COHEN CHAP 2. FREQUENCY TABLES, GRAPHS, & DISTRIBUTIONS

For EDUC/PSY 6600

ALWAYS PLOT YOUR DATA FIRST ! ! !

ALWAYS PLOT YOUR DATA 1st!

Graphical method ↔ Level of measurement

Label all axes, include figure caption!!!

Simplicity and clarity

- Avoid of 'chartjunk'
- Unless there are 3 or more variables, avoid 3D figures
- Black & white, grayscale/pattern fine for most simple figures

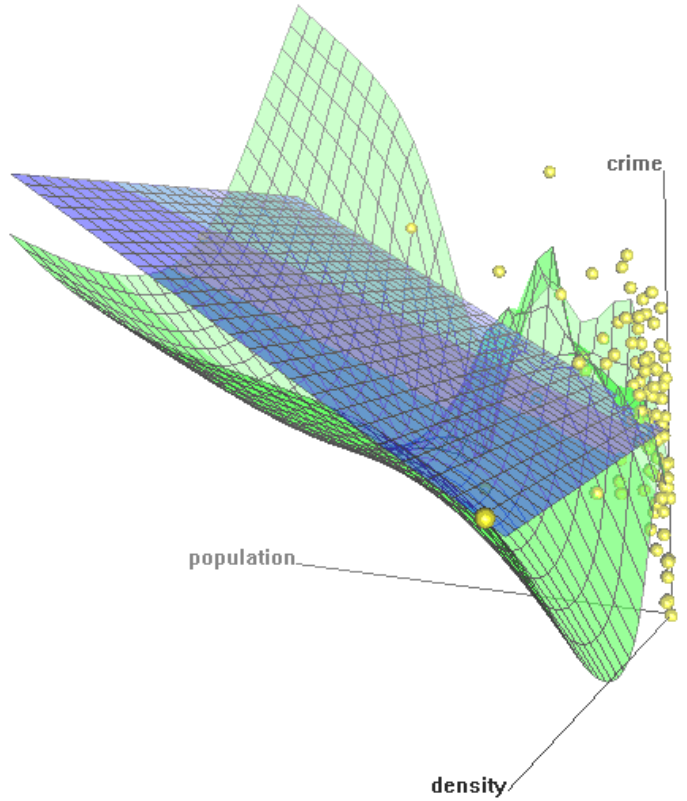


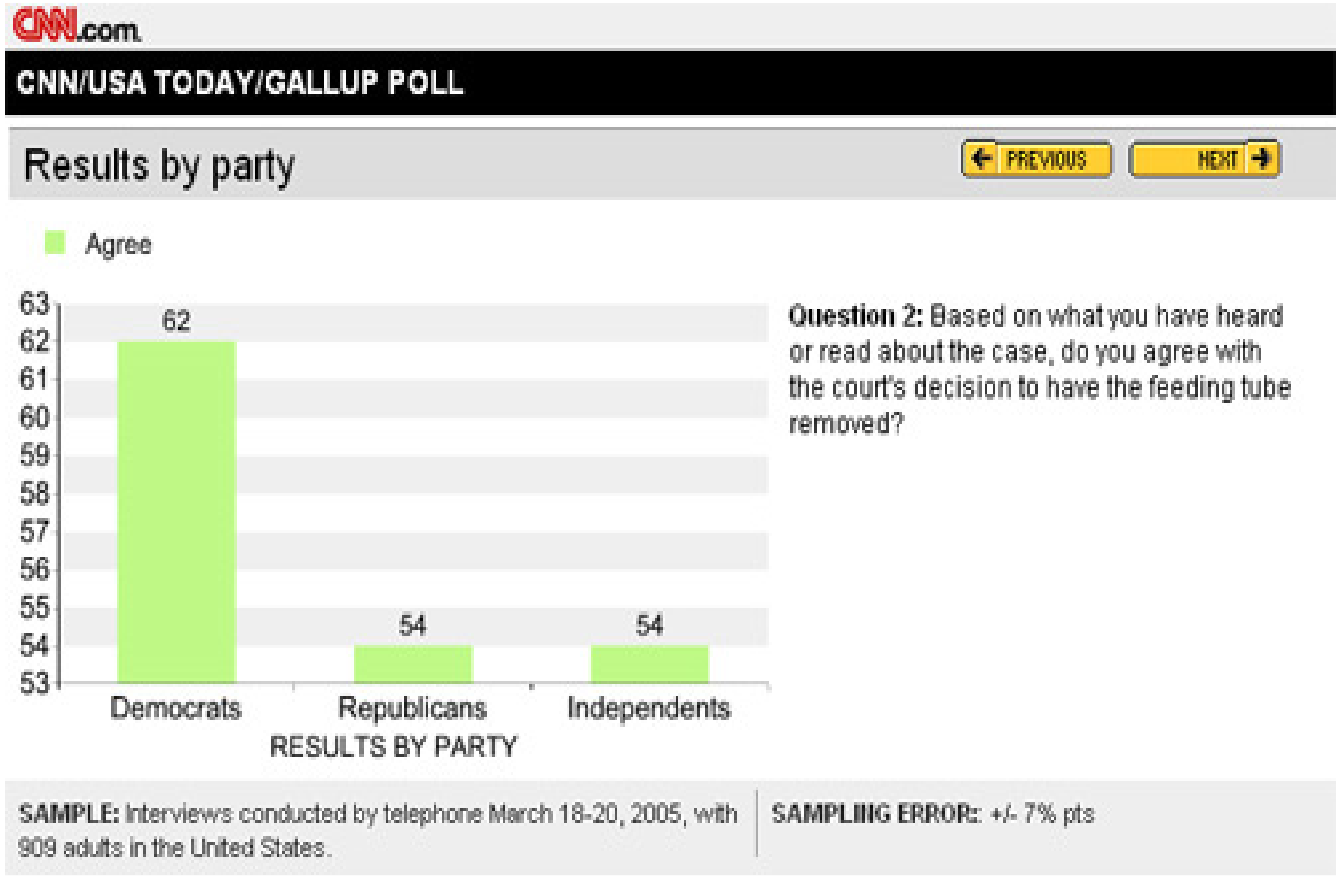
Figure 1. 3-D scatterplot with smoothed surface depicting relationship among violent crime, population, and population density in urban US cities, 2005 (*N* = 110).

DATA VISUALIZATION

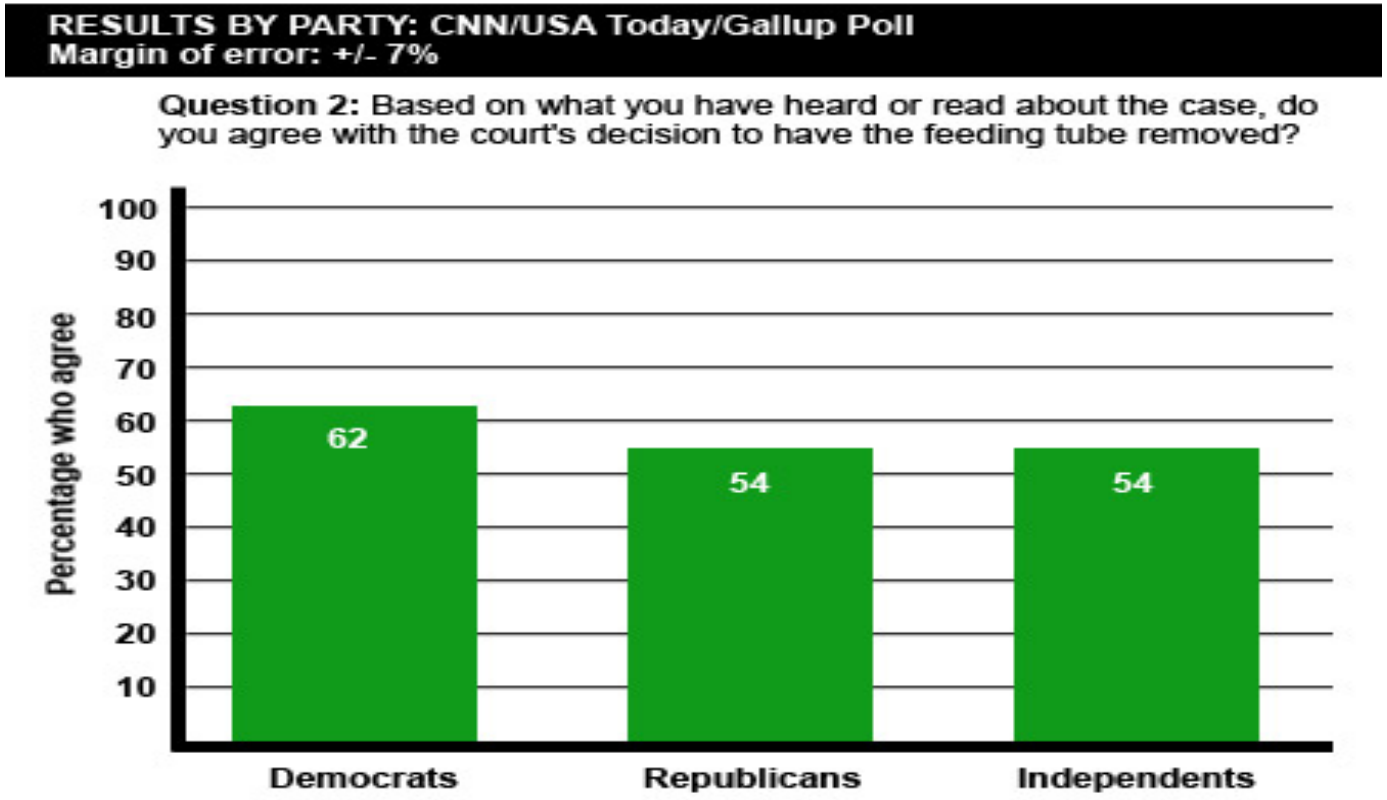
Most abused area of quantitative science: Making misleading charts and figures

