Statistics Tutorial Lab

- Open now
- Link is on our class website.
- Hours depend on the day of the week:
  - Times fall between 9:30am and 7:00pm
  - Very day-specific, see lab schedule link for specific hours
- Not all stat courses qualify, this one does.

Chapter 3: Visual Displays of Data

- 3.1 Frequency tables
- 3.2 Graphics and plots of data distributions
- 3.3 Graphics in the media
- 3.4 Some cautions about graphics

3.1 Frequency Tables

- A basic **frequency table** has two columns
  - The categories of the data
  - The frequency of each category (or the number of data values in the category).

We surveyed 1000 people about their political affiliation:

Political affiliation	**Frequency**
Democrat	517
Republican	371
Independent	112

EXAMPLE 1 Taste Test

The Rocky Mountain Beverage Company wants feedback on its new product, Coral Cola, and sets up a taste test with 20 people. Each individual is asked to rate the taste of the cola on a 5-point scale:

(bad taste) 1 2 3 4 5 (excellent taste)

The 20 ratings are as follows:

1 3 3 2 3 3 4 3 2 4 2 3 5 3 4 5 3 4 3 1

Construct a frequency table for these data.

EXAMPLE 1 Taste Test

Solution:

The variable of interest is *taste*, and this variable can take on five values: the taste categories 1 through 5. (Note that the data are qualitative and at the ordinal level of measurement.)

Taste scale	Frequency
1	2
2	3
3	9
4	4
5	2
Total	**20**

Tables on National Statistics

- These tables are often very dense with information (tables provided by federal organizations).
- Sometimes we just have to pick them apart to get the data we're looking for.

National Center for Education Statistics

Table 1.1 Number and percentage distribution of students ages 12 through 18 who reported being bullied at school and cyber-bullied anywhere, by type of bullying or cyber-bullying: School year 2010–11

Type of bullying	Number of students	Percent of students
Total bullied or not bullied	24,456,000	100.0
Bullied	6,809,000	27.8
Made fun of, called names, or insulted	4,303,000	17.6
Subject of rumors	4,469,000	18.3
Threatened with harm	1,232,000	5.0
Pushed, shoved, tripped, or spit on	1,923,000	7.9
Tried to make do things they did not want to do	804,000	3.3
Excluded from activities on purpose	1,355,000	5.5
Property destroyed on purpose	689,000	2.8
Not bullied	17,647,000	72.2
Total cyber-bullied or not cyber-bullied	24,411,000	100.0
Cyber-bullied	2,198,000	9.0
Hurtful information on Internet	884,000	3.6
Purposely shared private information[1]	263,000	1.1
Unwanted contact via e-mail	454,000	1.9
Unwanted contact via instant messaging	659,000	2.7
Unwanted contact via text messaging	1,073,000	4.4
Unwanted contact via online gaming	356,000	1.5
Purposeful exclusion from an online community	286,000	1.2
Not cyber-bullied	22,212,000	91.0

[1] This question is new in the 2010–11 survey.

NOTE: For bullying, “at school” includes the school building, school property, school bus, or going to and from school. Bullying and cyber-bullying types sum to more than totals because students could have experienced more than one type of bullying or cyber-bullying. Detail does not sum to total population of students because of rounding and missing data. The population size for all students ages 12–18 is 24,690,000.

SOURCE: U.S. Department of Justice, Bureau of Justice Statistics, School Crime Supplement (SCS) to the National Crime Victimization Survey (NCVS), 2011.

