Unit 2 – Data Visualization
Solutions – version 10/2/2022
Art of Stat Users

Question #1

The World Health Organization (WHO) records the annual number of confirmed cases of human Avian Influenza A/(H5N1). Following are a subset of their data:

Year:	2003	2004	2005	2006	2007	2008
# cases:	4	32	43	79	59	26

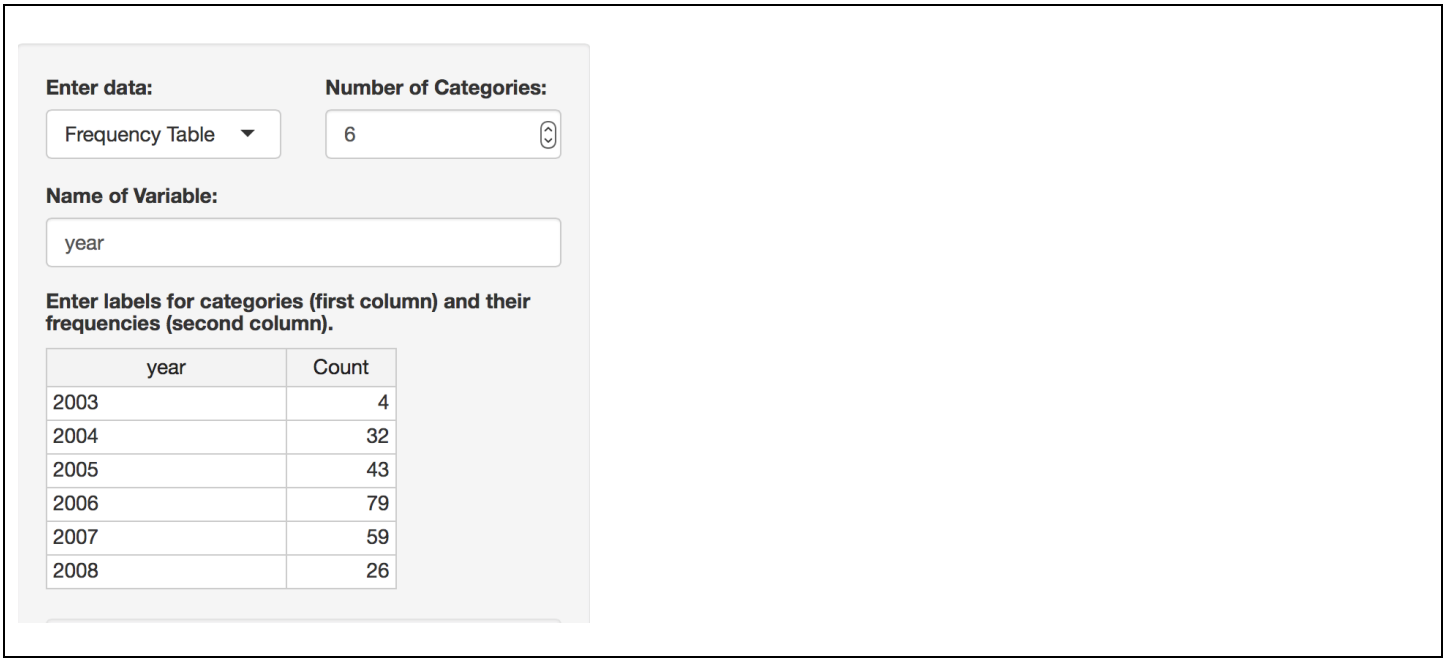
a) By any means you like, construct a **frequency/relative frequency table** summarization of these data.

b) By any means you like, construct a **bar graph** summarization of these data. Include a title and label your axes.

c) State the **facts** of your **bar graph**.

d) In 1-2 sentences, **interpret** the table and bar graph you have produced.