"If you can't convince them, confuse them!"

- President Harry S. Truman



REVISED…


RESULTS BY PARTY: CNN/USA Today/Gallup Poll
Margin of error: +/- 7%
Question 2: Based on what you have heard or read about the case, do you agree with the court's decision to have the feeding tube removed?
Percentage who agree
100
90
80
70
60
50
40
30
20
10
62
54
54
Democrats
Republicans
Independents

WHY PLOT YOUR DATA?

- Outliers and impossible values
- Determine correct statistical approach
- Assumptions and diagnostics
- Discover new relationships

FREQUENCY DISTRIBUTIONS

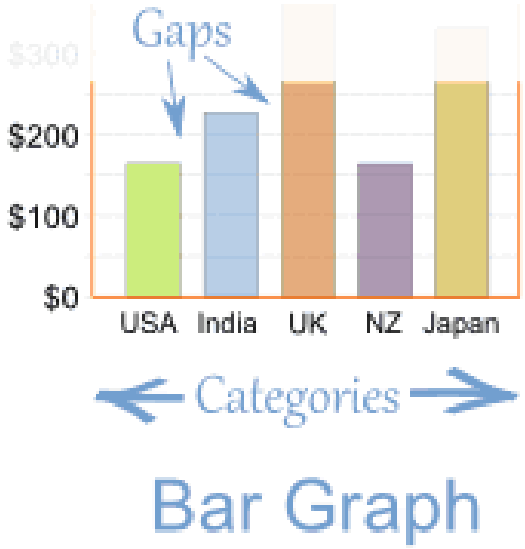
- Counting the number of occurrences of unique events
  - Categorical or continuous
- Can see central tendency (continuous data) or most common value (categorical data)
- Can see range and extremes

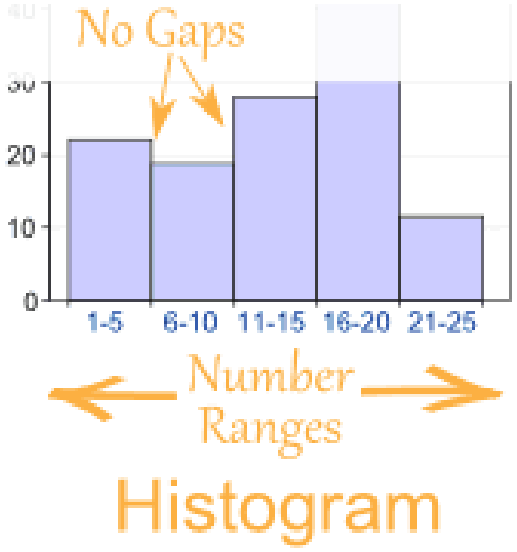
Table 1. Frequency Distribution of Students' Fear of Math

