber handguns. Only 2.1% of buyback handguns are large caliber, but this caliber is used in 10.8% of homicides. Finally, 2.2% of buyback handguns are of other calibers, while 14.2% of homicides are committed with handguns of other calibers. Generally, the handguns that are involved in

buyback programs are not the same caliber as handguns used in homicides in Milwaukee.

25. Movies by genre and rating.

a) The column totals are 100%.

b) 36.0%

c) 72.2%

d) **i)** 33.3%;

ii) can't tell;

iii) 0%;

iv) can't tell.

26. The last picture show.

a) Since neither the columns nor the rows total 100%, but the table itself totals 100%, these are table percentages.

b) The most common genre/rating combination was the R-rated drama. 17.98% of the 356 movies had this combination.

c) 6.18% of the 356 movies, or 22 movies, were PG-rated comedies.

d) A total of 4.21% of the 356 movies, or 15 movies, were rated G.

e) 4.21% of the movies were rated G, and 19.94% of them were rated PG. So patrons under 13 can see only 24.15% of these movies. This supports the assertion that approximately three-quarters of movies can only be seen by patrons 13 years old or older.

27. Seniors.

a) A table with marginal totals is to the right. There are 268 White graduates and 325 total graduates. $268/325 \approx 82.5\%$ of the graduates are white.

Plans	White	Minority	TOTAL
4-year college	198	44	242
2-year college	36	6	42
Military	4	1	5
Employment	14	3	17
Other	16	3	19
TOTAL	268	57	325

b) There are 42 graduates planning to attend 2-year colleges. $42/325 \approx 12.9\%$

c) 36 white graduates are planning to attend 2-year colleges. $36/325 \approx 11.1\%$

d) 36 white graduates are planning to attend 2-year colleges and there are 268 whites graduates. $36/268 \approx 13.4\%$

e) There are 42 graduates planning to attend 2-year colleges, and 36 of them are white.
$36/42 \approx 85.7\%$

28. Politics.

a) There are 192 students taking Intro Stats. Of those, 115, or about 59.9%, are male.

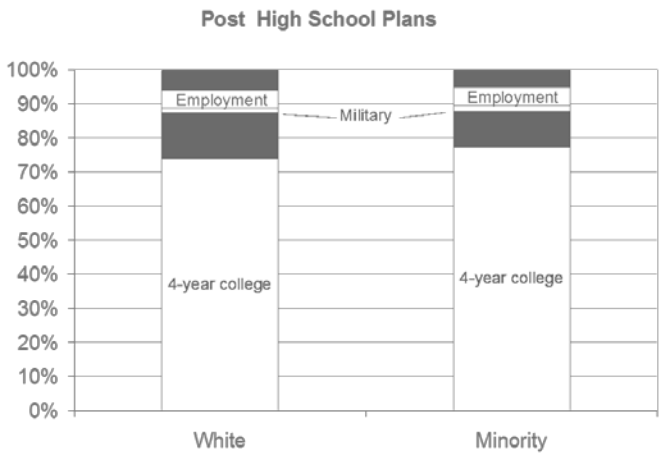
b) There are 192 students taking Intro Stats. Of those, 27, or about 14.1%, consider themselves to be "Conservative".

c) There are 115 males taking Intro Stats. Of those, 21, or about 18.3%, consider themselves to be "Conservative".

d) There are 192 students taking Intro Stats. Of those, 21, or about 10.9%, are males who consider themselves to be "Conservative".

29. More about seniors.

a) For white students, 73.9% plan to attend a 4-year college, 13.4% plan to attend a 2-year college, 1.5% plan on the military, 5.2% plan to be employed, and 6.0% have other plans.



b) For minority students, 77.2% plan to attend a 4-year college, 10.5% plan to attend a 2-year college, 1.8% plan on the military, 5.3% plan to be employed, and 5.3% have other plans.

c) A segmented bar chart is a good display of these data.