Preliminary – Direct entry of data
www.artofstat.com
> **online webapps** > **explore categorical data** > at tab: **ONE CATEGORICAL VARIABLE**
At left: enter data: **frequency table** > number of categories: **6**

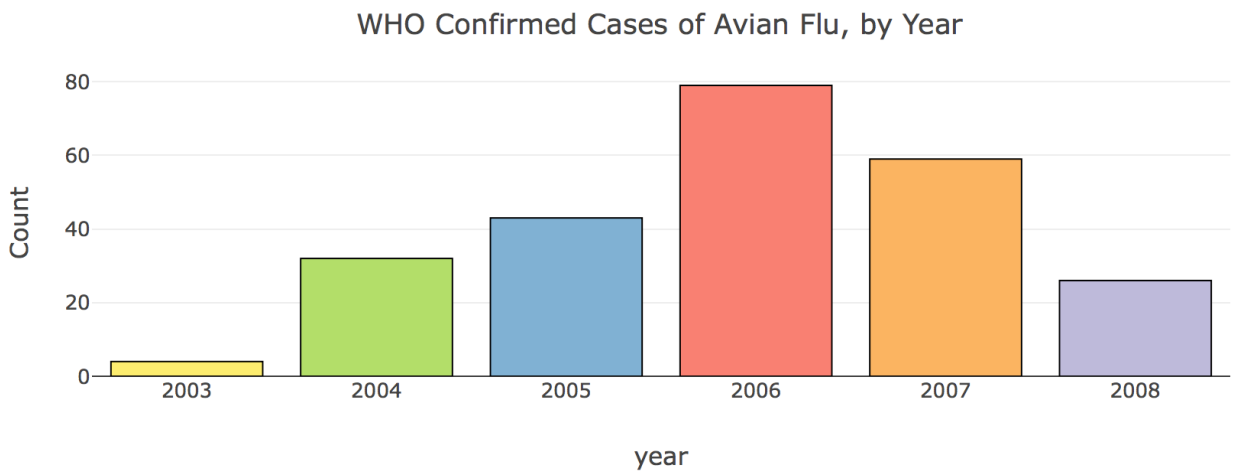


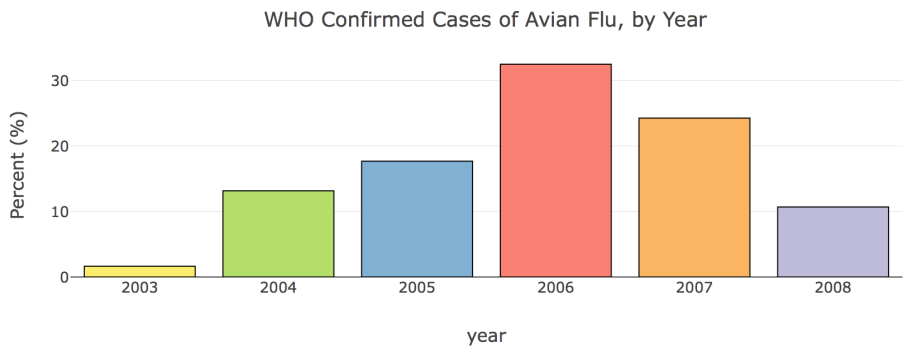
__ **1a) Frequency/relative frequency table** In Art of Stat, choose your options at left

Frequency Table:

year	Frequency	Relative Frequency	Percent (%)
2003	4	0.016	1.65
2004	32	0.132	13.17
2005	43	0.177	17.70
2006	79	0.325	32.51
2007	59	0.243	24.28
2008	26	0.107	10.70
Total:	243	1.000	100.00

__ **1b) Bar Graph** At left, I played with the options to change the colors….





__1c) Facts of Graph

This is a plot of the distribution of 243 cases of human avian influenza A/H5N1 recorded by the World Health Organization (WHO) over the six-year period 2003-2008. The plot shows that the annual number of cases ranged from a minimum of 4 in 2003 to a maximum of 79 in 2006. The annual numbers were: 4, 32, 43, 79, 59, and 26.

Interpretation

The 2003 number of confirmed cases, 4, is substantially lower than the other 5 years' records; the next higher annual number was 26 and was observed in 2008. Inspection of the annual numbers with advancing year shows that increases with year for the period 20013-2006 followed by a decline.

Question #2

A study examining the health risks of smoking measured the cholesterol (mg/dL) levels of people in two independent groups: 1) SMOKERS: those who had smoked for at least 25 years; and 2) NON-SMOKERS: persons of similar ages who had never smoked. The following are the data.

Data.