	Frequency	Percent
1	6	50.0
2	3	25.0
4	1	8.3
5	1	8.3
Total	11	91.7
Missing	1	8.3
Total	12	100.0

BAR GRAPH

- Graphical frequency distribution for DISCRETE variables
- Bars
  - Do NOT touch each other
  - Begin and terminate at real limits
  - Centered on the value
  - Height = frequency





HISTOGRAM

- Graphical frequency distribution for continuous variables
- Bars
  - Touch each other
  - Begin and terminate at real limits
  - Centered on interval midpoint
  - Height = frequency
- Interval size or 'bin' determines shape
  - Too narrow or too wide problematic
- Useful for checking distributional assumptions

HISTOGRAM - EXAMPLES

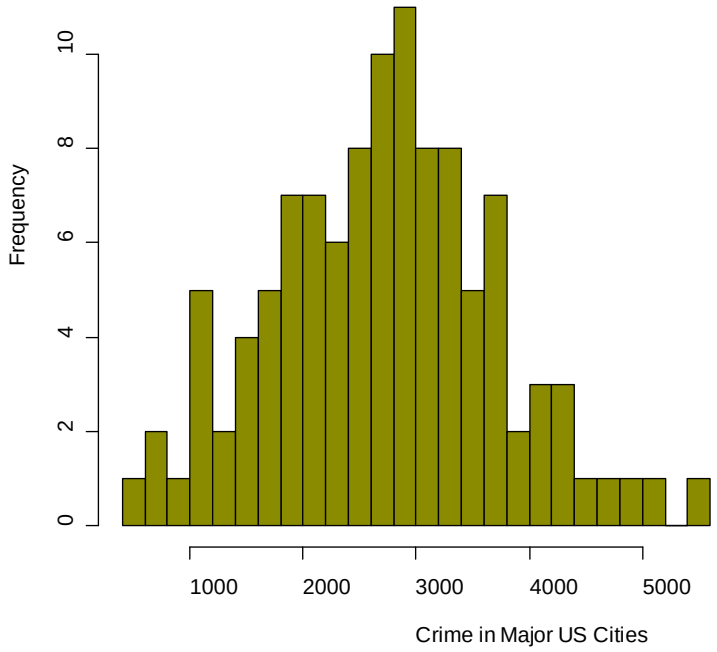


Figure 2. Histogram of violent crime incidents in major U.S. cities, 2005 (*N* = 110).

