Chapter 5: Describing Categorical Variables

Tallies

Two-Way Tables

Collapsing Tables

Types of data and variables (Review!)

Continuous: Can potentially take any value on an interval of numbers.

Categorical: indicates the group or category to which an item belongs.

Ordered Categories are those categories that cannot be displayed in an arbitray order without loss of information. A natural order exists. (page 115)

Examples:

- previous computer experience: none, a little, comfortable with computers.
- I heard this course was: easy, moderate, hard.
- disagree strongly, disagree, neutral, agree, strongly agree.

Nominal categorical data has no natural order.

Examples:

- hair color
- eye color
- type of music you prefer
- gender
- your major

Tallies (page 116)

A **tally** is a count of the number of cases (observations) in a category.

Example:

From a class survey we can produce the following tallies:

Gender	Count
male	229
female	155
Total	384

Hair Color	Count
black	31
brown	234
blond	100
red	9
other	6
Total	380

Frequently we add percentages to tallies such as

Gender	Count	Percent
male	229	60
female	155	40
Total	384	100

Hair Color	Count	Percent
black	31	8
brown	234	62
blond	100	26
red	9	2
other	6	2
Total	380	100

Two-Way Tables (page 119)

A **two-way table** shows the counts (or percentages) in each category formed by intersecting the categories of two variables. (page 119)

A **cell** is defined by the intersection of a category of one variable with the category of another variable.

Example: gender and hair color.

(hair colors: black, brown, blond, red, other)

Minitab Table

ROWS: Gender COLUMNS: Hair

	1	2	3	4	5	ALL
0	22	147	51	4	3	227
1	8	87	49	5	3	152
ALL	30	234	100	9	6	379

Counts

	black	brown	blond	red	other	All
male	22	147	51	4	3	227
female	8	87	49	5	3	152
All	30	234	100	9	6	379

Total Percents

	black	brown	blond	red	other	All
male	6	39	13	1	1	60
female	2	23	13	1	1	40
All	8	62	26	2	2	100

For example, 6 is a rounded
100(22/379) = 5.8047493%

and 39 is a rounded
100(147/379) = 38.7862796%

Marginal distributions, in percentage terms, are in the margins.

Row Percents

	black	brown	blond	red	other	All
male	10	65	22	2	2	100
female	5	57	32	3	2	100
All	8	62	26	2	2	100

For example, 10 is a rounded
100(22/227) = 9.6916299%
and
32 is a rounded
100(49/152) = 32.2368421%

Column Percents

	black	brown	blond	red	other	All
male	73	63	51	44	50	60
female	27	37	49	56	50	40
All	100	100	100	100	100	100

For example, 73 is a rounded
100(22/30)=73.33333333%
(**Conditional distributions**)

Minitab Table

```
ROWS: Gender     COLUMNS: Hair

          1        2        3        4        5      ALL

 0       22      147       51        4        3      227
       9.69    64.76    22.47     1.76     1.32   100.00
      73.33    62.82    51.00    44.44    50.00    59.89
       5.80    38.79    13.46     1.06     0.79    59.89

 1        8       87       49        5        3      152
       5.26    57.24    32.24     3.29     1.97   100.00
      26.67    37.18    49.00    55.56    50.00    40.11
       2.11    22.96    12.93     1.32     0.79    40.11

ALL      30      234      100        9        6      379
       7.92    61.74    26.39     2.37     1.58   100.00
     100.00   100.00   100.00   100.00   100.00   100.00
       7.92    61.74    26.39     2.37     1.58   100.00

 CELL CONTENTS --
                COUNT
                % OF ROW
                % OF COL
                % OF TBL
```


Collapsing Tables (page 121)

Frequently we combine categories (or collapse them) into fewer categories to simplify and gain further insight in to relationships.

Other Data Within Categories

(page 123)

Two categorical variables define the rows and columns but facts about another variable, possibly continuous, are placed in the cells of the table.

Example: Categorize by gender and lecture A or B but show mean age guess for Cryer in cell.

Mean Age Guess by Gender and Lecture

	Lecture		
	A	**B**	**All**
Male	48.000	53.888	50.531
Female	48.656	53.739	51.712
All	48.209	53.816	51.005

Minitab Table

```
MTB > Table 'Gender'  'Lec A';
SUBC>   Means 'Cryer';
SUBC>   Minimums 'Cryer';
SUBC>   Maximums 'Cryer';
SUBC>   StDev 'Cryer'.
Tabulated Statistics


 ROWS: Gender     COLUMNS: Lec A

           0        1      ALL

  0   48.000   53.888   50.531
      23.000   25.000   23.000
      90.000   65.000   90.000
       6.704    4.712    6.600

  1   48.656   53.739   51.712
      22.000   49.000   22.000
      72.000   62.000   72.000
       6.392    2.580    5.133

 ALL  48.209   53.816   51.005
      22.000   25.000   22.000
      90.000   65.000   90.000
       6.596    3.822    6.075

  CELL CONTENTS --
            Cryer:MEAN
                  MINIMUM
                  MAXIMUM
                  STD DEV
```