Good news! The data are available for you in excel format on the course website page, 2. Data Visualization.
Right click to download: https://people.umass.edu/biep540w/datasets/cholesterol_540.xlsx

Smokers			
225	211	209	284
258	216	196	288
250	200	209	280
225	256	243	200
213	246	225	237
232	267	232	216
216	243	200	155
216	271	230	309
183	280	217	305
287	217	246	351
200	280	209	

NON-Smokers		
250	213	300
249	213	310
175	174	328
160	188	321
213	257	292
200	271	227
238	163	263
192	242	249
242	267	243
217	267	218
217	183	228

a) By any means you like, produce a **histogram** of cholesterol for **SMOKERS**
b) By any means you like, produce a **histogram** of cholesterol for **NON-SMOKERS**
c) Produce side-by-side histograms *(Hint: at top click on "several groups")*
d) In 1-2 sentences**, state the facts** of these histograms.
e) In 1-2 sentences, provide an **interpretation** of the comparison of these histograms

Preliminary – Launch excel

- Open your previously downloaded file, ***cholesterol_540.xlsx***
- At bottom, activate the worksheet "data"
- Minimize excel but do not exit.

Preliminary – Launch Art of Stat

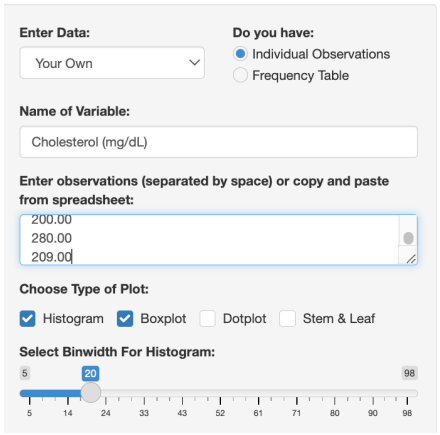
- Launch www.artofstat.com
- Minimize but do not exit.

2a. Histogram for smokers ONLY

- **In excel:** highlight and select values of cholesterol for group=1 (Smokers). Do a EDIT > COPY (control-C)
- **In art of stat**:
- **Online webapps > explore quantitative data >** at top, select tab **SINGLE GROUP**
- > at left enter data: YOUR OWN > at name of variable: CHOLESTEROL (mg/dL)
- > at enter observations: *paste (EDIT>PASTE or control-v) your smokers' data* > at choose type of plot HISTOGRAM

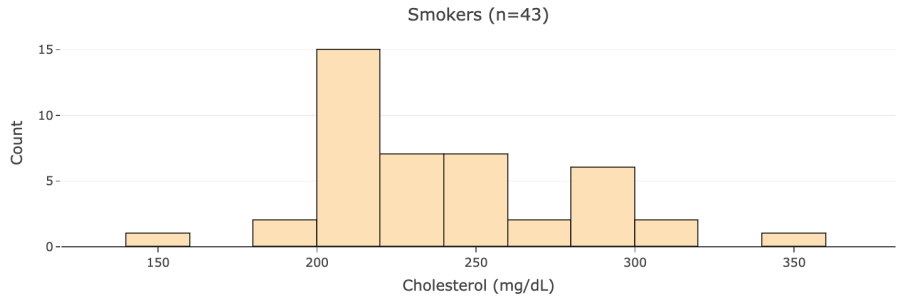
- Scroll down to change various options as you like: bin width, title, etc.

HISTOGRAM for SMOKERS



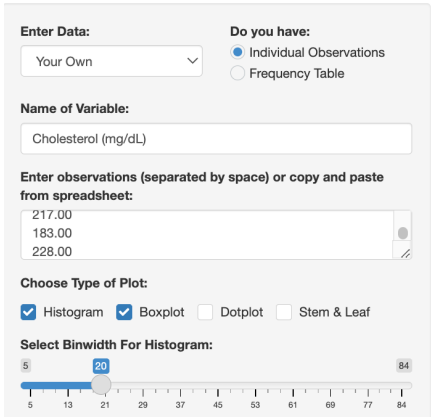
Descriptive Statistics:

Sample Size	Mean	Std. Dev.	Min.	Q1	Median	Q3	Max.	IQR
43	238	38.5	155	212	230	265	351	53.2