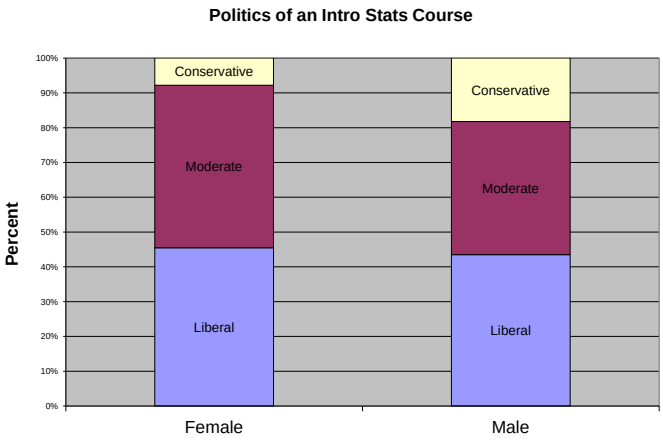
d) The conditional distributions of plans for Whites and Minorities are similar:
White – 74% 4-year college, 13% 2-year college, 2% military, 5% employment, 6% other.
Minority – 77% 4-year college, 11% 2-year college, 2% military, 5% employment, 5% other.
Caution should be used with the percentages for Minority graduates, because the total is so small. Each graduate is almost 2%. Still, the conditional distributions of plans are essentially the same for the two groups. There is little evidence of an association between race and plans for after graduation.

30. Politics revisited.

a) The females in this course were 45.5% Liberal, 46.8% Moderate, and 7.8% Conservative.

b) The males in this course were 43.5% Liberal, 38.3% Moderate, and 18.3% Conservative.

c) A segmented bar chart comparing the distributions is at the right.



d) Politics and sex do not appear to be independent in this course. Although the percentage of liberals was roughly the same for each sex, females had a greater percentage of moderates and a lower percentage of conservatives than males.

31. Magnet schools revisited.

a) There were 1755 qualified applicants to the Houston Independent School District's magnet schools program. Of those, 292, or about 16.6% were Asian.

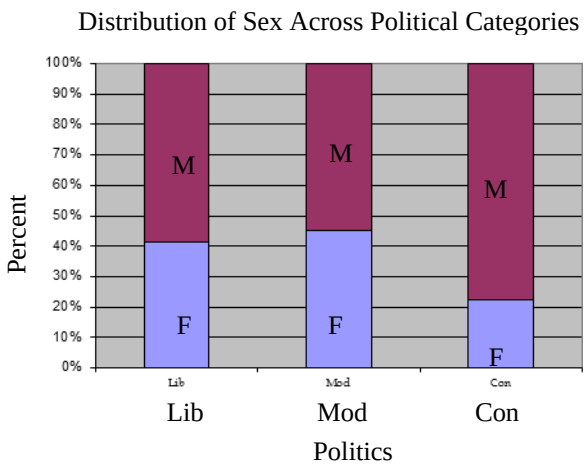
b) There were 931 students accepted to the magnet schools program. Of those, 110, or about 11.8% were Asian.

c) There were 292 Asian applicants. Of those, 110, or about 37.7%, were accepted.

d) There were 1755 total applicants. Of those, 931, or about 53%, were accepted.

32. More politics.

a)



b) The percentage of males and females varies across political categories. The percentage of self-identified Liberals and Moderates who are female is about

twice the percentage of Conservatives who are female. This suggests that *sex* and *politics* are not independent.

33. Back to school.

There were 1,755 qualified applicants for admission to the magnet schools program. 53% were accepted, 17% were wait-listed, and the other 30% were turned away. While the overall acceptance rate was 53%, 93.8% of Blacks and Hispanics were accepted, compared to only 37.7% of Asians, and 35.5% of whites. Overall, 29.5% of applicants were Black or Hispanics, but only 6% of those turned away were Black or Hispanic. Asians accounted for 16.6% of applicants, but 25.3% of those turned away. It appears that the admissions decisions were not independent of the applicant's ethnicity.

34. Parking lots.

a) Percentage of all the cars surveyed were American = (209/356) × 100 ≈ 58.71%.

b) Percentage of the American cars owned by students = (104/209) × 100 ≈ 49.76%.

c) Percentage of the students owned American cars = (104/192) × 100 ≈ 54.17%.