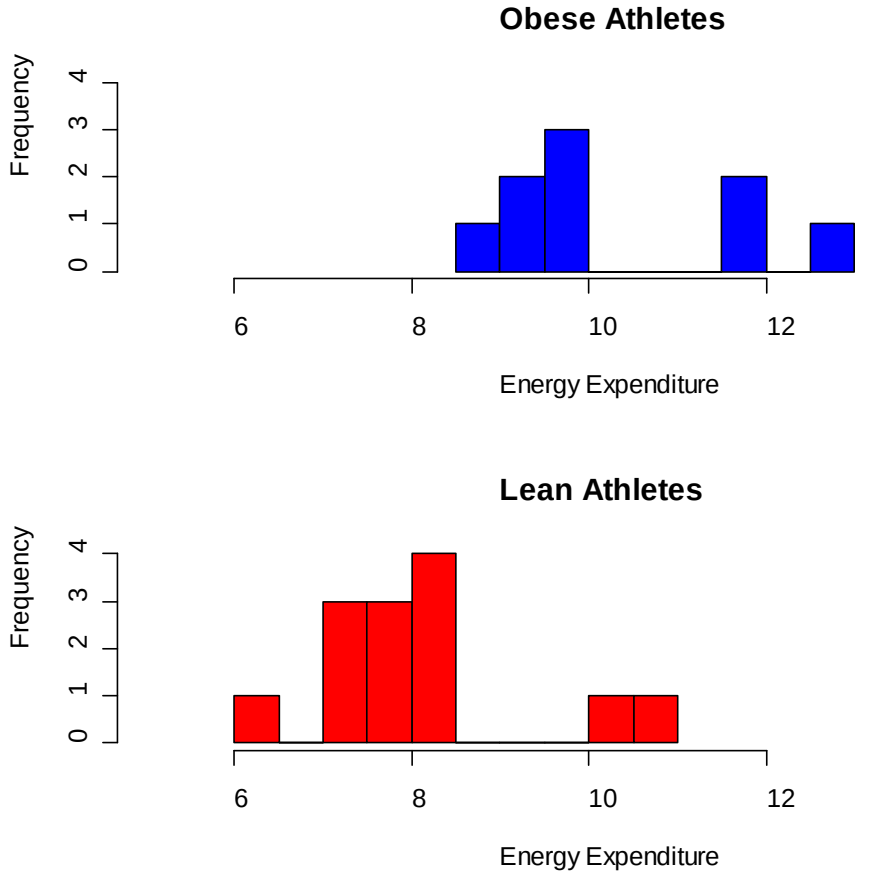


Figure 3. Histogram of energy expenditure by lean (*n* = 11) and obese athletes (*n* = 11).

STEM-AND-LEAF DISPLAY

- Histogram on side
- Continuous data
- Each score represented by stem and leaf
- Best when $N < 100$

```
  1 | 2: represents 12
   leaf unit: 1
           n: 50

     1    -0 | 3
     2     0 | 46
     2     1 | 68
     8     2 | 01467999
     9     3 | 456779999
   (10)    4 | 0012334678
     7     5 | 0022347
     5     6 | 02399
     5     7 | 01558
     1     8 | 0
```

WHAT DO WE MEAN BY **DISTRIBUTION**?

For a **Continuous** variable

- General shape
- Exceptions (outliers)
- Modes (peaks)
- Center & spread (chap 3)
- Histogram
- Cumulative polygon or ogive

For a **Categorical** variable

- Counts = raw number of ___

  Percent or Rate - adjusts for an 'out of' to compare
- Bar chart – should have space between bars, order?
- Pie chart - avoid!

"CANCER" DATASET FOR SPSS DEMO

The data set contains **part** of the data for a study of oral condition of cancer patients conducted at the Mid-Michigan Medical Center.

The **oral conditions** of the patients were measured and recorded at the **initial stage,** at the end of the **second week,** at the end of the **fourth week,** and at the end of the **sixth week.**

The variables **age, initial weight** and **initial cancer stage** of the patients were recorded.

Patients were divided into **two groups** at random: One group received a **placebo** and the other group received **aloe juice** treatment.

Sample size, n = 25 patients with neck cancer.

PREP THE DATASET

ID	TRT	AGE	WEIGHIN	STAGE	TOTAL...	TOTALCW2	TOTALCW4	TOTALCW6
1	0	52	124	2	6	6	6	7
5	0	77	160	1	9	6	10	9
6	0	60	137	4	7	9	17	19
9	0	61	180	1	6	7	9	3
11	0	59	176	2	6	7	16	13
15	0	69	168	1	6	6	6	11
21	0	67	186	1	6	11	11	10
26	0	56	158	3	6	11	15	15
31	0	61	213	1	6	9	6	8
35	0	51	189	1	6	4	8	7
39	0	46	149	4	7	8	11	11
41	0	65	157	1	6	6	9	6
45	0	67	186	1	8	8	9	10
2	0	46	164	2	7	16	9	10
12	1	56	227	4	6	10	11	9
14	1	42	163	1	4	6	8	7
16	1	44	261	2	6	11	11	14
22	1	27	225	1	6	7	6	6
24	1	68	226	4	12	11	12	9
34	1	77	164	2	5	7	13	12
37	1	86	140	1	6	7	7	7
42	1	73	182	0	8	11	16	.
44	1	67	187	1	5	7	7	7
50	1	60	164	2	6	8	16	.
58	1	54	173	4	7	8	10	8

```
VARIABLE LABELS
ID      "Identification number"
TRT     "Treatment Group"
AGE     "Incoming Age"
WEIGHIN "Incoming Weight in pounds"
STAGE   "Stage of Cancer".

VALUE LABELS TRT 0 "control" 1 "aleo treatment".

RECODE
TRT AGE WEIGHIN STAGE TOTALCIN TOTALCW2 TOTALCW4 TOTALCW6
   (SYSMIS = 999).
EXECUTE.

VALUE LABELS
TRT
   0 "control"
   1 "aloe treatment"
   999 "missing"/
AGE WEIGHIN STAGE TOTALCIN TOTALCW2 TOTALCW4 TOTALCW6
   999 "missing".

MISSING VALUES
TRT AGE WEIGHIN STAGE TOTALCIN TOTALCW2 TOTALCW4 TOTALCW6
   (999).
```