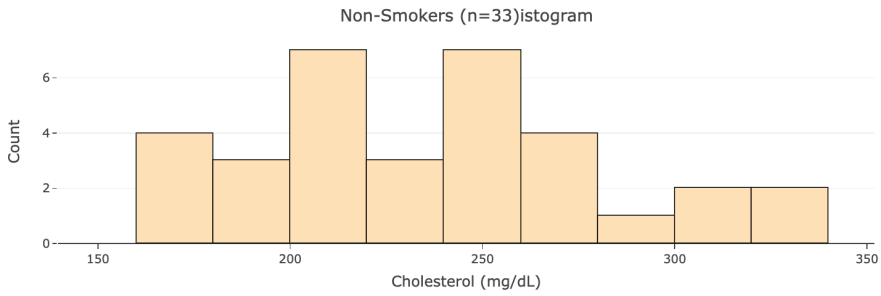
2b. Histogram for nonsmokers ONLY

- In Excel, highlight and select values of cholesterol for group=0 (Non-smokers).
- Proceed just as you did in #2a for the data on smokers.



Descriptive Statistics:

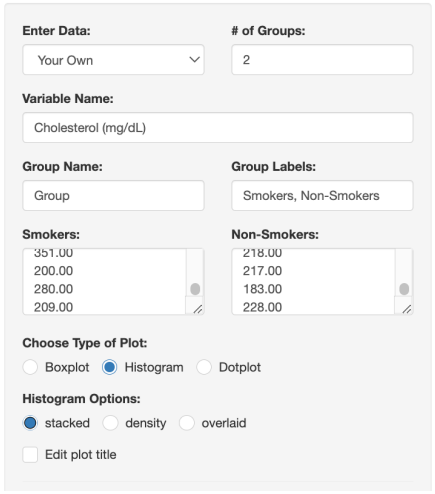
Sample Size	Mean	Std. Dev.	Min.	Q1	Median	Q3	Max.	IQR
33	235	44.4	160	210	238	264	328	54.2



__ 2c. Side-by-side Histograms for smokers and nonsmokers

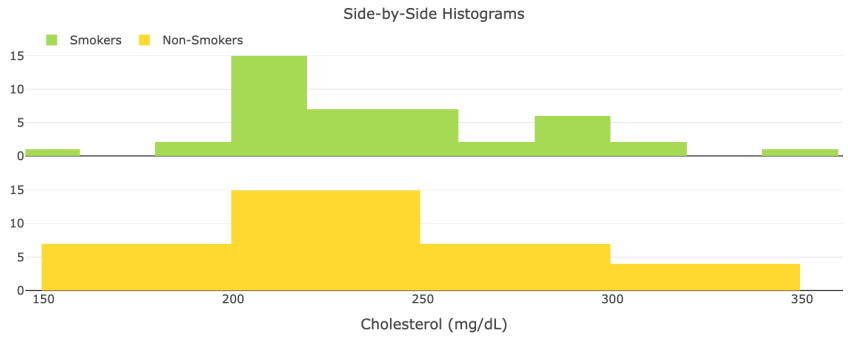
- **In art of stat**:
- **Online webapps > explore quantitative data** > at top, select tab **MULTIPLE GROUPS**
- > at left enter data: YOUR OWN > at number of groups: 2
- > at variable name: CHOLESTEROL (mg/dL) > at group name: GROUP
- > at group labels: SMOKERS, NON-SMOKERS
- > **Retrieve from excel**
  Now paste first your smokers' data into one box, then your non-smokers data into the other box

- > at choose type of plot HISTOGRAM
- > at histogram options: STACKED



Descriptive Statistics:

Group	n	Mean	Std. Dev.	Min.	Q1	Median	Q3	Max.	IQR
Smokers	43	238	38.5	155	212	230	265	351	53.2
Non-Smokers	33	235	44.4	160	210	238	264	328	54.2



__ 2d. Facts of Graph

Cholesterol (mg/dL) was measured in n=43 smokers and n=33 non-smokers. Among smokers, cholesterol ranged from 155 mg/dL to 351 mg/dL, with a mean and standard deviation of 238 mg/dL and 38.5 mg/dL, respectively. Among non-smokers, cholesterol ranged from 160 mg/dL to 328 mg/dL, with a mean and standard deviation of 235 mg/dL and 44.4 mg/dL, respectively. In both groups, the mean and median values were similar, suggesting that the distributions are each symmetric around their central values. Inspection of the histograms confirms this.