	Driver		
Origin	Student	Staff	Total
American	104	105	209
European	33	11	44
Asian	55	48	103
Total	192	164	356

d) The marginal distribution of origin is given below

	Driver			Marginal
Origin	Student	Staff	Total	distribution
American	104	105	209	209/356 = 0.5870
European	33	11	44	44/356 = 0.1236
Asian	55	48	103	103/356 = 0.2894
Total			356	

e) The conditional distribution of drivers for American cars is given below

		Driver		Total
		Student	Staff	
Origin	American	104	105	209
Conditional distribution		104/209 = 0.4976	105/209 = 0.5024	

f) Let, the origin of the car is independent of the type of driver.

Against the alternative

The origin of the car is not independent of the type of driver.

Here the contingency table is given below

		Driver		Total
		Student	Staff	
Origin	American	104 (112.72)	105 (96.28)	209
	European	33 (23.73)	11 (20.27)	44
	Asian	55 (55.55)	48 (47.45)	103
Total		192	164	356

The test statistics is

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(104 - 112.72)^2}{112.72} + \frac{(105 - 96.28)^2}{96.28} + \frac{(33 - 23.73)^2}{23.73} + \frac{(11 - 20.27)^2}{20.27} + \frac{(55 - 55.5)^2}{55.5} + \frac{(48 - 47.45)^2}{47.45}$$

$$= 0.67 + 0.79 + 3.62 + 4.24 + 0.0045 + 0.064$$

$$= 9.33$$

Since $\chi^2_{cal} = 9.33 > \chi^2_{tab} = 5.991$ (for 2 d.f. al 5% probability level) we reject the null hypothesis and hence the origin of the car is dependent of the type of driver.

35. Weather forecasts.

a) Percentage of days it actually rain = (36/365) × 100 = 9.86%.

b) Percentage of days predicted rain = (92/365) × 100 = 25.21%.

		Actual Weather		
		Rain	No Rain	**Total**
Forecast	Rain	30	62	92
	No Rain	6	267	273
	Total	36	329	365

c) Percentage of the time the correct forecast = ((30 + 267)/365) × 100 = 81.37%.

d) Let, there is no association between the type of weather and the ability of forecasters to make an accurate prediction.

Against the alternative

There is an association between the type of weather and the ability of forecasters to make an accurate prediction.

Here the contingency table is given below

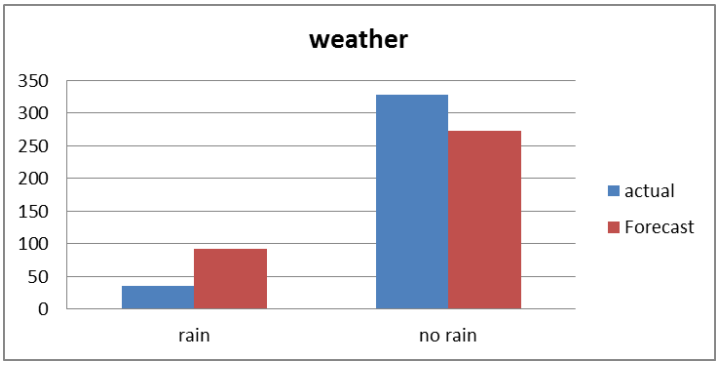
		Actual weather		Total
		Rain	No rain	
Forecast	Rain	30 (9.07)	62 (82.93)	92
	No rain	6 (26.93)	267 (246.07)	273
Total		36	329	365

The test statistics is

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$
$$= \frac{(30-9.07)^2}{9.07} + \frac{(62-82.93)^2}{82.93} + \frac{(6-26.93)^2}{26.93} + \frac{(267-246.07)^2}{246.07}$$
$$= 48.29 + 5.28 + 16.27 + 1.78$$
$$= 71.62$$

Since $\chi^2_{cal} = 71.62 > \chi^2_{tab} = 3.841$ (for 1 d.f. al 5% probability level) we reject the null hypothesis and hence there is an association between the type of weather and the ability of forecasters to make an accurate prediction.

