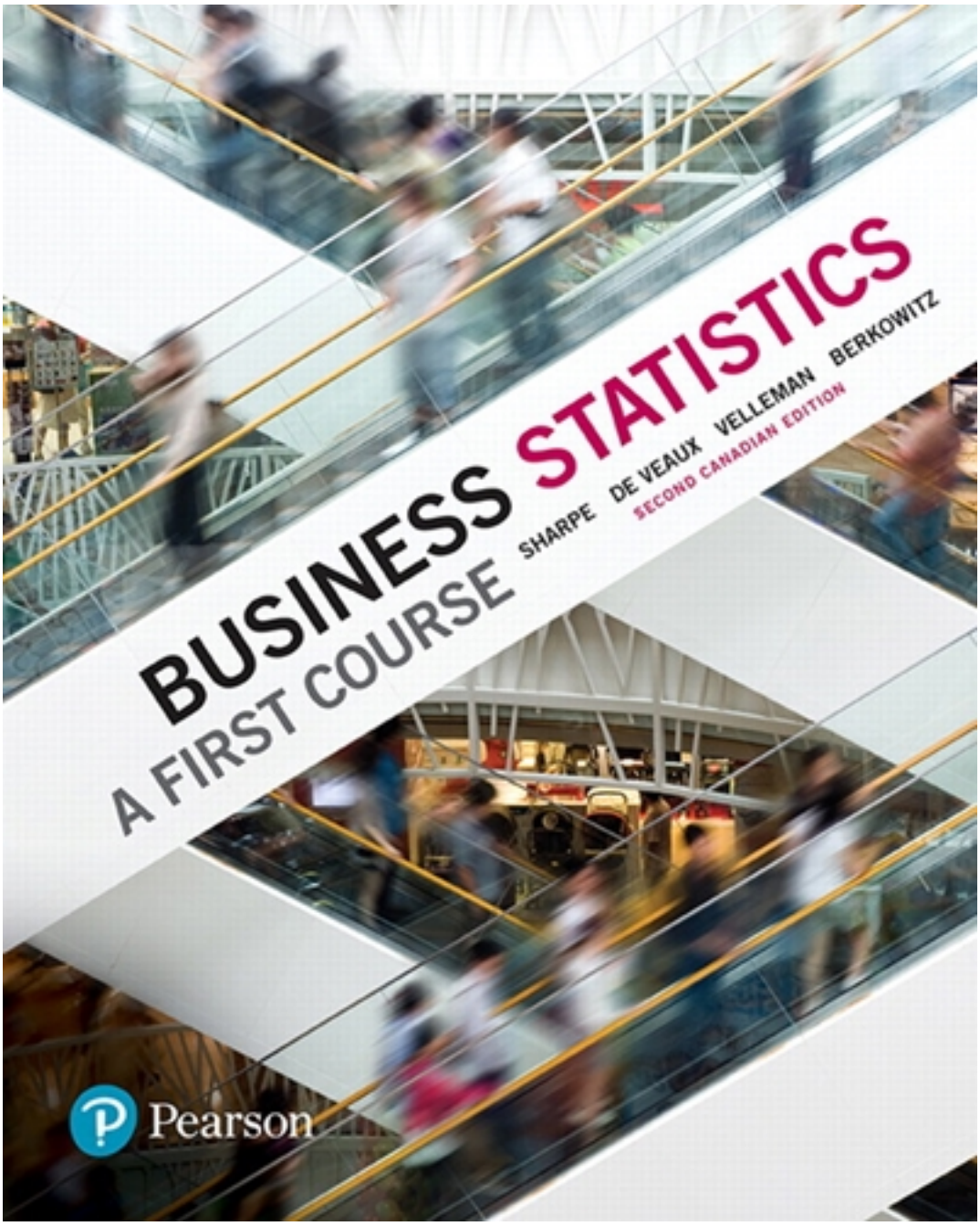
Solutions for Business Statistics A First Course 2nd Edition by Sharpe

CLICK HERE TO ACCESS COMPLETE Solutions



Solutions

SOLUTIONS: Chapter 2 – Displaying and Describing Categorical Data

1. **Graphs in the news.** Answers will vary.

2. **Graphs in the news, part 2.** Answers will vary.

3. **Tables in the news.** Answers will vary.

4. **Tables in the news, part 2.** Answers will vary.

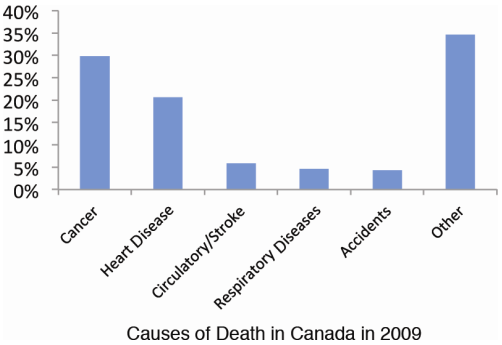
5. **Canadian market share.**
   - **a.** Yes, this is an appropriate display for these data because all categories of one variable (type of non-alcoholic beverage) are displayed. The categories divide the whole, while the category Other combines the smaller shares.
   - **b**. Carbonated soft drinks look to have the same market share as coffee, and both only slightly larger than milk and tea, with bottled water ranked fifth.
   - **c.** The "All Others" category is slightly less than 25%.

6. **World market share.**
   - **a.** Yes, this is an appropriate display for these data. The percentages add up to 100% and there are not too many categories. All categories of one variable (distributors of confectionery products) are displayed. The categories divide the whole and the category "Other" combines the smaller distributors.
   - **b**. The company with the largest share is Cadbury & Kraft who just edges out Mars.

7. **Canadian market share again.**
   - **a.** The pie chart does a better job of comparing portions of the whole.
   - **b**. The "All Others" category is missing and without it, the results could be misleading.

8. **World market share again.**
   - **a.** Either a bar chart or a pie chart would be acceptable, but probably the bar chart does a better job because the "Other" category is so large and takes up almost two thirds of the pie. In addition, the close categories are hard to compare directly because they are so close.
   - **b.** It is very difficult to tell which has bigger market share. You would probably need to look at the bar chart to verify.

9. **Insurance company.**
   - **a.** Yes, it is reasonable to conclude that deaths due to heart OR respiratory diseases is equal to 20.7% plus 4.6%, which equals 25.3%. The percentages can be added because the categories do not overlap. There can only be one cause of death.
   - **b.** The percentages listed in the table only add up to 65.3%. Therefore, other causes must account for 34.7% of Canadian deaths.
   - **c.** An appropriate display could either be a bar graph or a pie graph, using an "Other" category for the remaining 34.7% causes of death.



10. Education levels.

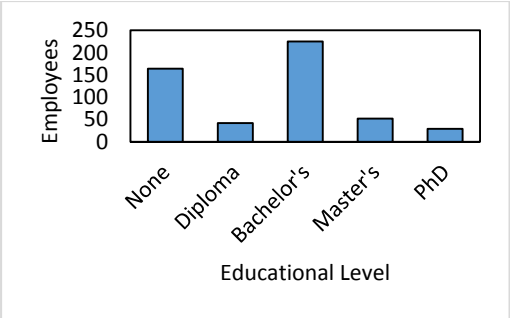
a. Frequency table:

None	Diploma	Bachelor's	Master's	PhD
164	42	225	52	29

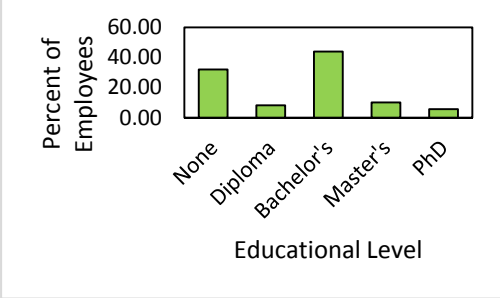
b. Relative frequency table (divide each number by 512 and multiply by 100)

None	Diploma	Bachelor's	Master's	PhD
32.03%	8.20%	43.95%	10.16%	5.66%

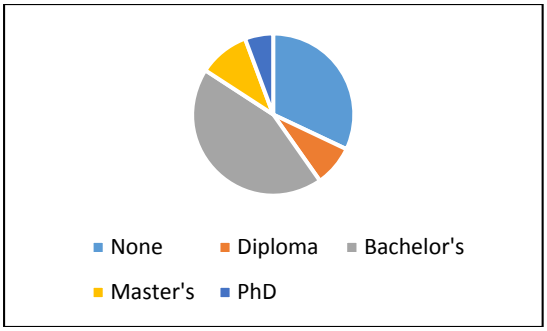
c. Bar chart with counts



d. Relative frequency bar chart



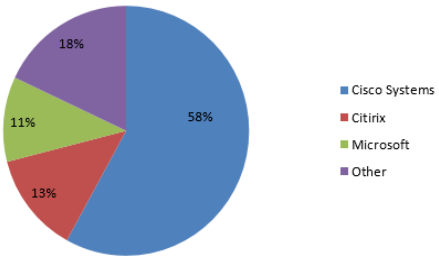
e. Pie chart



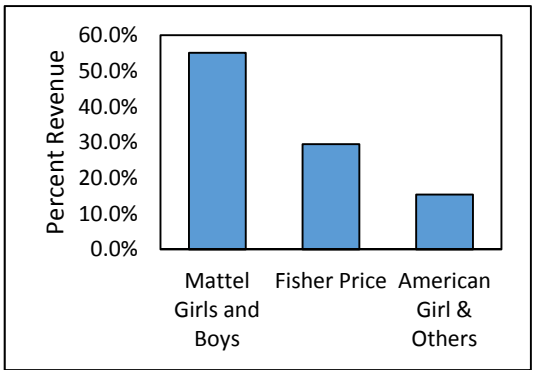
f. Most employees have either a bachelor's degree (44%) or no college degree (32%). About 10% have master's degrees, 8% have associate's degrees, and nearly 6% have PhDs.

g. It is difficult to generalize these results to any other division of the company or to any other company. These data were collected from only one division. Other divisions and companies might have vastly different educational requirements for their employees and therefore distributions of educational levels.

11. SaaS. Cisco systems continues to dominate the market for desktop conferencing. Citrix and Microsoft are battling for second place.



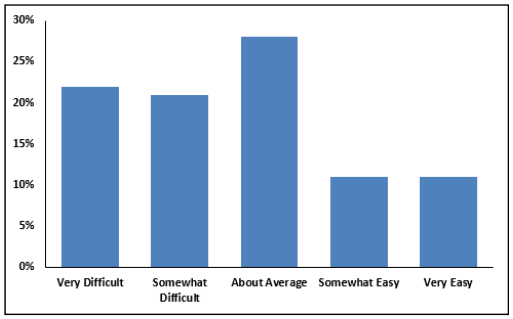
12. Mattel. Mattel received the largest revenue from their Mattel Girls and Boys brand (55.1%). They received 29.5% from their Fisher-Price brand and the rest (15.4%) from their American Girl/ Construction/ Arts&Crafts brand. Either a pie chart or bar chart would be appropriate.



13. Small business financing.

a. The percentages total 93%, not 100%. Others either refused to answer or didn't know.

b. Bar chart:



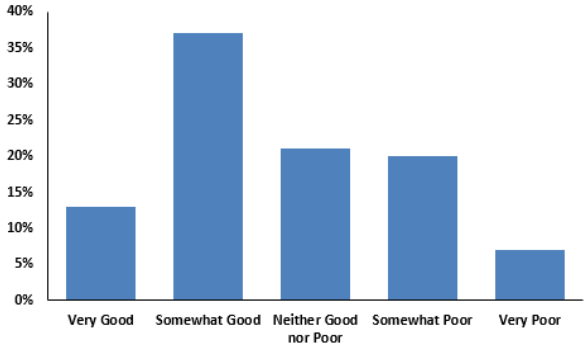
c. A pie chart would not be appropriate because the percentages do not represent parts of a whole and do not total 100% unless an "Other" category is added.

d. (Answers will vary). Nearly half (43%) of business owners said that it would be somewhat or very difficult to obtain credit. Only 22% said it would be somewhat or very easy. Of the remaining, 28% said it would be about average and 7% didn't answer

14. Small business cash flow.

a. The percentages total 98%. The other 2% either didn't answer or didn't know.

b. Bar chart:



c. A pie chart would not be appropriate because the percentages do not represent parts of a whole and do not total 100%. An "Other" category would have to be added.

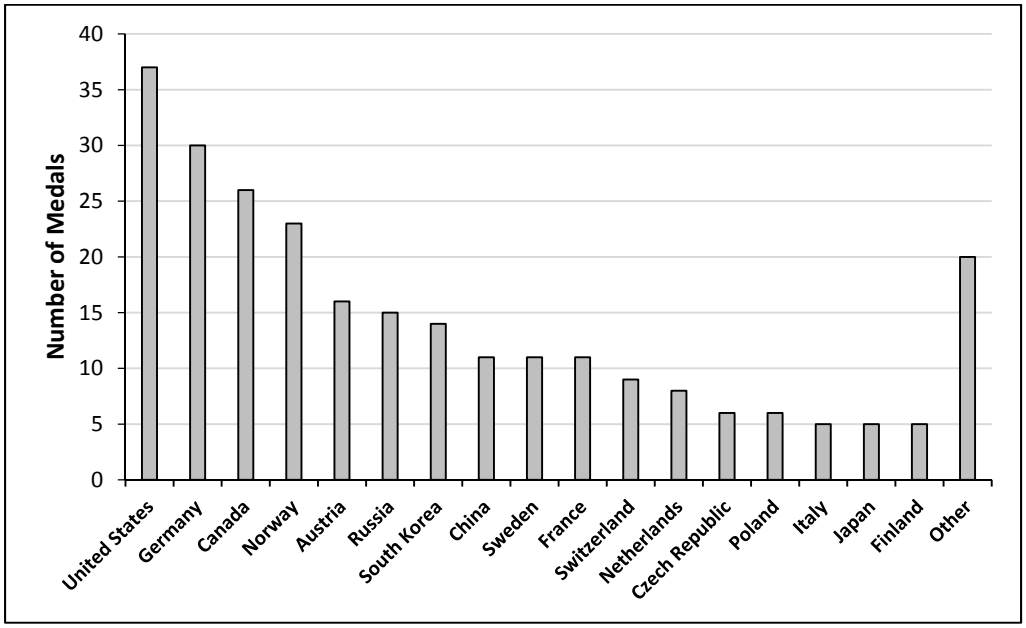
d. (Answers will vary) Half (50%) of the respondents said that their cash flow was very or somewhat good (37% said somewhat). Only 27% said somewhat or very poor.