__ 2e. Interpretation

These samples of cholesterol (mg/dL) among smokers and non-smokers, on the basis of simple descriptive statistics and side-by-side histograms, suggest that the distributions are similar and thus, not different depending on smoking status.

Question #3

Consider again the cholesterol (mg/dL) data in smokers and non-smokers introduced in question #2.

a) By any means you like, produce a **side-by-side stem and leaf** diagram.
b) By any means you like, complete the following table:

Solution #3

Dear class – Oh too bad! I see that artofstat.com does not provide an option for doing a side-by-side stem and leaf diagram in a single plot. Therefore, I had to do them separately and then do two screen captures which I put side-by-side by brute force. Not convenient! Better would have been to do the side-by-side boxplot requested in question #4, so I show that below too.

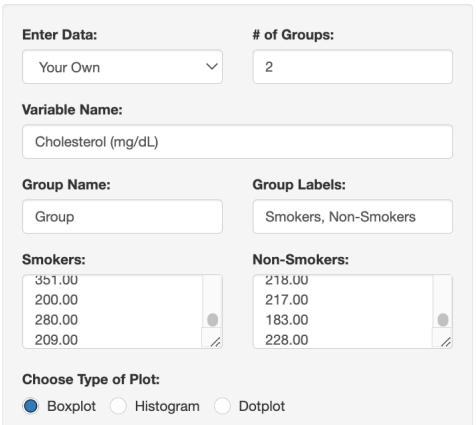
3a. Side-by-side Stem & Leaf Diagram for Smokers and Non-smokers

Smokers (n=43)	Non-Smokers (n=33)
The decimal point is 1 digit(s) 14 | 5 16 | 18 | 36 20 | 000099913666677 22 | 5550227 24 | 3366068 26 | 71 28 | 000478 30 | 59 32 | 34 | 1	The decimal point is 1 digit(s) 16 | 0345 18 | 382 20 | 0333778 22 | 788 24 | 2239907 26 | 3771 28 | 2 30 | 00 32 | 18

Hi again. Comment – it's not as good to do the two graphs separately rather than to do them side-by-side in a single graph. This is because the stems don't quite match, making the comparison a bit of a challenge.

__ 3b. HACK for you Art of Stat users! Two hacks actually.

Hack #1 - The best way to obtain what you need to complete the table is to construct a side-by-side boxplot.
Hack #2 – Then, before exiting art of stat position your cursor over the box plot and hover around. Art of Stat rewards you with the values of key statistics, including most that you need to fill in the table.



Descriptive Statistics:

Group	n	Mean	Std. Dev.	Min.	Q1	Median	Q3	Max.	IQR
Smokers	43	238	38.5	155	212	230	265	351	53.2
Non-Smokers	33	235	44.4	160	210	238	264	328	54.2

	Smokers	NON-Smokers
Number in group, n =	43	33
P_{25} = Q1 = Lower Quartile =	**211.5**	**209.75**
P_{50} = Q2 = Median Quartile =	230	238
P_{75} = Q3 = Upper Quartile =	**264.75**	**264**
Interquartile Range (IQR) =	**53.25**	**54.25**
1.5* IQR=	**79.875**	**81.375**
Value of Lower Fence = P25 – 1.5 IQR	**131.625**	**128.375**
Lower Whisker = smallest observation that is greater than lower fence	**155**	**160**
Value of Upper Fence = P75 + 1.5 IQR	**344.625**	**345.375**
Upper Whisker = largest observation that is smaller than upper fence	**309**	**328**
Outliers (if any) below lower fence (LIST) =	*none*	*none*
Outliers (if any) above upper fence (LIST) =	351	*none*

Dear Class – Some notes:
(1) I hovered over the box plot to get more exact figures
(2) Art of Stat is showing whiskers, but is calling them "fences"
(3) Sometimes, the calculation of the lower (or upper fence) yields the actual minimum (or maximum), as occurred here.
(4) I calculated the 1.5*IQR by hand

Question #4

Consider again the cholesterol (mg/dL) data in smokers and non-smokers introduced in question #2. By any means you like, produce a **side-by-side boxplot**

Solution #4

4. Side-by-side Boxplot for Smokers and Non-smokers

Here it is again, for completeness.