Graphical representation of type of weather



36. Twin births.

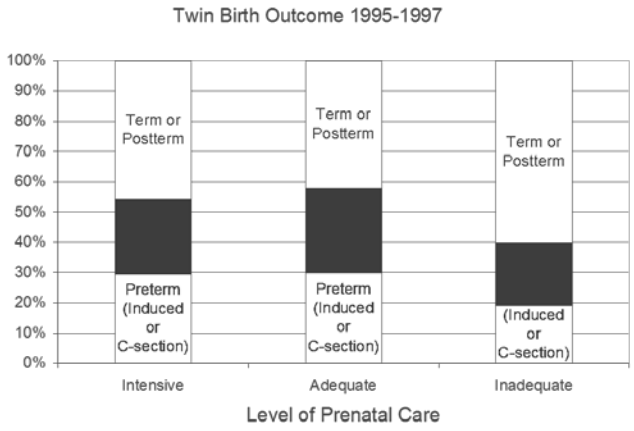
a) Of the 278,000 mothers who had twins in 1995-1997, 63,000 had inadequate health care during their pregnancies. 63,000/278,000 = 22.7%

Twin Births 1995-97 (in thousands)				
Level of Prenatal Care	**Preterm (Induced or Caesarean)**	**Preterm (without procedures)**	**Term or Postterm**	**Total**
Intensive	18	15	28	61
Adequate	46	43	65	154
Inadequate	12	13	38	63
Total	76	71	131	278

b) There were 76,000 induced or Caesarean births and 71,000 preterm births without these procedures. (76,000 + 71,000)/278,000 = 52.9%

c) Among the mothers who did not receive adequate medical care, there were 12,000 induced or Caesarean births and 13,000 preterm births without these procedures. 63,000 mothers of twins did not receive adequate medical care. (12,000 + 13,000)/63,000 = 39.7%

d)



e) 52.9% of all twin births were preterm, while only 39.7% of births in which inadequate medical care was received were preterm. This is evidence of an association between level of prenatal care and twin birth outcome. If these variables were independent, we would expect the percentages to be roughly the same. Generally, those mothers who received adequate medical care were more likely to have preterm births than mothers who received intensive medical care, who were in turn more likely to have preterm births than mothers who received inadequate health care. This does *not* imply that mothers should receive inadequate health care do decrease their chances of having a preterm birth, since it is likely that women that have some complication *during* their pregnancy (that might lead to a preterm birth), would seek intensive or adequate prenatal care.

37. Blood pressure.

a) The marginal distribution of blood pressure level is given below

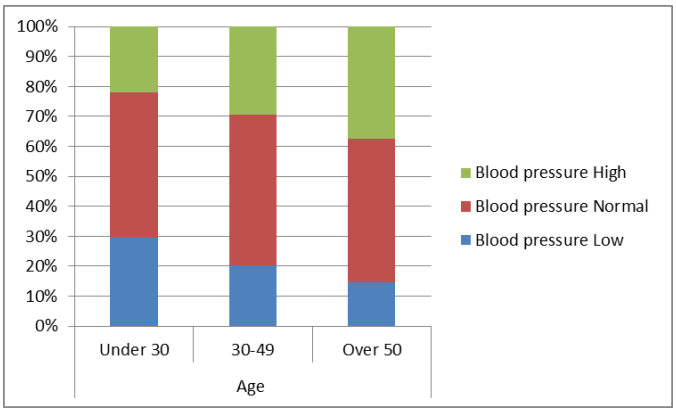
Blood pressure	**under 30**	**30 - 49**	**over 50**	**Total**
low	28	37	29	94
normal	46	92	94	232
high	21	54	74	149
Total	95	183	197	475

		Age				
		Under 30	30-49	Over 50	Total	Marginal distribution
Blood pressure	Low	28	37	29	94	94/475 = 0.1979
	Normal	46	92	94	232	232/475 = 0.4884
	High	21	54	74	149	149/474 = 0.3137

b) The conditional distribution of blood pressure level within each age group is given below

		Age			Total
		Under 30	30-49	Over 50	
Blood pressure	Low	28/94 = 0.2979	37/94 = 0.3936	29/94 = 0.3085	94
	Normal	46/232 = 0.1983	92/232 = 0.3966	94/232 = 0.4052	232
	High	21/149 = 0.1409	54/149 = 0.3624	74/149 = 0.4967	149

c) Compare these distributions with a segmented bar graph.



d) Let, there is no association between age and blood pressure among the employees.