15. Environmental hazard. The bar chart shows that grounding is the most frequent cause of oil spillage (149) for these 455 spills, while collisions (134) are ranked a close second. The other causes with lower values (18–60) were due to hull failures, fires and explosions, equipment failures, and other or unknown causes. In order to differentiate between close counts, a bar chart is easier to read unless the pie chart gives the actual percentages. Even with the actual percentages the bar chart is easier to read. It is difficult to determine differences between similar areas in the pie chart. To showcase the causes of oil spills as a fraction of all 455 spills, the pie chart could be a reasonable choice.

16. Winter Olympics 2010.

a. There are too many categories to make a meaningful display of the data—too many bars and too many slices in the pie. It would be especially difficult to read the countries with low numbers.

b. One way to simplify the graph would be to showcase only those who earned five or more medals and include the smaller awards under an "Other" category.



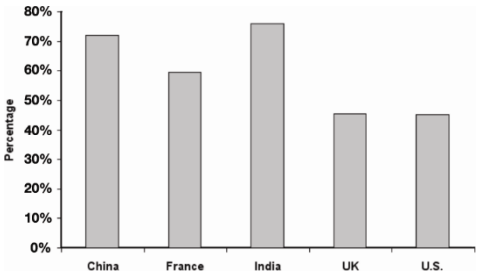
17. Importance of wealth.

a. India 76.1% – USA 45.3% = 30.8%

b. The vertical axis on the display starts at 40%, which makes the comparison between countries difficult and the areas disproportionate. For example, the India bar looks about 5–6 times as big as the USA bar but the actual values are not even twice as big.

c. The display would be improved by starting the vertical axis at 0%, not 40%.

d. The percentage of people who say that wealth is important to them is highest in China and India (over 70%), followed by France (close to 60%) and then the USA and U.K. where the percentages were close to 45%.

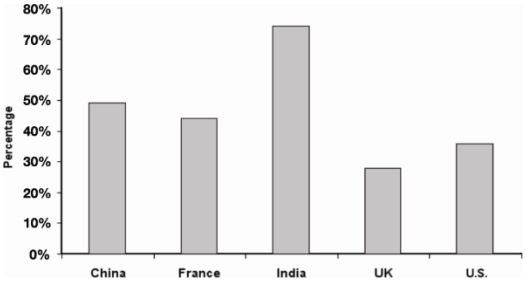


18. Importance of power.

a. The percentages don't add up to 100% so a pie chart is not appropriate. Showing the pie chart three dimensionally on a slant violates the area principle and makes it much more difficult to compare fractions of the whole.

b. A bar chart is more appropriate.

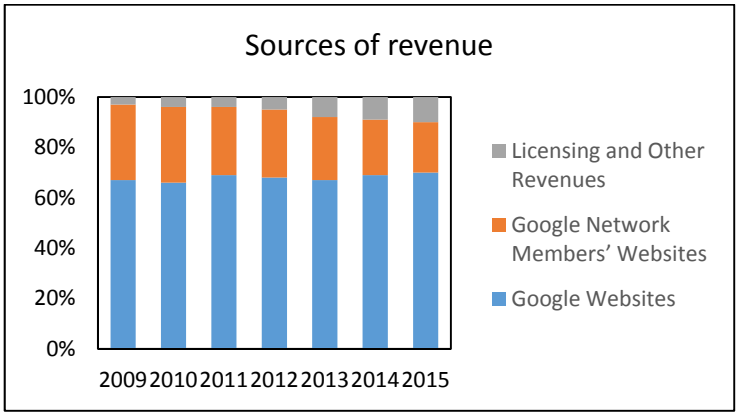
c. The percentage of people who say that power is important to them is highest in India (almost 75%), followed by China (close to 50%) and then France (almost 45%). The lowest percentages occur in U.S. and the U.K. (28%–36%).



19. Google financials.

a. These are column percentages because the column sums add up to 100% and the row percentages add up to more than 100%.

b. A stacked bar chart is appropriate.



c. The main source of revenue for Google is from their own websites, which in 2009 was 67%, and stayed fairly stable through 2015. The second largest source of revenue is from network members' websites. While the Google websites have remained the main source of revenue, the revenue from the Google network websites has been decreasing. Licensing and other revenue has risen from 3% in 2009 to 10% in 2015.

20 Real estate pricing.

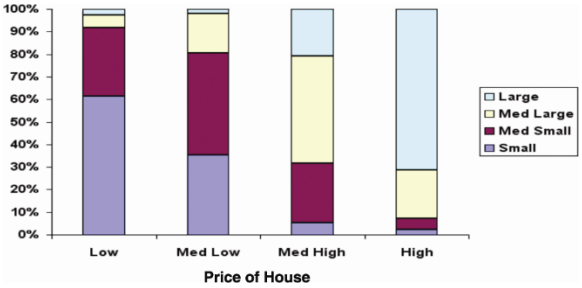
a. These are column percentages because the column sums add up to 100% and the row percentages add up to more than 100%.

b. 2.4%

c. This cannot be determined. We are only given the percentages of size within each *Price* category.

d. Small 61.5% + Med Small 30.4% = 91.9%.

e. The higher the Price of a house, the higher the Size. A stacked bar chart is shown below illustrating the changing conditional distributions.



21. Stock performance.

a. There does not appear to be much, if any, relationship between the performance of a stock on a single day and its performance over the previous year. 45.1% ([164+48]/470)

b. 34.9% (164/470)

c. 5.3% (25/470)

d. 59.8% ([48+233]/470)

e. 41.3% (164/397)

f. 65.8% (48/[48+25])

g. Companies that reported a positive change on October 24 were more likely to report a negative change for the year than companies who reported a negative change on October 24

22. New product.

a. 4.0% (56/1415)

b. 34% (481/1415)

c. 3.7% (18/481)

d. 32.1% (18/56)

e. Marginal Distributions—total % of the categories: Students 64%; Faculty/Staff 23.9%; Alumni 4%; Town Residents 8.2%.

f. Conditional Distributions—percentages for *Very Likely* column: Students 66.5%; Faculty/Staff 20.4%; Alumni 3.7%; Town Residents 9.4%.

g. The likelihood to buy seems independent of campus group (compare percentages for *Very Likely* in each category). However, there are more students, so focusing advertising in that group may have a greater impact on revenue.

23. Real estate.

a. 25.3% (219/865); 30.2% (265/877)

b. 5.9% (51/865); 6.7% (59/877)

c. Sales increased from 865 to 877, which represents an increase of 1.4%

24. Google financials, part 2.

a. 12.9% (1984/15,338); 16.3% (9047/55,629)

b. 18.5% (2843/15,338); 22.1% (12,282/55,629)

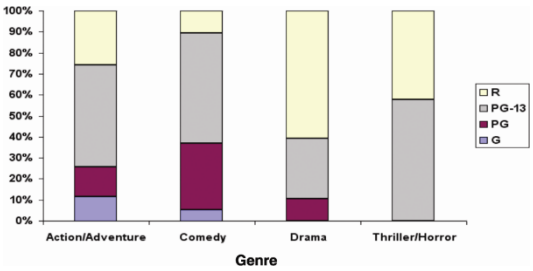
c. As a percentage over time: 10.9%. 10.4%, 10.4%, 10.3%, 10.5%, 11.8%, 11.0% The costs have been relatively consistent except for a slight rise in 2014.

25. Movie ratings.

a. Conditional distribution (in percentages) of movie ratings for action/adventure films: G 11.4%; PG 14.3%; PG-13 48.6%; R 25.7%.

b. Conditional distribution (in percentages) of movie ratings for thriller/horror films: G 0%; PG 0%; PG-13 57.9%; R 42.1%.

c. Stacked bar chart:



d. *Genre* and *Rating* are not independent, in other words, *Genre* and *Rating* are related to each other. Thriller/Horror movies are all PG-13 or R and Drama is similar. For example, there is a 5% chance (6/120) that a randomly selected movie is rated G. However, if you were told that the movie was a Thriller/Horror film, there would be a 0% change that it was rated G. Thus, knowing the genre does affect the rating—they are not independent.

26. Smartphone use.

a.

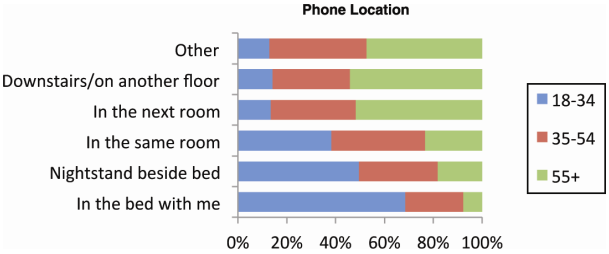
	Age			
Phone location at night	**18–34**	**35–54**	**55+**	**Total**
In the bed with me	24	11	4	39
Nightstand beside bed	162	138	92	392
In the same room	35	46	33	114
In the next room	22	74	129	225
Downstairs/on another floor	22	64	129	215
Other	5	21	29	55
Total	270	354	416	1040

Marginal distribution of *Age*: 18–34: 26.0%, 34.0%, 40.0%
Marginal distribution of *Phone Location*: 3.8%, 37.7%, 11.0%, 21.6%, 20.7%, 5.3%

b. Compute column percentages.

	Age		
Phone location at night	**18–34**	**35–54**	**55+**
In the bed with me	8.9%	3.1%	1.0%
Nightstand beside bed	60.0%	39.0%	22.1%
In the same room	13.0%	13.0%	7.9%
In the next room	8.1%	20.9%	31.0%
Downstairs/on another floor	8.1%	18.1%	31.0%
Other	1.9%	5.9%	7.0%

c. Stacked bar chart:



d. The younger age groups are much more likely to have their phones close by at night. More than two-thirds of the 18–34 group have their phone in or beside the bed, compared with 42% of the 35–54 group and 23% of the 55+ group. The other groups are more likely to have their phones in another room.

27. MBAs.

a. 47.6% (235/494)
b. 45.4% (154/339)
c. 52.3% (81/155)
d. The marginal distribution of region of birth: 47.6% from North America; 36.6% from Asia/Pacific Rim; 8.7% from Europe; 3.8% from the Middle East; and 3.2% from Other.
e. The column percentages:

	Full-time	Part-time	Total
North America	45.4	52.3	47.6
Asia/Pacific Rim	41.0	27.1	36.6
Europe	8.0	10.3	8.7
Middle East	1.8	8.4	3.8
Other	3.8	1.9	3.2
Total	100.0	100.0	100.0

f. They are not independent. Compared with the full-time program, the part-time program has a higher percentage of students born in North America and in the Middle East, but a lower percentage of students born in Asia/Pacific Rim. Thus, knowing the type of MBA program does affect the likelihood of the region of birth of the MBA student.

28. MBAs, part 2.

a. 32.0% (158/494)
b. 32.2% (109/339)
c. 31.6% (49/155)
d. No. The percentage of women in each program is similar, just over 30%.

29. Top producing movies.

a. 3.5% (7/200)
b. 5.0% (1/20)
c. 5.5% (11/200)
d. 58.0% (58/100)
e. 66.0% (66/100)
f. Overall, differences between the two periods are small. However, PG-13 films increased from 55% in 2006–2010 to 58% in 2011–2015 while PG films decreased from 29% in 2006–2010 to 24% in 2011–2015. R films also increased, from 11% to 16%.

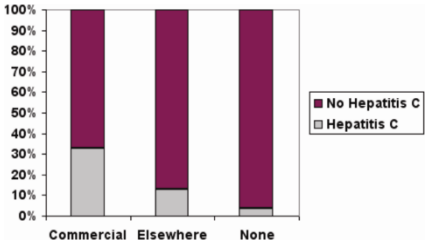
	G	PG	PG-13	R	Total
2011–2015	2%	24%	58%	16%	100%
2006–2010	5%	29%	55%	11%	100%

30. Movie admissions 2013.

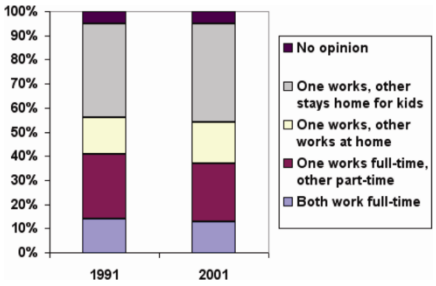
a. 36.4% (65.5/179.8)
b. 62.9% (22.0/35.0)
c. 3.5% (6.3/179.8)
d. 12.3% (4.3/35.1)
e. 2.4% (4.3/179.8)
f. The conditional age distribution: each value is divided by the total for that year. The age distribution of movie admissions stayed fairly constant throughout this time period. The age group 12–17 showed a slight decrease and children's attendance increased in 2013.

	2–11	12–17	18–24	25–39	40–49	50–59	60+
2009	8.78 (2.8/31.9)	17.9	19.7	19.7	14.1	9.09	10.7
2010	8.83 (3.1/35.1)	17.4	21.1	21.9	9.97	8.55	12.3
2011	7.14 (2.5/35)	16.3	18.9	27.7	9.43	8.86	11.7
2012	6.76 (2.8/41.4)	15.2	21.0	23.9	14.0	7.97	11.1
2013	11.8 (4.3/36.4)	15.1	19.8	22.5	8.79	11.5	10.4

31. Tattoos. The study by the medical centre provides evidence of an association between having a tattoo and contracting hepatitis C. Approximately 33% of the subjects who were tattooed in a commercial parlor had hepatitis C, compared with 13% of those tattooed elsewhere, and only 3.5% 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.