National Center for Education Statistics (Zoomed-in)

Table 1.1 Number and percentage distribution of students ages 12 through 18 wh school and cyber-bullied anywhere, by type of bullying or cyber-bullyin

Type of bullying	Number of students
Total bullied or not bullied	24,456,000
Bullied	6,809,000
Made fun of, called names, or insulted	4,303,000
Subject of rumors	4,469,000
Threatened with harm	1,232,000
Pushed, shoved, tripped, or spit on	1,923,000
Tried to make do things they did not want to do	804,000
Excluded from activities on purpose	1,355,000
Property destroyed on purpose	689,000
Not bullied	17,647,000

Total # students in survey

Two categories

National Center for Education Statistics (Zoomed-in)

Table 1.1 Number and percentage distribution of students ages 12 through 18 wh school and cyber-bullied anywhere, by type of bullying or cyber-bullyin

Type of bullying	Number of students
Total bullied or not bullied	24,456,000
Bullied	6,809,000
Made fun of, called names, or insulted	4,303,000
Subject of rumors	4,469,000
Threatened with harm	1,232,000
Pushed, shoved, tripped, or spit on	1,923,000
Tried to make do things they did not want to do	804,000
Excluded from activities on purpose	1,355,000
Property destroyed on purpose	689,000
Not bullied	17,647,000

Total # students in survey

Frequency of each category

6,809,000 + 17,647,000 = 24,456,000

bullied + # not bullied = total

NOTE: respondents can choose more than one sub-category of bullying, so the total of subcategories will be more than 6,809,000

Relative frequency

- When we include a column expressing the fraction, or proportion, or percentage of observations in each category, we are including the **relative frequencies**.

Survey question:
When is premarital sex wrong?

Category	Count	%
Always	452	452/1895 = 23.85
Almost Always	183	183/1895 = 9.66
Sometimes	429	429/1895 = 22.64
Never	831	831/1895 = 43.85
Total	1895	100.00

Relative frequency

- The sum of the **relative frequencies** must equal 1 if using fractions (or decimals) or 100% if using percentages.

Survey question:
When is premarital sex wrong?

Category	Count	%
Always	452	23.85
Almost Always	183	9.66
Sometimes	429	22.64
Never	831	43.85
Total	1895	100

Relative frequency

- **relative frequency =** $\dfrac{\text{frequency in category}}{\text{total frequency}}$

Survey question:
When is premarital sex wrong?

452/1895 = 0.2385

Category	Count	%
Always	452	23.85
Almost Always	183	9.66
Sometimes	429	22.64
Never	831	43.85
Total	1895	100

Relative frequency

- **relative frequency =** $\dfrac{\text{frequency in category}}{\text{total frequency}}$

Survey question:
Have you been bullied?

6,809,000/24,456,000 = 0.2784

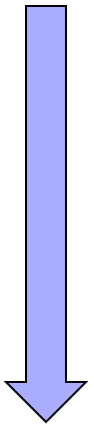
Category	Count	%
Bullied	6,809,000	27.84
Not Bullied	17,647,000	72.16
Total	24,456,000	100

Cumulative frequency

- It's sometimes useful to include information on the accumulation of observations as we move down a table. This is called the **cumulative frequency**.

Survey question:
When is premarital sex wrong?

Category	Frequency	Cumulative Frequency
Always	452	452
Almost Always	183	635
Sometimes	429	1064
Never	831	1895
Total	1895	1895



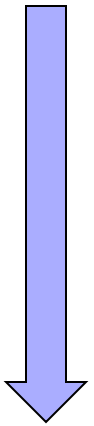
Accumulating

Definition:

- The **cumulative frequency** of any category is the number of data values in that category and all preceding categories

Survey question:
When is premarital sex wrong?

Category	Frequency	Cumulative Frequency
Always	452	452
Almost Always	183	635
Sometimes	429	1064
Never	831	1895
Total	1895	1895



Accumulating

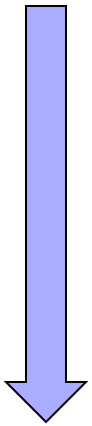
EXAMPLE 1 (return) Taste Test

Taste scale	Frequency	Relative frequency	Cumulative frequency
1	2	2/20 = 0.10	2
2	3	3/20 = 0.15	3 + 2 = 5
3	9	9/20 = 0.45	9 + 3 + 2 = 14
4	4	4/20 = 0.20	4 + 9 + 3 + 2 = 18
5	2	2/20 = 0.10	2 + 4 + 9 + 3 + 2 = 20
Total	**20**	**1**	**20**