Against the alternative

There is an association between age and blood pressure among the employees.

Here the contingency table is given below

		Age			Total
		Under 30	30-49	Over 50	
Blood pressure	Low	28 (18.8)	37 (36.21)	29 (38.98)	94
	Normal	46 (46.40)	92 (89.38)	94 (96.22)	232
	High	21 (29.80)	54 (57.40)	74 (61.80)	149
	Total	95	183	197	475

The test statistics is

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(28-18.8)^2}{18.8} + \frac{(37-36.21)^2}{36.21} + \frac{(29-38.98)^2}{38.98} + \frac{(46-46.40)^2}{46.40} + \frac{(92-89.38)^2}{89.38} + \frac{(94-96.22)^2}{96.22} + \frac{(21-29.80)^2}{29.80} + \frac{(54-57.40)^2}{57.40} + \frac{(74-61.80)^2}{61.80}$$

$$= 4.50 + 0.02 + 2.56 + 0.0034 + 0.08 + 0.05 + 2.6 + 0.2 + 2.41$$

$$= 12.42$$

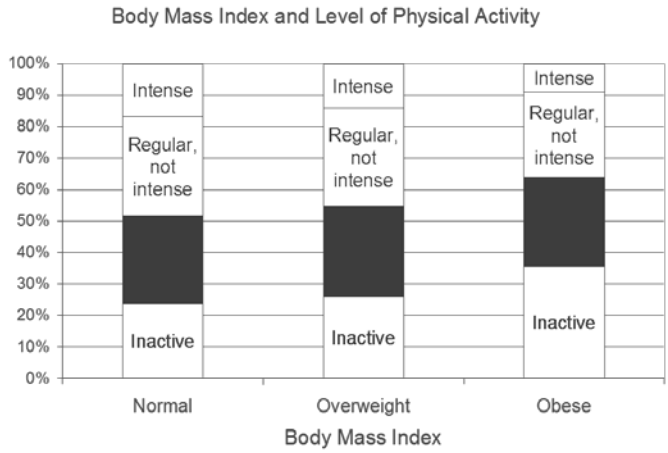
Since $\chi^2_{cal} = 12.42 > \chi^2_{tab} = 9.488$ (for 4 d.f. al 5% probability level) we reject the null hypothesis and hence there is an association between age and blood pressure among the employees.

e) It cannot be said that people's blood pressure increases as their age increases as other variable may also be related to increasing blood pressure.

38. Obesity and exercise.

a) Participants were categorized as Normal, Overweight or Obese, according to their Body Mass Index. Within each classification of BMI (column), participants self reported exercise levels. Therefore, these are column percentages. The percentages sum to 100% in each column, *not* across each row.

b) A segmented bar chart of the conditional distributions of level of physical activity by Body Mass Index category is at the right.



c) No, even though the graphical displays provide strong evidence that lack of exercise and BMI are not independent. All three BMI categories have nearly the same percentage of subjects who report "Regular, not intense" or "Irregularly active", but as we move from Normal to Overweight to Obese we see a decrease in the percentage of subjects who report "Regular, intense" physical activity (16.8% to 14.2% to 9.1%), while the percentage of subjects who report themselves as "Inactive" increases. While it may seem logical that lack of exercise causes obesity, association between variables does not imply a cause-and-effect relationship. A lurking variable (for example, overall health) might influence both BMI and level of physical activity, or perhaps lack of exercise is *caused by* obesity. Only a controlled experiment could isolate the relationship between BMI and level of physically activity.

39. Anorexia.

No, there's no evidence that Prozac is effective. The relapse rates were nearly identical: 31.5% among the people treated with Prozac, compared to 31.1% among those who took the placebo.

40. Antidepressants and bone fractures.

These data provide evidence that taking a certain class of antidepressants (SSRI) might be associated with a greater risk of bone fractures. Approximately 10% of

the patients taking this class of antidepressants experience bone fractures. This is compared to only approximately 5% in the group that were not taking the antidepressants.