32. Working parents. The Gallup poll doesn't provide strong evidence of a change in people's opinions regarding the ideal family in today's society between the years of 1991 and 2001. The conditional distributions of opinion by year appear roughly the same. For example, the percentage of respondents in 1991 who thought the ideal family had two parents that worked full-time was 14%, and in 2001, the percentage was 13%. There does not seem to be strong evidence to conclude that opinions have changed in the two years of study.



33. Education levels II.

a.

	Totals
< 1 Year	95
1–5 Years	205
Over 5 Years	212

b.

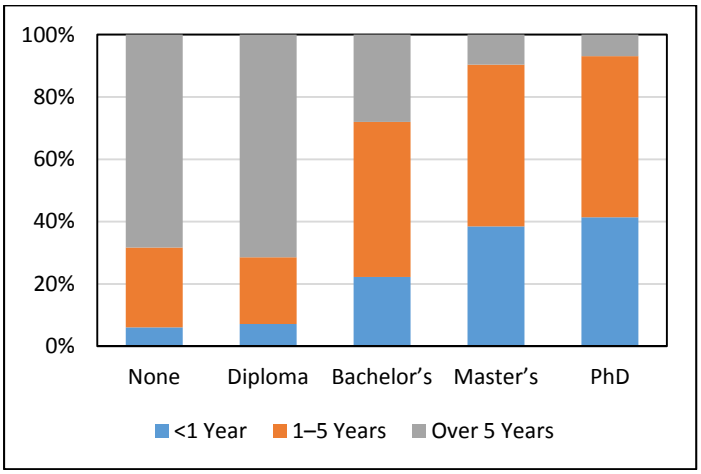
None	Diploma	Bachelor's	Master's	PhD
164	42	225	52	29

c.

(%)	None	Diploma	Bachelor's	Master's	PhD
<1 Year	6.1	7.1	22.2	38.5	41.4
1–5 Years	25.6	21.4	49.8	51.9	51.7
Over 5 Years	68.3	71.4	28.0	9.6	6.9

d. No. The distributions look quite different. More than 2/3 of those with no degree have been with the company longer than 5 years, but almost none of the PhDs (less than 7%) have been there that long. It appears that within the last few years the company has hired better educated employees.

e.



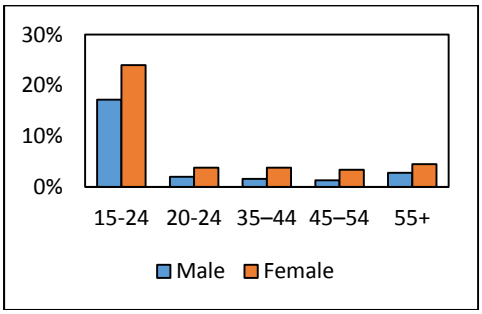
f. It is easier to see the differences in the distributions in the stacked bar chart.

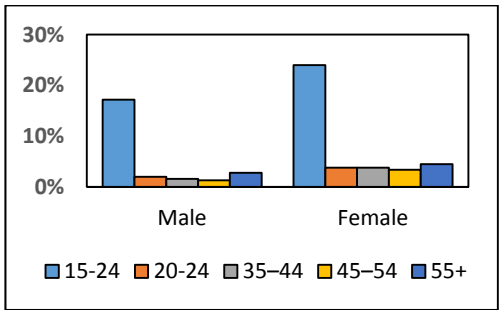
g. A mosaic plot would display the different counts for each degree type. Areas of the plot representing each cell would then reflect the cell counts accurately.

34. Low wage workers.

a. This is not a contingency table; the percentages are neither row nor column percentages. They are the percentages of each subgroup (age by sex) who earn low wages. A stacked bar chart is not appropriate since the percentages do not sum to 100%.

b. A clustered bar chart is appropriate. There are two possibilities.





35. Moviegoers and ethnicity.

a.

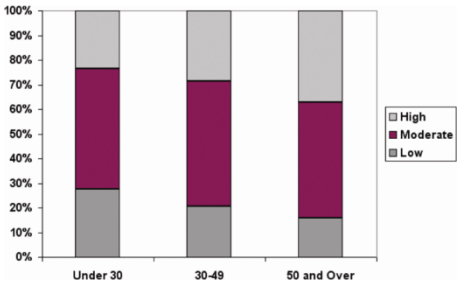
	Caucasian	Hispanic	African-American	Other
Population	66.0% (204.6/310)	16.0% (49.6/310)	12.0% (37.2/310)	6.0% (18.6/310)
Moviegoers	63.0% (88.8/141)	19.0% (26.8/141)	12.0% (16.9/141)	6.0% (8.5/141)
Tickets	56.0% (728/1300)	26.0% (338/1300)	11.0% (143/1300)	7.0% (91/1300)

b. The distributions of moviegoers are quite similar to the population as a whole, but Hispanics appear to buy proportionally more tickets and Caucasians fewer. Hispanics appear to go to the movies more often, on average, than Caucasians.

36. Department store.

a. Low 20.0%; Moderate 48.9%; High 31.0%.

b. Under 30: Low 27.6%; Moderate 49.0%; High 23.5%
30–49: Low 20.7%; Moderate 50.8%; High 28.5%
Over 50: Low 15.7%; Moderate 47.2%; High 37.1%



c. As age increases, the percentage of customers reporting a high frequency of shopping increases, and the percentage who report a low frequency of shopping decreases.

d. No. An association between two variables does not imply a cause-and-effect relationship.

37. Being successful.

a. 66% (18%+48%)

b. It is higher. Young men: 58% (11%+47%)

c. No, because we are not given counts or totals.

d. Young women appear to consider being professionally successful more important in their lives than do young men. Older respondents showed no difference by sex.

e. We don't know the answer. What is your opinion?

38. Advertising.

a. No, the income distributions of households by pet ownership wouldn't be expected to be the same. Caring for a horse is much more expensive, generally, than caring for a dog, cat, or bird. Households with horses as pets would be expected to be more common in the higher income categories.

b. Column percentages (add up to 100%).

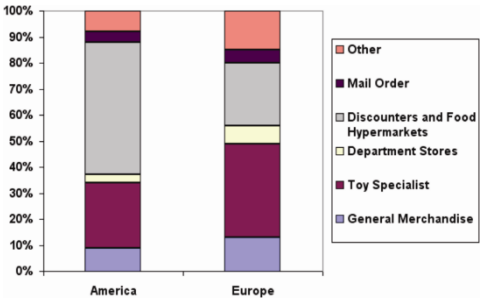
c. No. Among horse owners, there are relatively fewer households in the lowest income bracket and relatively more households in the highest income bracket. In the middle income ranges, the percentages are about the same for each of the different types of pets.

39. Worldwide toy sales.

a. Row percentages (add up to 100%).

b. No. We are given only the conditional distributions. We have no idea how much are sold in either Europe or America.

c.

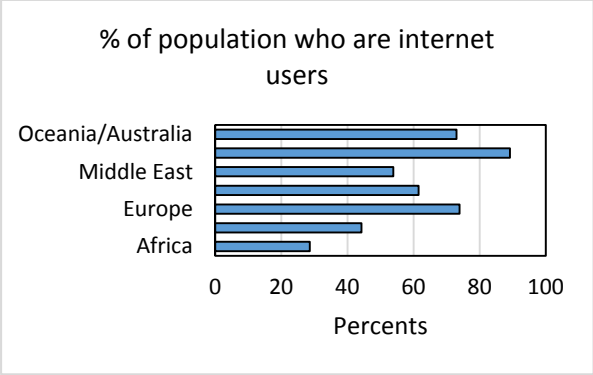


d. In America, more than 50% of all toys are sold by large mass merchant discounters and food hypermarkets and only 25% are sold in toy specialty stores. In Europe, 36% of all toys are sold in toy specialty stores whereas a relatively small 24% are sold in the large discount and hypermarket chains.

40. Internet users.

	Users as percent of population
Africa	28.6%
Asia	44.2%
Europe	73.9%
Latin America/Caribbean	61.5%
Middle East	53.9%
North America	89.1%
Oceania/Australia	73.0%
World Total	49.2%

a. 89.1% (320/359)
b. 8.9% (320/3611)
c. Distribution of users by region:



41. Health care.
a. The marginal totals were added. 160 of 1300 or 12.3% had a delayed discharge.

	Large Hospital	Small Hospital	Total
Major surgery	120 of 800	10 of 50	**130 of 850**
Minor surgery	10 of 200	20 of 250	**30 of 450**
Total	**130 of 1000**	**30 of 300**	**160 of 1300**

b. Major surgery patients were delayed 15.3% of the time. Minor surgery patients were delayed 6.7% of the time.
c. Large Hospital had a delay rate of 13%. Small Hospital had a delay rate of 10%. The small hospital has the lower overall rate of delayed discharge.
d. Large Hospital: Major Surgery 15% and Minor Surgery 5%.
Small Hospital: Major Surgery 20% and Minor Surgery 8%.
e. Yes, while the overall rate of delayed discharge is lower for Small Hospital, Large Hospital did better with *both* major and minor surgery.
f. Small Hospital performs a higher percentage of minor surgeries than major surgeries. 250 of 300 surgeries at Small Hospital were minor (83%). Only 200 of Large Hospitals' 1000 surgeries were minor (20%). Minor surgery had a lower delay rate than major surgery (6.7% to 15.3%), so the small hospital's overall rate was artificially inflated. Large Hospital does better when comparing discharge delay rates.

42. Delivery service.

a. Pack Rats has delivered a total of 28 late packages (12 Regular + 16 Overnight), out of a total of 500 deliveries (400 Regular + 100 Overnight). 28/500 = 5.6% of the packages are late. Boxes R Us has delivered a total of 30 late packages (2 Regular + 28 Overnight) out of a total of 500 deliveries (100 Regular + 400 Overnight). 30/500 = 6% of the packages are late.

b. The company should have hired Boxes R Us instead of Pack Rats. Boxes R Us only delivers 2% (2 out of 100) of its Regular packages late, compared to Pack Rats, who deliver 3% (12 out of 400) of its Regular packages late. Additionally, Boxes R Us only delivers 7% (28 out of 400) of its Overnight packages late, compared to Pack Rats, who delivers 16% of its Overnight packages late. Boxes R Us is better at delivering Regular and Overnight packages.

c. This is an instance of Simpson's Paradox, because the overall late delivery rates are unfair averages. Boxes R Us delivers a greater percentage of its packages Overnight, where it is comparatively harder to deliver on time. Pack Rats delivers many Regular packages, where it is easier to make an on-time delivery.

43. Graduate admissions.

a. 1284 applicants were admitted out of a total of 3014 applicants. 1284/3014 = 42.6%

b. 1022 of 2165 (47.2%) of males were admitted. 262 of 849 (30.9%) of females were admitted.

c. Because there are four comparisons to make, the table below organizes the percentages of males and females accepted in each program. Females are accepted at a higher rate in every program.

d.

Program	Males Accepted (of applicants)	Females Accepted (of applicants)	Total
1	511 of 825	89 of 108	600 of 933
2	352 of 560	17 of 25	369 of 585
3	137 of 407	132 of 375	269 of 782
4	22 of 373	24 of 341	46 of 714
Total	**1022 of 2165**	**262 of 849**	**1284 of 3014**

Program	Males	Females
1	61.9%	82.4%
2	62.9%	68.0%
3	33.7%	35.2%
4	5.9%	7%

e. The comparison of acceptance rate within each program is most valid. The overall percentage is an unfair average. It fails to take the different numbers of applicants and different acceptance rates of each program. Women tended to apply to the programs in which gaining acceptance was difficult for everyone. This is an example of Simpson's Paradox.

44. Simpson's Paradox. 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%

Mini Case Study Project: Eddie's Hang-up Display

Report:
Data from *Google Analytics* indicate that about half of *Sessions* on Eddie's website originate from British Columbia with about 29% from Alberta and 12% from Ontario. There are some notable differences between March and October; traffic from BC rose from 49% to 55%, with all other regions declining slightly to make up that difference. Only about 10% of traffic comes from elsewhere in Canada. Similar results are seen with *Pages* (# of pages visited). British Columbia also has the largest percentage of *New Users*, but the gap between BC and Alberta or Ontario is slightly smaller. This may indicate increasing visibility in other regions of Canada. Eddie's could perhaps contemplate increased advertising activities or a storefront presence in Ontario. Website activity in British Columbia and Alberta shows similar patterns across the year. Both have increased activity in March and October/November, in preparation for increased retail activity in spring and pre-Christmas. Advertising and marketing campaigns should take account of these peaks.