Cumulative relative frequency

- We can also consider the **cumulative <u>relative</u> frequency** (sometimes more useful than just cumulative counts):

Category	Frequency	Relative Frequency (as decimal)	Cumulative Relative Frequency (as decimal)
Always	452	0.2385	0.2385
Almost Alw.	183	0.0966	0.2385+0.0966=0.3351
Sometimes	429	0.2264	0.3351+0.2264=0.5615
Never	831	0.4385	0.5615+0.4385=1.0000
Total	1895	1.0000	



EXAMPLE 1 (return) Taste Test

Taste scale	Frequency	Relative frequency	Cumulative frequency	Cumulative Relative Frequency (as decimal)
1	2	2/20 = 0.10	2	0.10
2	3	3/20 = 0.15	3 + 2 = 5	0.25
3	9	9/20 = 0.45	9 + 3 + 2 = 14	0.70
4	4	4/20 = 0.20	4 + 9 + 3 + 2 = 18	0.90
5	2	2/20 = 0.10	2 + 4 + 9 + 3 + 2 = 20	1.00
Total	**20**	**1**	**20**	

Binning Data

- When there are many possible values, or when many of the categories have only 1 response, we may want to **bin** or pool some of the values.
- In such cases, the researcher has some choice in how to **bin** the data.
- Usually, these bins represent a range of numeric values.

Example: Binning Compressive Strengths

- Consider the following set of **80 data points** which are compressive strengths in pounds per square inch of 80 specimens of a new aluminum-lithium alloy undergoing evaluation.

105	97	245	163	207	134	218	199	160	196
221	154	228	131	180	178	157	151	175	201
183	153	174	154	190	76	101	142	149	200
186	174	199	115	193	167	171	163	87	176
121	120	181	160	194	184	165	145	160	150
181	168	158	208	133	135	172	171	237	170
180	167	176	158	156	229	158	148	150	118
143	141	110	133	123	146	169	158	135	149

Example: Binning Compressive Strengths

- The values range from 76 to 245.
- There are 62 unique values.
  - That's way too many bins to be useful.
  - Many of those bins would have 1 observation.
  - We should create bins that will pool some of the observations in useful way.

105	97	245	163	207	134	218	199	160	196
221	154	228	131	180	178	157	151	175	201
183	153	174	154	190	76	101	142	149	200
186	174	199	115	193	167	171	163	87	176
121	120	181	160	194	184	165	145	160	150
181	168	158	208	133	135	172	171	237	170
180	167	176	158	156	229	158	148	150	118
143	141	110	133	123	146	169	158	135	149

Example: Binning Compressive Strengths

- Here is one option for binning the data into 10 non-overlapping bins:

Class Interval	Class midpoint	Frequency f	Relative frequency	Cumulative Relative frequency
61-80	70.5	1	$1/80 = 0.0125$	0.0125
81-100	90.5	2	$2/80 = 0.0250$	0.0375
101-120	110.5	6	$6/80 = 0.0750$	0.1125
121-140	130.5	8	$8/80 = 0.1000$	0.2125
141-160	150.5	23	$23/80 = 0.2875$	0.5000
161-180	170.5	19	$19/80 = 0.2375$	0.7375
181-200	190.5	12	$12/80 = 0.1500$	0.8875
201-220	210.5	4	$4/80 = 0.0500$	0.9375
221-240	230.5	4	$4/80 = 0.0500$	0.9875
241-260	250.5	1	$1/80 = 0.0125$	1.0000

Bin 1 (low values)

Bin 10 (high values)

We observe many values in the mid-value bins.

Example: Compressive Strengths

- Each bin covers a range of 20 values.
- I've provided a '*class representative'* column as the class midpoint of the range of values.
- We can see that the bin from 141-160 is the most common bin for this data set.

Example: Compressive Strengths

- There are an infinite number of ways to choose the bins (or ranges for bins).
  - Too many bins is not useful.
  - Too few bins is not useful.
  - Software will usually (wisely) choose the number of bins for you.

- See worksheet on frequency tables for our STAT:1010 students: 1) Majors and 2) Class in school…