41. Driver's licenses 2011.

a) Percentage of total drivers under 20 = (10/208.3) × 100 = 4.8%.

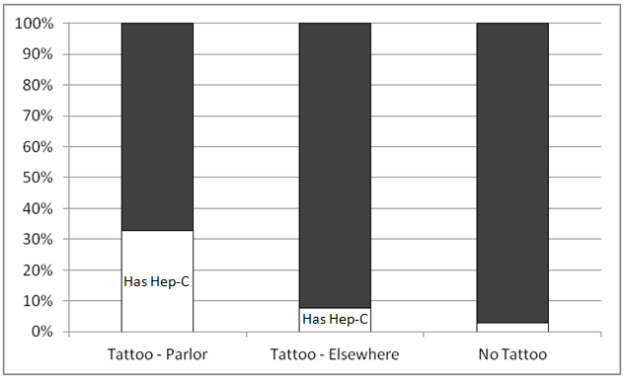
b) Percentage of total male drivers = (103/208.3) × 100 = 49.69%.

c) In each subgroup up to the age of 44 years, percentage of male drives are slightly higher than that of female but this reversed after the age of 44 years.

d) Driver's age and sex appear to be dependent. Younger drivers are slightly more likely to be male and older drivers are slightly more likely to be female.

42. Tattoos.

The study provides evidence of an association between having a tattoo and contracting hepatitis C. Around 33% of the subjects who were tattooed in a commercial parlor had hepatitis C, compared with 12% of those tattooed elsewhere, and only around 3% of those with no tattoo. If having a tattoo and having hepatitis C were independent, we would have expected these percentages to be roughly the same.



43. Hospitals.

a) 165 of 1760, or 9.4%

b) Yes. Major surgery: 13.7% vs. minor surgery: 4.7%.

c) Large hospital: 10.4%; small hospital: 6.9%.

d) Large hospital: Major 13.5% vs. minor 3.8%.

Small hospital: Major 16.7% vs. minor 5.6%.

e) No. Smaller hospitals have a higher rate for both kinds of surgery, even though it's lower "overall."

f) The small hospital has a larger percentage of minor surgeries (88.2%) than the large hospital (32%). Minor surgeries have a lower delay rate, so the small hospital looks better "overall."

44. Delivery service.

a) Pack Rats deliver 700 numbers out of that 28 numbers ate late deliveries i.e. $(28/700) \times 100 = 4\%$ are late deliveries. Boxes R Us deliver 700 numbers out of that 30 numbers ate late deliveries i.e. $(30/700) \times 100 = 4.3\%$ are late deliveries.

b) No. Comparing the individual rate of late deliveries according to type of services, it is seen that Pack Rats has higher rate of late deliveries.

c) This is an instance of Simpson's paradox.

45. Graduate admissions.

a) Percentage of total applicants admitted = $(1284/3014) \times 100 = 42.6\%$

b) Percentage of males admitted = $(1022/2165) \times 100 = 47.20\%$
Percentage of females admitted = $(262/849) \times 100 = 30.86\%$
Hence overall higher percentage of males was admitted.

c) Percentage of males and females admitted in each program are given below.

Program	Males accepted (of applicants)	Females accepted (of applicants)
1	$(511/825) \times 100 = 61.94\%$	$(89/108) \times 100 = 82.41\%$
2	$(352/560) \times 100 = 62.86\%$	$(17/25) \times 100 = 68\%$
3	$(137/407) \times 100 = 33.99\%$	$(132/375) \times 100 = 35.2\%$
4	$(22/373) \times 100 = 5.9\%$	$(24/341) \times 100 = 7.04\%$

Percentage of female applicants accepted is higher in each program.

d) Comparisons of acceptance rate within each program are valid but overall percentage is an unfair average.

46. Be a Simpson!

Answers will vary. The three-way table below shows one possibility. The number of local hires out of new hires is shown in each cell.

	Company A	**Company B**
Full-time New Employees	40 of 100 = 40%	90 of 200 = 45%
Part-time New Employees	170 of 200 = 85%	90 of 100 = 90%
Total	210 of 300 = 70%	180 of 300 = 60%