Frequency Tables: Visits, Pages, New Visits: March 2015

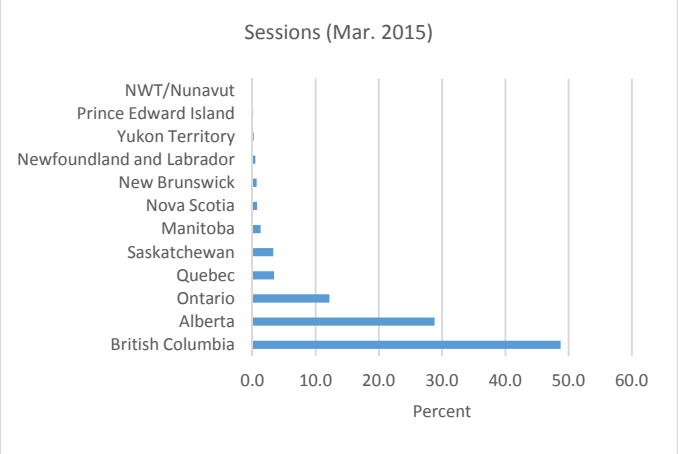
MARCH 2015

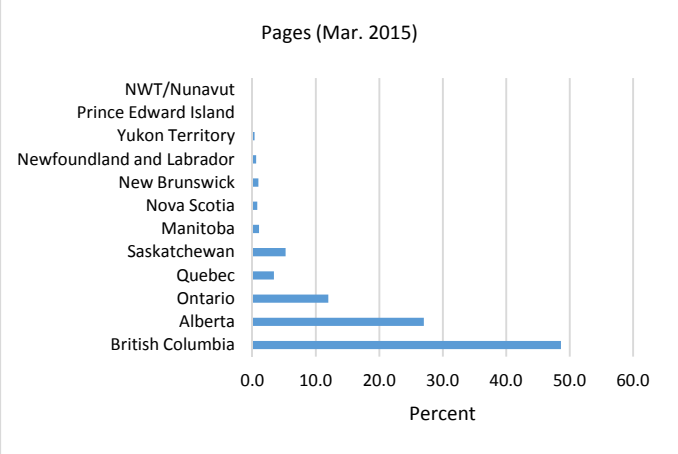
Region	**Sessions**		**Pages**		**New Users**	
	Count	**Pct**	**Count**	**Pct**	**Count**	**Pct**
British Columbia	9,236	48.73	108,523	48.63	4,953	45.01
Alberta	5,461	28.81	60,289	27.02	3,225	29.31
Ontario	2,308	12.18	26,680	11.96	1,602	14.56
Quebec	649	3.42	7,541	3.38	422	3.83
Saskatchewan	624	3.29	11,694	5.24	372	3.38
Manitoba	245	1.29	2,303	1.03	175	1.59
Nova Scotia	142	0.75	1,693	0.76	95	0.86
New Brunswick	126	0.66	2,116	0.95	75	0.68
Newfoundland and Labrador	87	0.46	1,288	0.58	51	0.46
Yukon Territory	38	0.20	761	0.34	12	0.11
Prince Edward Island	20	0.11	115	0.05	16	0.15
NWT/Nunavut	18	0.09	154	0.07	6	0.05
Total	**18,954**	**100.00**	**223,157**	**100.00**	**11,004**	**100.00**

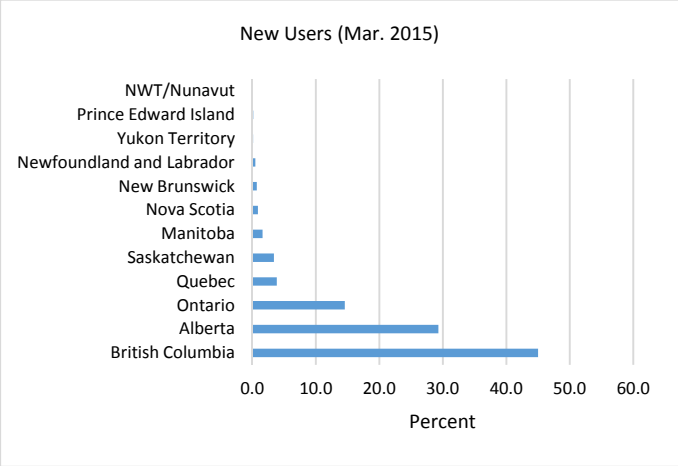
Frequency Tables: Visits, Pages, New Visits: March 2015

	OCTOBER 2015					
Region	**Sessions**		**Pages**		**New Users**	
	Count	**Pct**	**Count**	**Pct**	**Count**	**Pct**
British Columbia	9,713	54.53	86,543	54.22	4,882	51.94
Alberta	4,732	26.57	41,405	25.94	2,432	25.87
Ontario	1,842	10.34	16,136	10.11	1,205	12.82
Quebec	426	2.39	4,026	2.52	259	2.76
Saskatchewan	455	2.55	5,792	3.63	233	2.48
Manitoba	235	1.32	1,455	0.91	134	1.43
Nova Scotia	121	0.68	1,029	0.64	80	0.85
New Brunswick	100	0.56	1,076	0.67	64	0.68
Newfoundland and Labrador	112	0.63	1,306	0.82	59	0.63
Yukon Territory	30	0.17	513	0.32	20	0.21
Prince Edward Island	18	0.10	133	0.08	14	0.15
NWT/Nunavut	28	0.16	211	0.13	18	0.19
	17,812	**100.00**	**159,624**	**100.00**	**9,400**	**100.00**

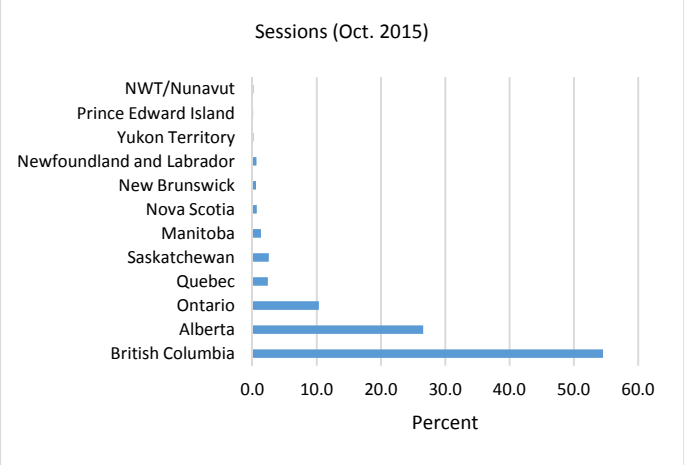
Bar Charts: Sessions, Pages, New Users—March 2015

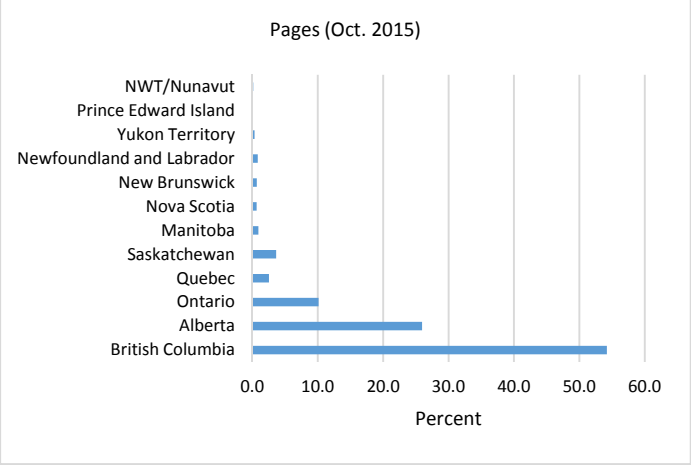


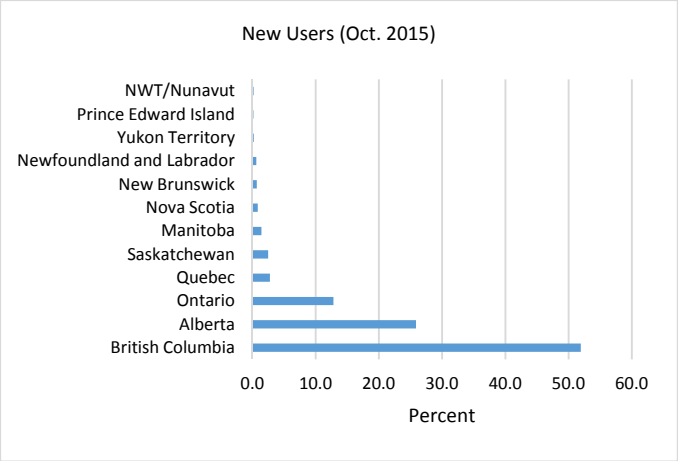




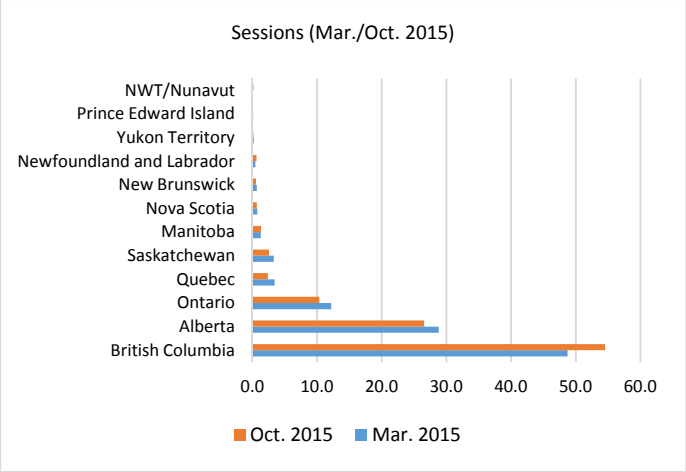
Bar Charts: Sessions, Pages, New Users—October 2015

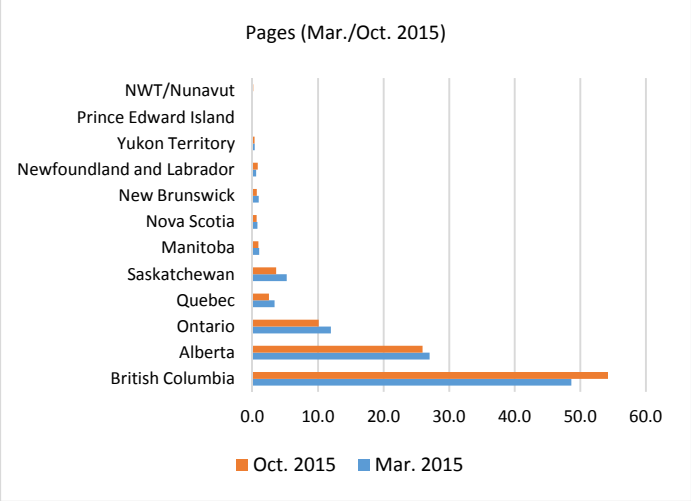


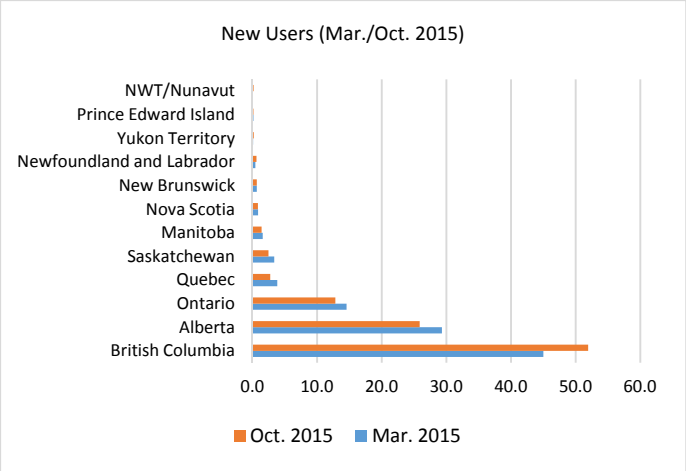




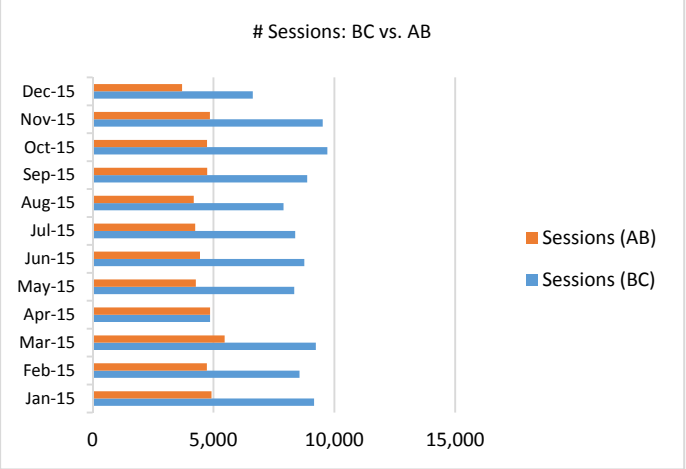
Clustered Bar Charts: Sessions, Pages, New Users—March and October 2015

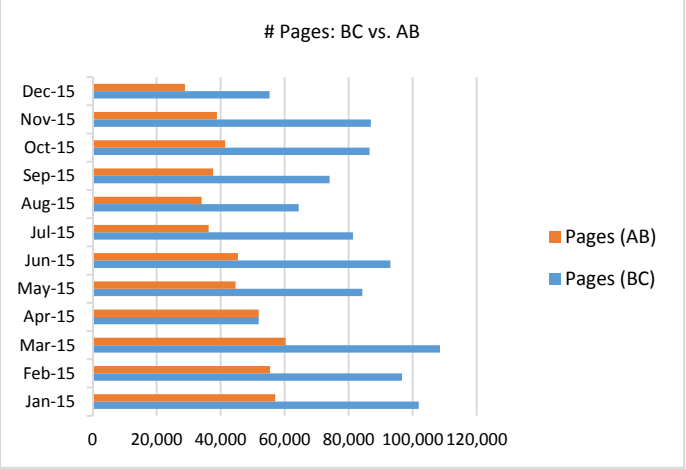


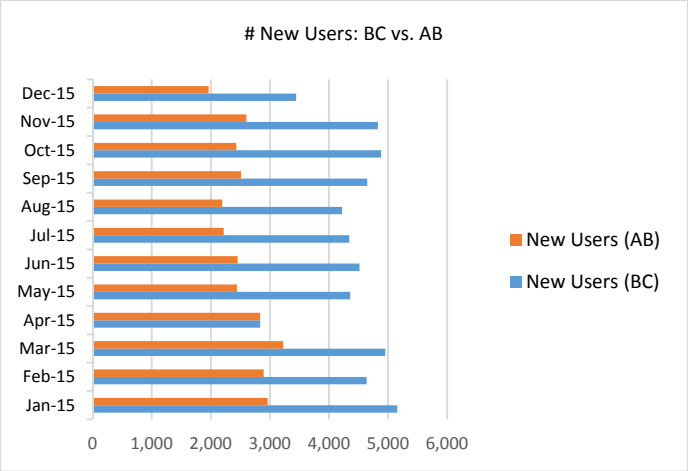




Clustered Bar Charts: # Users,# Sessions, # New Users—BC vs. Alberta, by month, 2015







Chapter 2
Displaying and Describing Categorical Data

What's It About?

We introduce students to distributions of categorical variables. The mathematics is easy (summaries are just percentages) and the graphs are straightforward (pie charts and bar graphs). We challenge students to uncover the story the data tell and to write about it in complete sentences in context.

Then we up the ante, asking them to compare distributions in two-way tables. Constructing comparative graphs, discussing conditional distributions, and considering (informally) the idea of independence give students a look at issues that require deeper thought, careful analysis, and clear writing.

Comments

Most texts do not deal with conditional distributions, independence, and confounding (Simpson's paradox) so soon. Our experience is that students can get lulled into a false sense of security in the early part of this course if all they see is things like means and histograms that they have dealt with since middle school. They think the course is going to be easy, and they may not recognize the level of sophistication that is required until it's too late. The ideas in Chapter 4 are not hard and are introduced only informally, but they do require some thought. Students will find it difficult to make clear explanations. We want these ideas to be interesting, to engage imaginations, and to challenge students. We hope the level of thought required will get their attention and arouse their interest.

It is probably beginning to dawn on your students that this isn't a math class. At the very least, they are going to be expected to write often and clearly. In business, the numeric solution alone is never sufficient. There is always a real-world understanding or conclusion.

Looking Ahead

There are many important skills and ideas here that prepare students for later topics. They need to think about the type of data, checking a condition before plunging ahead. They need to think about what comparisons will answer the questions posed, and write clear explanations in context. They begin to think about independence, one of the most important issues in Statistics. And, in Simpson's paradox, they see the need to think more deeply to avoid being misled by lurking or confounding variables.

Class Do's

Continue to emphasize precision of vocabulary (and notation). These are an important part of clear communication and critical to success.

Emphasize *Plan-Do-Report* right from the start. The key to doing well in Statistics is to plan carefully by understanding the business questions on the table and ask what statistical techniques can address those issues before starting to write an answer. And then, after doing some calculations or other work, to write clear and concise report of what it all means. Your students may rebel at first at having to write sentences, much less paragraphs, in a course they may have thought was a math class. They are used to just doing the *Do*. *Report* is at least 50% of each solution.

If you make that point consistently right from the start of the course it becomes second nature soon, and puts each student in the right mindset for writing solid answers—and for using statistics to make business decisions. Continually remind them: ***Answers are sentences, not numbers.***

Weave the key step of checking the assumptions and conditions into the fabric of doing Statistics. It's easy: have students check that the data are being treated as categorical before they proceed with pie charts, conditional distributions, and the like. As the course goes on, thinking about assumptions and conditions will help students Plan appropriate statistical procedures. Start now.

Discuss categorical data and appropriate summaries: numerical (counts/percentages) and graphical (pie charts, bar graphs). Discuss *distribution, frequency, relative frequency*.

It gets more interesting when we make comparisons (using bivariate data): for example, purchase preferences by gender. Discuss two-way tables, *marginal* and *conditional* distributions. Purchase choices may be interesting, but looking at the differences in purchase choice by gender adds much more to the discussion. You can emphasize the vocabulary by asking things like "What is the marginal frequency distribution of gender?" vs. "What is the conditional relative frequency distribution of gender among customers interested in electronics?"

Make sure students can correctly sort out answers to similar sounding questions:

1. What percent of the class are women who plan to purchase electronics?
2. What percent of the electronics customers are women?
3. What percent of the women are electronics customers?

Raise the issue of independence. It's not formal independence yet, just the general idea that if gender and purchase preferences were independent, the percentages for either gender would mirror the class as a whole, or the percentages of interested in each item would be the same for both genders. If they are not, statisticians would say this indicates that purchase preference is not independent of gender.

Pay attention in each chapter to the What Can Go Wrong? (WCGW) sections. Helping students avoid common pitfalls is one of the keys to success in this course.

Simpson's Paradox is fun, but don't overemphasize it. It's not a critical issue, but it's a good discussion point about making valid comparisons, and not overlooking lurking or confounding variables.

The Importance of What You Don't Say

Probability. You can see that we are patrolling the perimeter of probability. Concepts like relative frequency, conditional relative frequency, and independence cry out for a formal discussion in probabilistic terms. Don't heed the cry. You and we know that we are setting up the habits of thought that students will need for learning about probability. But this isn't the time to discuss the formalities. Or even to say the word "probability" out loud. (Notice that the book doesn't use the term in this chapter at all.) Talk about "relative frequency" instead. In this class probability is a relative frequency, so we are encouraging students to think about the concepts correctly.

We'll get to the basics of probability in Chapter 5.

Class Examples

1. If you collected class data about gender and political view, you can use it here. Help students develop their Plan-Do-Report skills with questions like:
   - What percent of the class are women with liberal political views?
   - What percent of the liberals are women?
   - What percent of the women are liberals?
   - What is the marginal frequency distribution of political views?
   - What is the conditional relative frequency distribution of gender among conservatives?
   - Are gender and political view independent?
2. Is the color distribution of M&Ms independent of the type of candy? Break open bags of plain and peanut M&Ms and count the colors. (Then eat the data…)
3. Simpson's paradox example:
   It's the last inning of an important game. Your team is a run down with the bases loaded and two outs. The pitcher is due up, so you'll be sending in a pinch-hitter. There are 2 batters available
   on the bench. Whom should you send in to bat? First show the students the overall success history of the two players.

Player	Overall	vs LHP	vs RHP
A	33 for 103	28 for 81	5 for 22
B	45 for 151	12 for 32	33 for 119

   A's batting average is higher than B's (.320 vs. .298), so he looks like the better choice. Someone, though, will raise the issue that it matters whether the pitcher throws right- or left- handed. Now add the rest of the table. It turns out that B has a higher batting average against both right- and left-handed pitching, even though his overall average is lower. Students are stunned.

   Here's an explanation. B hits better against both right- and left-handed pitchers. So no matter the pitcher, B is a better choice. So why is his batting "average" lower? Because B sees a lot more right-handed pitchers than A, and (at least for these guys) right-handed pitchers are harder to hit. For some reason, A is used mostly against left-handed pitchers, so A has a higher average.

   Suppose all you know is that A bats .227 against righties and .346 against lefties. Ask the students to guess his overall batting average. It could be anywhere between .227 and .346, depending on how many righties and lefties he sees. And B's batting average may slide between .277 and .375. These intervals overlap, so it's quite possible that A's batting average is either higher or lower than B's, depending on the mix of pitchers they see.

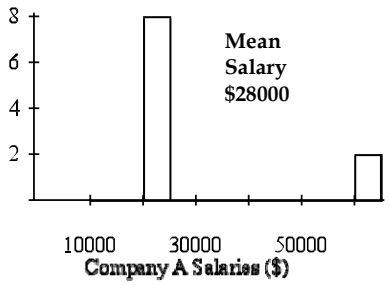
   Pooling the data loses important information and leads to the wrong conclusion. We always should take into account any factor that might matter.

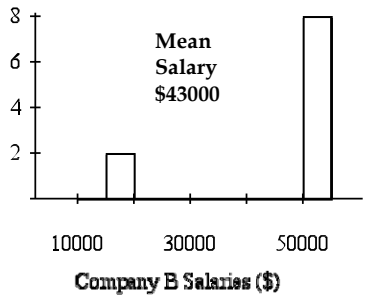
4. Refer to Simpson, again. Here's a nice thought problem to pose to the class; give them a few minutes to work it out. Two companies have labor and management classifications of employees. Company A's laborers have a higher average salary than company B's, as do Company A's managers. But overall company B pays a higher average salary. How can that be? And which is the better way to compare earning potential at the two companies?

Solution:

Make sure you first point out that this example deals with quantitative variables, not categorical. The paradox can be explained when you realize that Company A must employ a greater percentage of laborers than Company B. Also, Company A must employ a smaller percentage of managers than Company B. If laborers earn salaries that are considerably lower than managers, the salaries of Company A's laborers will pull the company average down, and the salaries of Company B's managers will pull the company average up.

The proper way to compare the companies is to use the salaries that are broken down by job type. Using the overall average salary leads to a misleading conclusion.





Resources

ActivStats[*]

- Lesson 4-1: Displaying Categorical Variables - examines displays of categorical data.

[*] ActivStats (0-321-57719-1) can be purchased from Pearson at www.pearsonhighered.com or bundled with your textbook.

Basic Exercises

1. The following data show responses to the question "What is your primary source for news?" from a sample of college students.

Internet	Newspaper	Internet	TV	Internet
Newspaper	TV	Internet	Internet	TV
Newspaper	TV	TV	Newspaper	TV
Internet	Internet	Internet	Internet	Internet
TV	Internet	Internet	TV	TV

a. Prepare a frequency table for these data.
b. Prepare a relative frequency table for these data.
c. Based on the frequencies, construct a bar chart.
d. Based on relative frequencies, construct a pie chart.

2. A cable company surveyed its customers and asked how likely they were to bundle other services, such as phone and Internet, with their cable TV. The following data show the responses.

Very Likely	Unlikely	Unlikely	Very Likely
Likely	Unlikely	Likely	Likely
Unlikely	Unlikely	Likely	Likely
Very Likely	Unlikely	Unlikely	Very Likely
Unlikely	Unlikely	Unlikely	Likely

a. Prepare a frequency table for these data.
b. Prepare a relative frequency table for these data.
c. Based on frequencies, construct a bar chart.
d. Based on relative frequencies, construct a pie chart.

3. A membership survey at a local gym asked whether weight loss or fitness was the primary goal for joining. Of 200 men surveyed, 150 responded fitness and the rest responded weight loss. Of 250 women surveyed, 175 responded weight loss and the rest responded fitness.

a. Construct a contingency table.
b. How many members have fitness as their primary goal for joining the gym?
c. How many members have weight loss as their primary goal for joining the gym?
d. Based on the results, should the owner of the gym emphasize one goal over the other? Explain.

4. The following contingency table shows a random sample of 1000 students from a Western Canada University. The students are classified by major and towns they are from.

Major Program of Study

Home Town	*Biology*	*Accounting*	*History*	*Education*
Edmonton	80	65	55	100
Beaumont	50	40	65	95
Leduc	75	50	45	80
Camrose	65	55	40	40

a. Give the marginal frequency distribution for home state.
b. Give the marginal frequency distribution for major program of study.
c. What percentage of students major in accounting and come from Edmonton?
d. What percentage of students major in education and come from Leduc?

5. The following contingency table shows a random sample of 1000 students from a Western Canada University. The students are classified by major and towns they are from.

Major Program of Study

Home Town	*Biology*	*Accounting*	*History*	*Education*
Edmonton	80	65	55	100
Beaumont	50	40	65	95
Leduc	75	50	45	80
Camrose	65	55	40	40

a. Find the conditional distribution (in percentages) of major distribution for the home town of Beaumont.
b. Find the conditional distribution (in percentages) of major distribution for the home town of Camrose.
c. Construct segmented bar charts for these two conditional distributions.
d. What can you say about these two conditional distributions?

6. The following contingency table shows a random sample of 1000 students from a Western Canada University. The students are classified by major and towns they are from.

Major Program of Study

Home Town	*Biology*	*Accounting*	*History*	*Education*
Edmonton	80	65	55	100
Beaumont	50	40	65	95
Leduc	75	50	45	80
Camrose	65	55	40	40

a. Find the conditional distribution (in percentages) of home town distribution for the biology major.
b. Find the conditional distribution (in percentages) of home town distribution for the education major.
c. Construct segmented bar charts for these two conditional distributions.
d. What can you say about these two conditional distributions?

ANSWERS

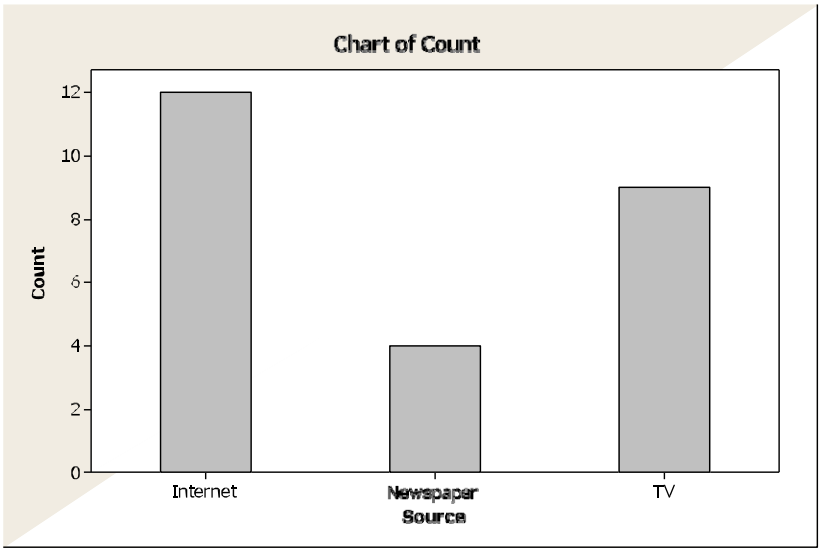
1. a.

News Source	***Number of Students***
Internet	12
Newspaper	4
TV	9

b.

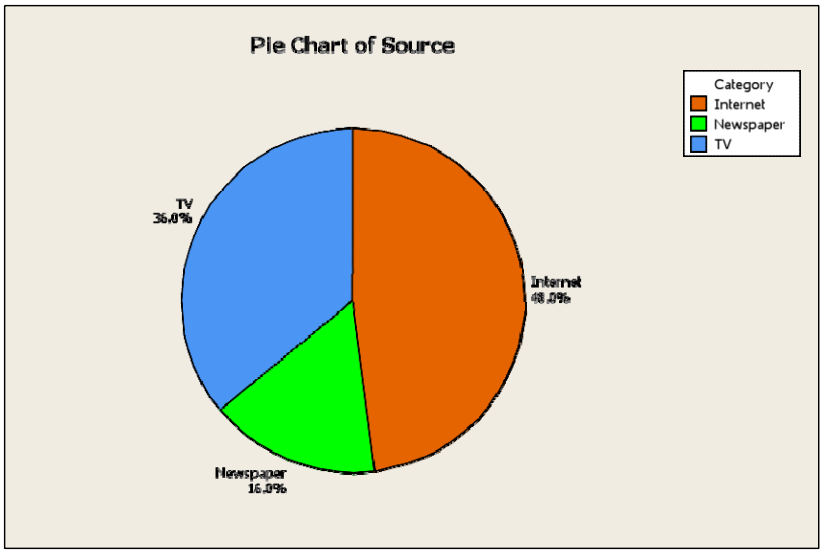
News Source	***% of Students***
Internet	48 %
Newspaper	16 %
TV	36 %

c.



2-8 Part 1 Exploring And Collecting Data

d.



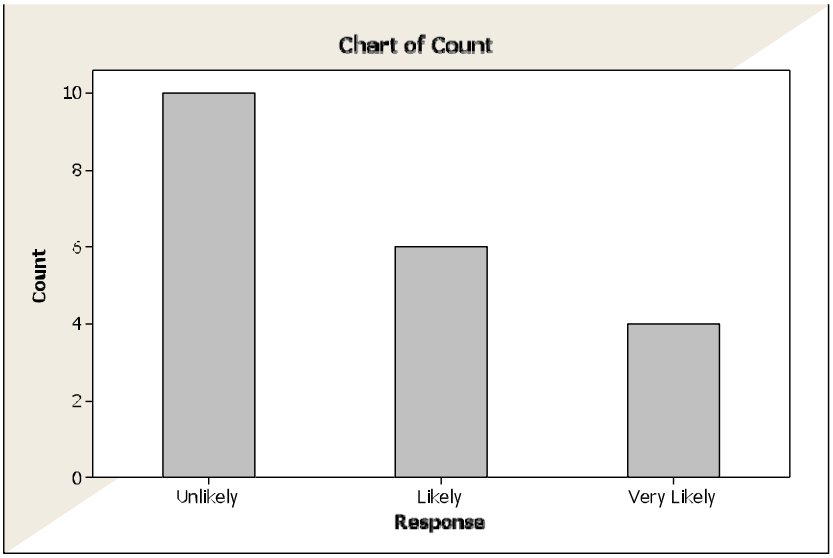
2. a.

Response	***Number of Consumers***
Unlikely	10
Likely	6
Very Likely	4

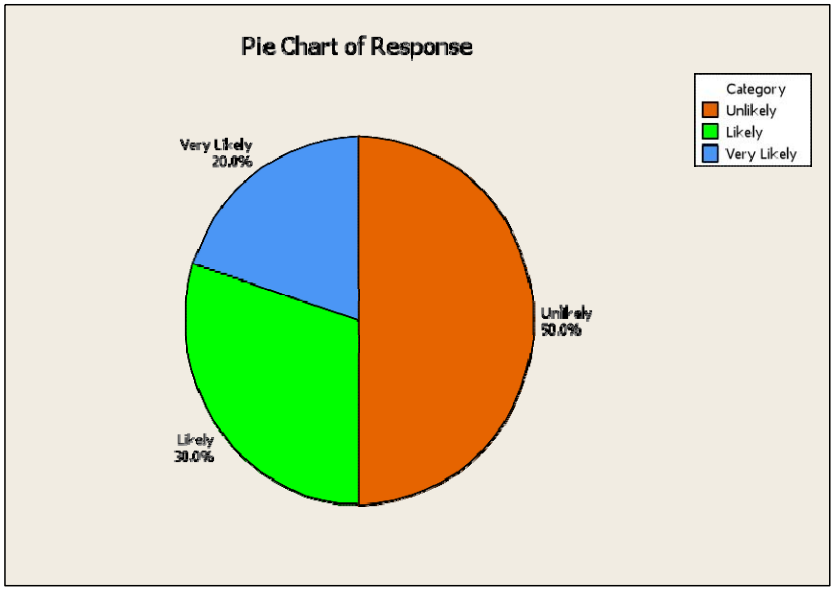
b.

Response	***% of Consumers***
Unlikely	50 %
Likely	30 %
Very Likely	20 %

c.



d.



3. a.

Goal for Gym Membership

Gender	Fitness	Weight Loss	Total
Men	150	50	200
Women	75	175	250
Total	225	225	450

b. 225
c. 225
d. No. 50% of the membership is pursuing each goal.

4. a.

Home Town	***Number of Students***
Edmonton	300
Beaumont	250
Leduc	250
Camrose	200

b.

Major	***Number of Students***
Biology	270
Accounting	210
History	205
Education	315

c. 6.5 %
d. 8 %

2-10 Part 1 Exploring And Collecting Data

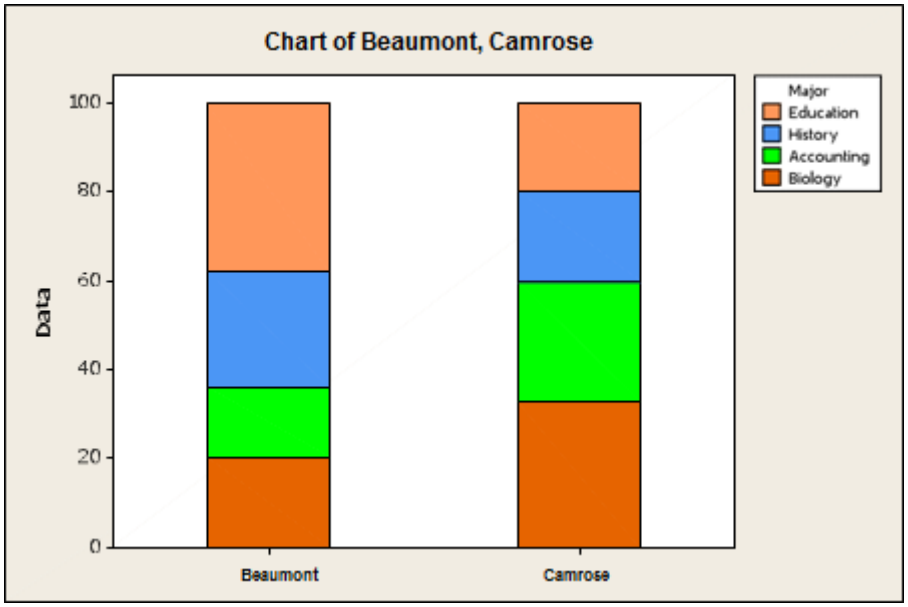
5. a.

Major	***Conditional for Beaumont***
Biology	20 %
Accounting	16 %
History	26 %
Education	38 %

b.

Major	***Conditional for Camrose***
Biology	32.5 %
Accounting	27.5 %
History	20 %
Education	20 %

c.



d. More biology and accounting majors come from Camrose compared to Beaumont

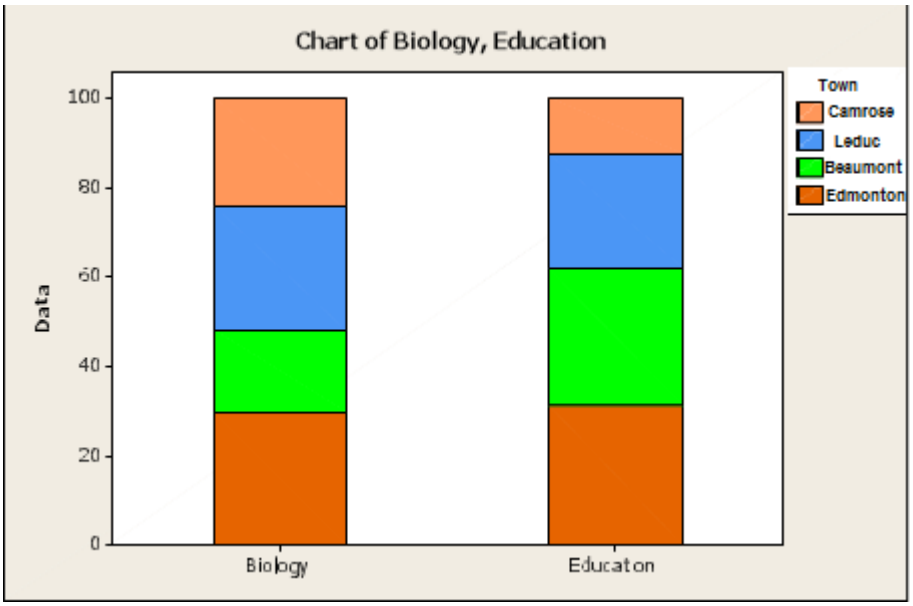
6. a.

Home Town	***Conditional for Biology***
Edmonton	29.6 %
Beaumont	18.5 %
Leduc	27.8 %
Camrose	24.1 %

b

Home Town	***Conditional for Education***
Edmonton	31.7 %
Beaumont	30.2 %
Leduc	25.4 %
Camrose	12.7 %

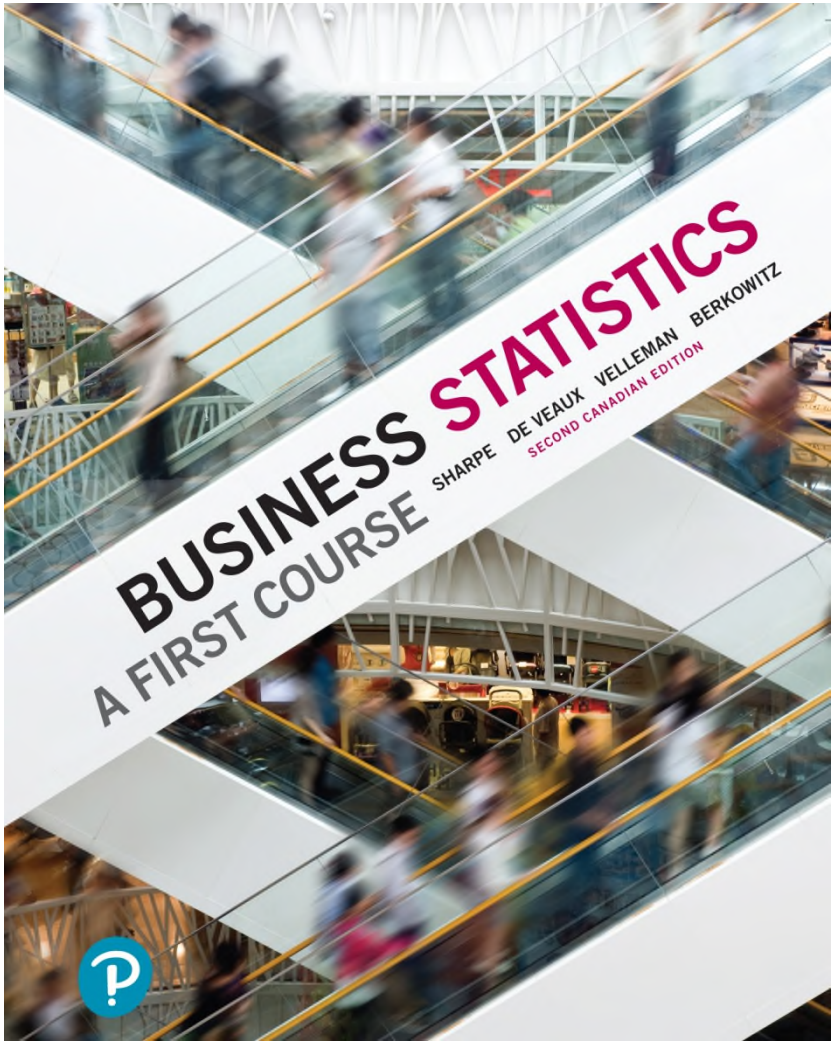
c.



d. Fewer education majors are from Camrose and more are from Beaumont compared with biology majors.

Business Statistics: A First Course

Second Canadian Edition



Chapter 2

Displaying and Describing Categorical Data

Learning Objectives

1. Choose an appropriate display of categorical data and determine its effectiveness
2. Analyze a contingency table of counts or percentages
3. Create and analyze relative frequency distributions from tabulated data
4. Compute and interpret marginal and conditional distributions from contingency tables
5. Identify misleading results that are due to data aggregation (Simpson's paradox)

2.1 The Three Rules of Data Analysis

The Three Rules of Data Analysis:

1. Make a picture. 2. Make a picture. 3. Make a picture.

Pictures …

- reveal things that can't be seen in a table of numbers.
- show important features and patterns in the data.
- provide an excellent means for reporting findings to others.

2.2 Frequency Tables (1 of 3)

A *frequency table* organizes data by recording totals and category names as in the table below.

The names of the categories label each row in the frequency table.

2.2 Frequency Tables (2 of 3)

Table 2.1 Frequency table of organic search traffic to MEC.ca, Jan. 1–Dec. 31, 2012, by province. An organic search visit originates from a search engine, not from an advertisement.

Province	Organic Search Visits
British Columbia	1 609 160
Alberta	1 031 830
Manitoba	208 185
Ontario	2 108 643
Quebec	1 441 269
Nova Scotia	138 393
Total	**6 537 470**

Source: Based on MEC and Google Analytics, Feb. 2013

2.2 Frequency Tables (3 of 3)

A *relative frequency table* displays the *percentages,* rather than the counts, of the values in each category. (See the table below.)

Table 2.2 A relative frequency table for the same data.

Province	Organic Search Visits (%)
British Columbia	24.61%
Alberta	15.78%
Manitoba	3.18%
Ontario	32.25%
Quebec	22.05%
Nova Scotia	2.12%
Total	**100.00%**

2.3 Displaying a Categorical Variable (1 of 5)

The Area Principle

The best data displays observe the *area principle*: the area occupied by a part of the graph should correspond to the magnitude of the value it represents.

2.3 Displaying a Categorical Variable (2 of 5)

Bar Charts

A *bar chart* displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison.

The bar graph here gives a more *accurate* visual impression of the distribution.

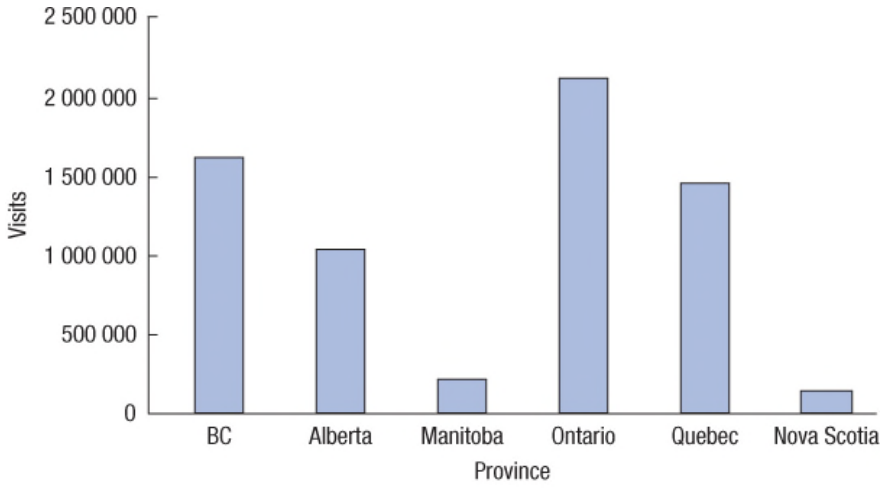


Figure 2.3 Visits to MEC website by *Province*. With the area principle satisfied, the true distribution is clear.

Source: Based on Myanmar Economic Corporation

2.3 Displaying a Categorical Variable (3 of 5)

Bar Charts

If the counts are replaced with percentages, the data can be displayed in a *relative frequency bar chart*.

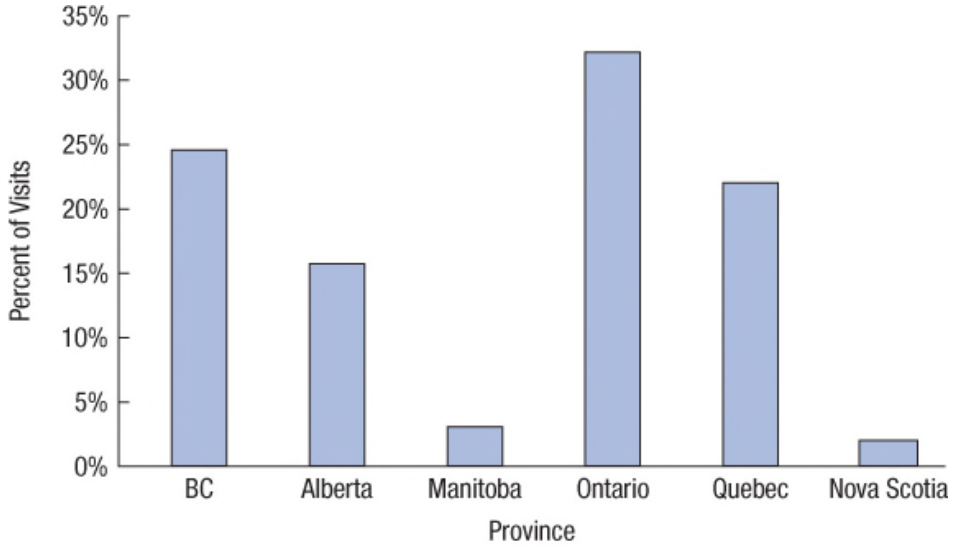
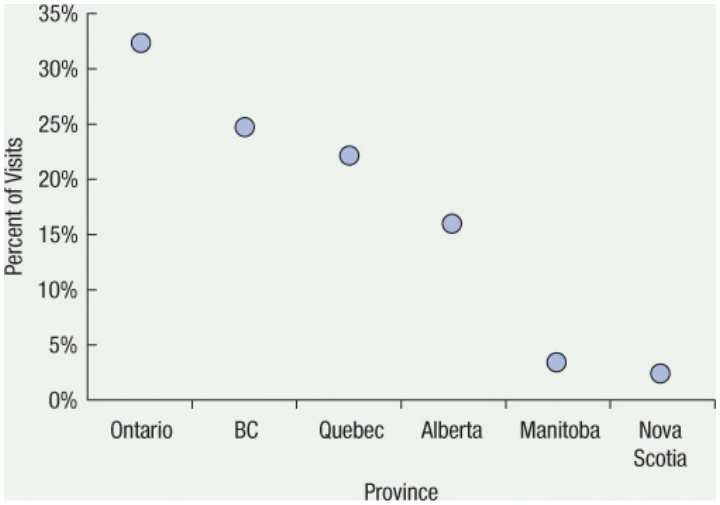


Figure 2.4 The relative frequency bar chart looks the same as the bar chart (Figure 2.3) but shows the proportion of visits in each category rather than the counts.

2.3 Displaying a Categorical Variable (4 of 5)

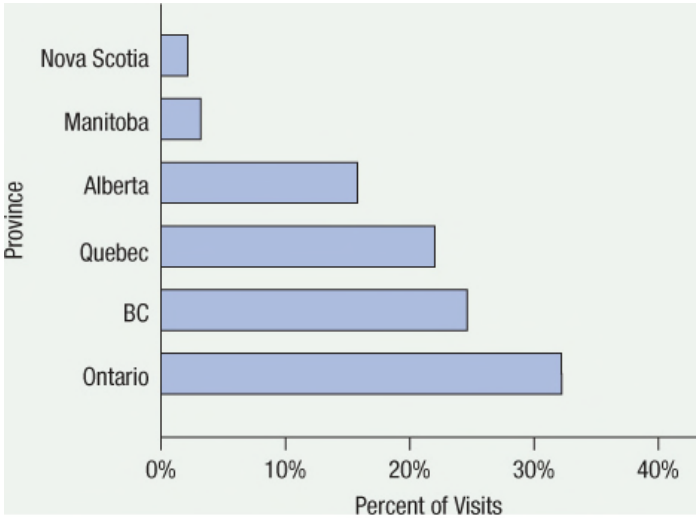
Alternative and Variations on Bar Charts

Dot Plot: the dots replace the bars



Pareto Chart

It is useful to rearrange the order of the categories, and the bars, so that they go from highest to lowest (or vice-versa).



2.3 Displaying a Categorical Variable (5 of 5)

Before making a bar chart or pie chart, …

- the data must satisfy the *Categorical Data Condition*: the data are counts or percentages of individuals in categories.
- be sure the categories don't overlap.

2.4 Exploring Two Categorical Variables (1 of 15)

Example: Data was collected on the strength of consumers' preferences for regional foods in their country. The data is displayed in the frequency table and clarified with a pie chart.

Table 2.3 A combined frequency and relative frequency table for the responses (from all five countries represented: China, France, India, the United Kingdom, and the United States) to the statement "I have a strong preference for regional or traditional products and dishes from where I come from."

Response to *Regional Food Preference* Question	**Counts**	**Relative Frequency**
Agree Completely	2346	30.51%
Agree Somewhat	2217	28.83%
Neither Disagree nor Agree	1738	22.60%
Disagree Somewhat	811	10.55%
Disagree Completely	498	6.48%
Don't Know	80	1.04%
Total	**7690**	**100.00%**

2.4 Exploring Two Categorical Variables (2 of 15)

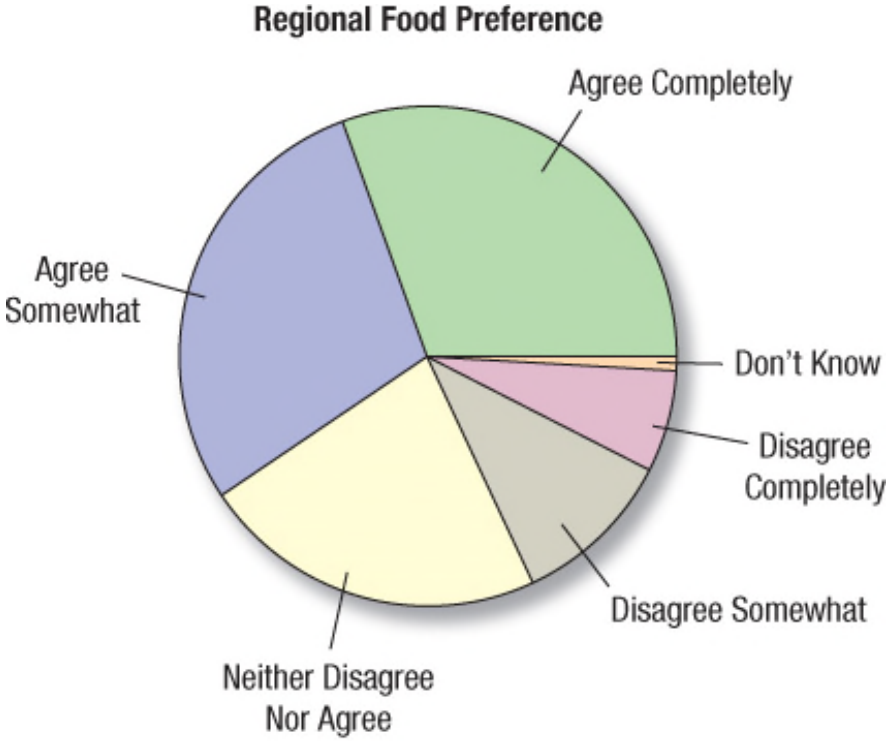


Figure 2.7 It's clear from the pie chart that the majority of respondents identify with their local foods.

2.4 Exploring Two Categorical Variables (3 of 15)

To show how opinions on regional foods varied by countries, we can display the data in a *contingency table* where we have added the countries as a new variable.

2.4 Exploring Two Categorical Variables (4 of 15)

Table 2.4 Contingency table of *Regional Preference* and *Country.* The bottom line "Totals" are the values that were in Table 2.3.

	Regional Preference						
Country	**Agree Completely**	**Agree Somewhat**	**Neither Disagree nor Agree**	**Disagree Somewhat**	**Disagree Somewhat**	**Don't Know**	**Total**
China	518	576	251	117	33	7	**1502**
France	347	475	400	208	94	15	**1539**
India	960	282	129	65	95	4	**1535**
U.K.	214	407	504	229	175	28	**1557**
U.S.	307	477	454	192	101	26	**1557**
Total	**2346**	**2217**	**1738**	**811**	**498**	**80**	**7690**

2.4 Exploring Two Categorical Variables (5 of 15)

The *marginal distribution* of a variable in a contingency table is the total count that occurs when the value of that variable is held constant.

Here the marginal distribution indicated shows that there were 1502 respondents from China.

2.4 Exploring Two Categorical Variables (6 of 15)

Table 2.4 Contingency table of *Regional Preference* and *Country.* The bottom line "Totals" are the values that were in Table 2.3.

	Regional Preference						
Country	**Agree Completely**	**Agree Somewhat**	**Neither Disagree nor Agree**	**Disagree Somewhat**	**Disagree Somewhat**	**Don't Know**	**Total**
China	518	576	251	117	33	7	**1502**
France	347	475	400	208	94	15	**1539**
India	960	282	129	65	95	4	**1535**
U.K.	214	407	504	229	175	28	**1557**
U.S.	307	477	454	192	101	26	**1557**
Total	**2346**	**2217**	**1738**	**811**	**498**	**80**	**7690**

2.4 Exploring Two Categorical Variables (7 of 15)

Each *cell* of a contingency table (any intersection of a row and column of the table) gives the count for a combination of values of the two variables.

The circled cell shows that 4 respondents from India didn't know how they felt about the question asked.

2.4 Exploring Two Categorical Variables (8 of 15)

Table 2.4 Contingency table of *Regional Preference* and *Country.* The bottom line "Totals" are the values that were in Table 2.3.

	Regional Preference						
Country	**Agree Completely**	**Agree Somewhat**	**Neither Disagree nor Agree**	**Disagree Somewhat**	**Disagree Somewhat**	**Don't Know**	**Total**
China	518	576	251	117	33	7	**1502**
France	347	475	400	208	94	15	**1539**
India	960	282	129	65	95	4	**1535**
U.K.	214	407	504	229	175	28	**1557**
U.S.	307	477	454	192	101	26	**1557**
Total	**2346**	**2217**	**1738**	**811**	**498**	**80**	**7690**

2.4 Exploring Two Categorical Variables (9 of 15)

Rather than displaying the data as counts, a table may display the data as a percentage – as a *total percent*, *row percent,* or *column percent*, which show percentages with respect to the total count, row count, or column count, respectively.

We see that 6.74% of all respondents were from China and agreed completely with the question asked.

2.4 Exploring Two Categorical Variables (10 of 15)

Table 2.6 A contingency table of *Regional Preference* and *Country* showing only the total percentages.

Regional Preference—Percentage of Total

Country	Agree Completely	Agree Somewhat	Neither Disagree nor Agree	Disagree Somewhat	Disagree Completely	Don't Know	Total
China	*6.74*	*7.49*	*3.26*	*1.52*	*0.43*	*0.09*	**19.53**
France	*4.51*	*6.18*	*5.20*	*2.70*	*1.22*	*0.20*	**20.01**
India	*12.48*	*3.67*	*1.68*	*0.85*	*1.24*	*0.05*	**19.96**
U.K.	*2.78*	*5.29*	*6.55*	*2.98*	*2.28*	*0.36*	**20.25**
U.S.	*3.99*	*6.20*	*5.90*	*2.50*	*1.31*	*0.34*	**20.25**
Total	**30.51**	**28.83**	**22.60**	**10.55**	**6.48**	**1.04**	**100.00**

2.4 Exploring Two Categorical Variables (11 of 15)

Conditional Distributions

Variables may be restricted to show the distribution for just those cases that satisfy a specified condition. This is called a *conditional distribution*.

Here are the preferences of the respondents from India and the U.K, which allows comparison of these responses.

2.4 Exploring Two Categorical Variables (12 of 15)

Table 2.7 The conditional distribution of *Regional Preference* conditioned on two values of *Country*: India and the United Kingdom. This table shows the row percentages.

	Regional Preference						
Country	**Agree Completely**	**Agree Somewhat**	**Neither Disagree nor Agree**	**Disagree Somewhat**	**Disagree Completely**	**Don't Know**	**Total**
India	960	282	129	65	95	4	**1535**
Row percentage	62.54	18.37	8.40	4.23	6.19	0.26	**100%**
U.K.	214	407	504	229	175	28	**1557**
Row percentage	13.74	26.14	32.37	14.71	11.24	1.80	**100%**

2.4 Exploring Two Categorical Variables (13 of 15)

Conditional Distributions

We may display the results of a conditional distribution as a pie chart or as a bar graph.

The data from the previous table are displayed here as a side-by-side bar chart.

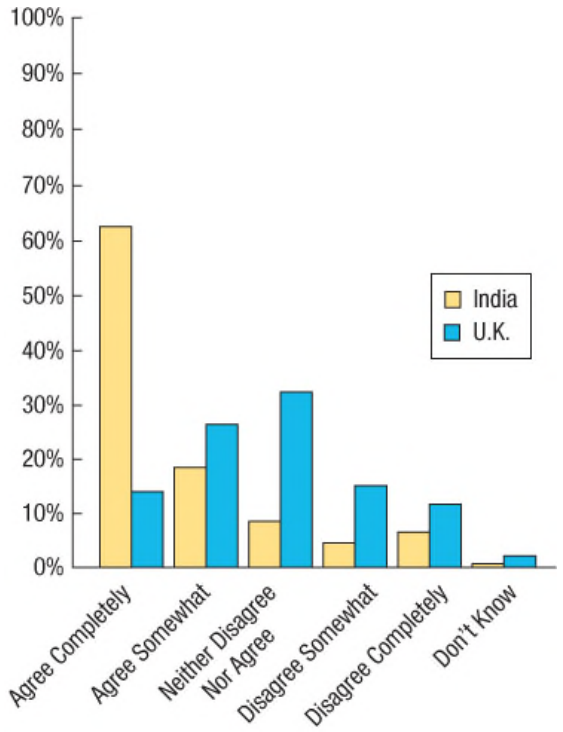


Figure 2.8 Side-by-side bar charts of the conditional distributions of *Regional Food Preference* importance for India and the United Kingdom. The percentage of people who agree is much higher in India than in the United Kingdom.

2.4 Exploring Two Categorical Variables (14 of 15)

Conditional Distributions

Variables can be associated in many ways, so it is typically easier to ask if they are *not* associated.

In a contingency table, when the distribution of one variable is the same for all categories of another variable, we say that the variables are *independent*.

2.4 Exploring Two Categorical Variables (15 of 15)

Segmented Bar Charts

Data can be displayed by dividing up bars rather circles. The result is a *segmented bar chart* where a bar is divided proportionally into segments corresponding to the percentage in each group. The data from the conditional distribution pertaining to India and the U.K. are displayed here as segmented bar charts.

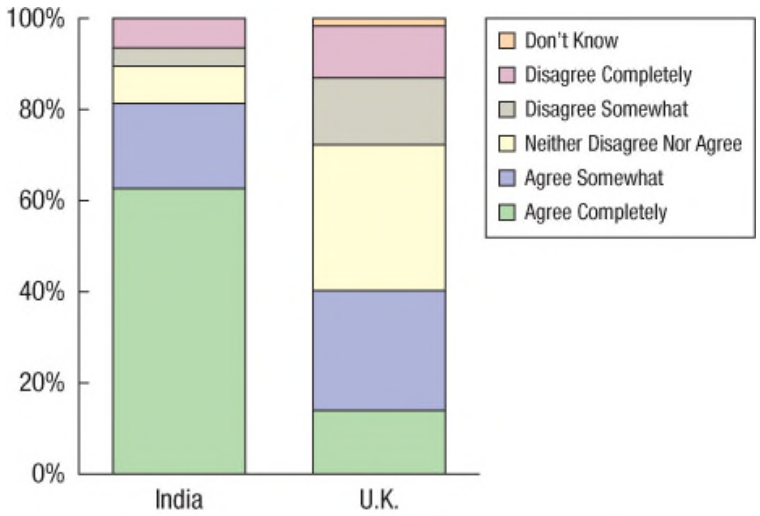
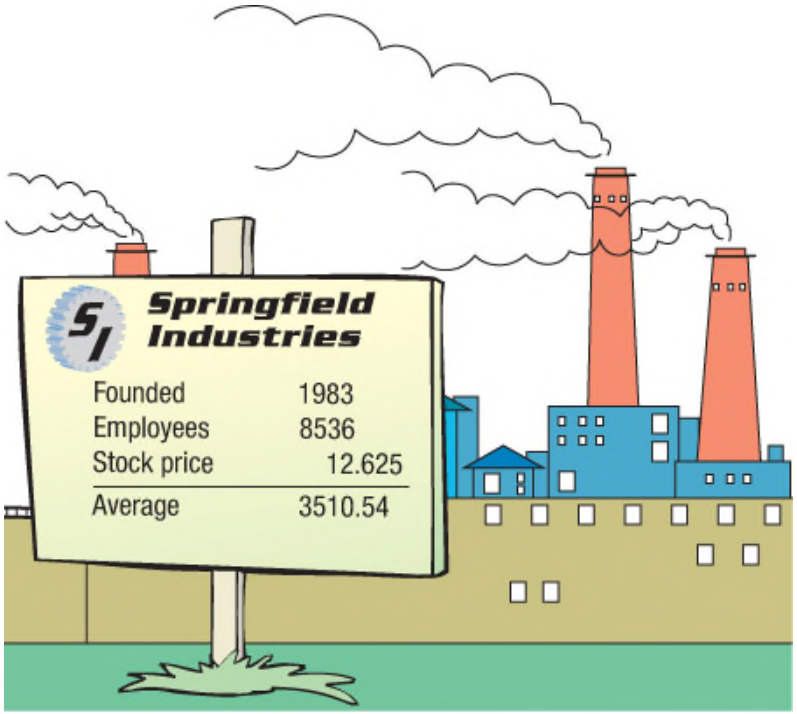


Figure 2.10 Although the totals for India and the United Kingdom are different, the bars are the same height because we have converted the numbers to percentages. Compare this display with the side-by-side bar charts in Figure 2.8 and the side-by-side pie charts of the same data in Figure 2.9.

2.5 Simpson's Paradox (1 of 5)

Simpson's Paradox

Combining percentages across very different values or groups can give confusing results. This is known as *Simpson's Paradox* and occurs because percentages are inappropriately combined.



2.5 Simpson's Paradox (2 of 5)

Example

Suppose there are two sales representatives, Peter and Katrina. Peter argues that he's the better salesperson, since he managed to close 83% of his last 120 prospects compared with Katrina's 78%. Let's look at the data more closely:

2.5 Simpson's Paradox (3 of 5)

Table 2.8 Look at the percentages within each Product category. Who has a better success rate closing sales of paper? Who has the better success rate closing sales of flash drives? Who has the better performance overall?

	Product		
Sales Rep	**Printer Paper**	**USB Flash Drive**	**Overall**
Peter	90 out of 100	10 out of 20	100 out of 120
	90%	50%	83%
Katrina	19 out of 20	75 out of 100	94 out of 120
	95%	75%	78%

2.5 Simpson's Paradox (4 of 5)

Example

Katrina is outperforming Peter in both products, but when combined, Peter has a better overall performance. This is an example of Simpson's paradox, and results from inappropriately combining percentages of different groups.

2.5 Simpson's Paradox (5 of 5)

Table 2.8 Look at the percentages within each Product category. Who has a better success rate closing sales of paper? Who has the better success rate closing sales of flash drives? Who has the better performance overall?

	Product		
Sales Rep	**Printer Paper**	**USB Flash Drive**	**Overall**
Peter	90 out of 100	10 out of 20	100 out of 120
	90%	50%	83%
Katrina	19 out of 20	75 out of 100	94 out of 120
	95%	75%	78%

2.6 How to Make a Table That Has Legs! (1 of 2)

Here are a few suggestions about how to make a good table:

- Use informative headings.
- Arrange rows and columns in the most meaningful way.
- Use white space and lines to organize rows and columns. If you can't use white space, use subtle fill colours.
- Use horizontal text orientation.
- Align numbers to the right, with decimal points aligned. Align dates and text to the left.

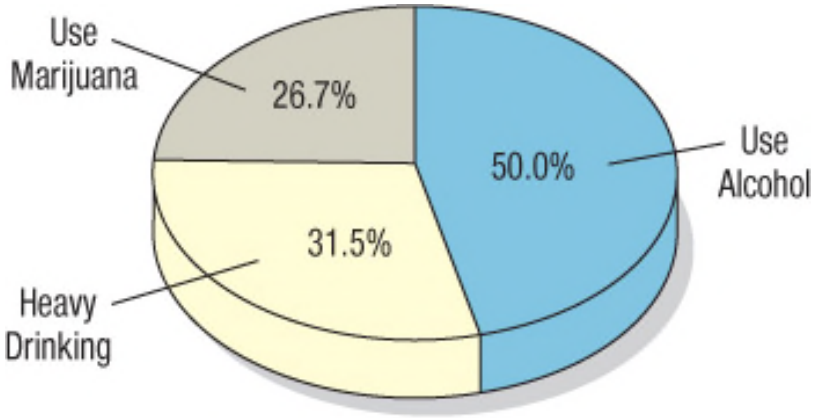
2.6 How to Make a Table That Has Legs! (2 of 2)

Here are a few suggestions about how to make a good table:

- Limit the number of significant digits. Use commas for “thousands” separators. Use percentage signs to the right of every percentage value.
- Use a legible font and keep the same one throughout the table. Use boldface or italics or colour to highlight.
- Make the table as self-contained as possible.

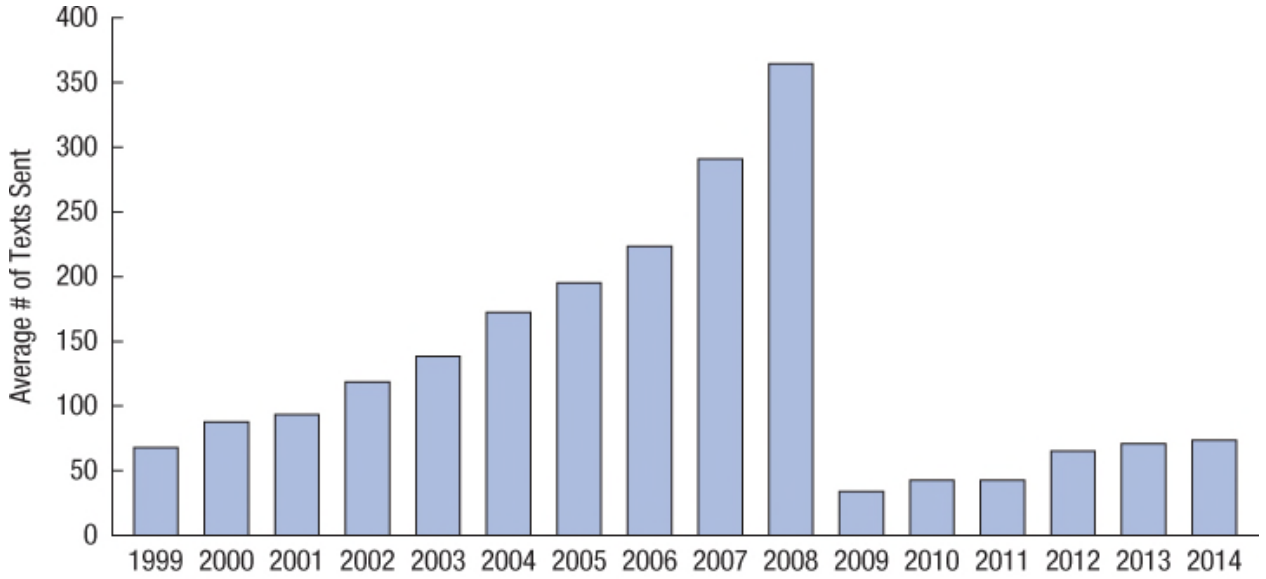
What Can Go Wrong? (1 of 3)

- Don’t violate the area principle.
- Keep it honest.
  - The pie chart below is confusing because the percentages add up to more than 100% and the 50% piece of pie looks smaller than 50%.



What Can Go Wrong? (2 of 3)

- Keep it honest.
  - The denominator used to calculate average is not the same from 2009 on. After 2008 they reported the data for average texts sent per day, not per month.



What Can Go Wrong? (3 of 3)

- Don’t confuse percentages – differences in what a percentage represents needs to be clearly identified.
- Don’t forget to look at the variables separately in contingency tables and through marginal distributions.
- Be sure to use enough individuals in gathering data.
- Don’t overstate your case. You can only conclude what your data suggests. Other studies under other circumstances may find different results.
- Don’t use unfair or inappropriate percentages.

What Have We Learned?

- Categorical data can be summarized by counting the number of cases (or percentages) in each category.
- Distributions can be displayed in a bar chart or a pie chart.
- Categorical variables can be compared using a contingency table to …
  - consider marginal distributions.
  - consider conditional distributions of a variable within each category of the other variable.
  - create bar charts or pie charts.
  - determine if the variables are independent.