NEW VARIABLE

```
COMPUTE weight_oz = 16 * weighin.
COMPUTE total_add = TOTALCIN + TOTALCW2 + TOTALCW4 + TOTALCW6.
COMPUTE total_sum = sum(TOTALCIN, TOTALCW2, TOTALCW4, TOTALCW6).
EXECUTE.
```

ID	TRT	AGE	WEIGHIN	STAGE	TOTAL...	TOTALCW2	TOTALCW4	TOTALCW6	weight_oz	total_add	total_sum
1	0	52	124	2	6	6	6	7	1984.00	25.00	25.00
5	0	77	160	1	9	6	10	9	2560.00	34.00	34.00
6	0	60	137	4	7	9	17	19	2184.00	52.00	52.00
9	0	61	180	1	6	7	9	3	2873.60	25.00	25.00
11	0	59	176	2	6	7	16	13	2812.80	42.00	42.00
15	0	69	168	1	6	6	6	11	2681.60	29.00	29.00
14	1	42	163	1	4	6	8	7	2601.60	25.00	25.00
16	1	44	261	2	6	11	11	14	4182.40	42.00	42.00
22	1	27	225	1	6	7	6	6	3606.40	25.00	25.00
24	1	68	226	4	12	11	12	9	3616.00	44.00	44.00
34	1	77	164	2	5	7	13	12	2624.00	37.00	37.00
37	1	86	140	1	6	7	7	7	2240.00	27.00	27.00
42	1	73	182	0	8	11	16	999	2904.00	.	35.00
44	1	67	187	1	5	7	7	7	2992.00	26.00	26.00
50	1	60	164	2	6	8	16	999	2624.00	.	30.00
58	1	54	173	4	7	8	10	8	2764.80	33.00	33.00

SPSS: THE MULTIPURPOSE "FREQUENCIES" COMMAND

Frequency Table

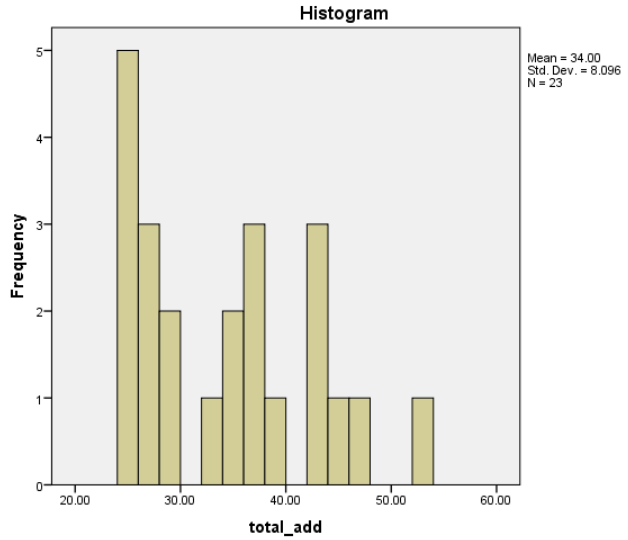
```
FREQUENCIES total_add.
```

total_add

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	25.00	5	20.0	21.7	21.7
	26.00	1	4.0	4.3	26.1
	27.00	2	8.0	8.7	34.8
	29.00	2	8.0	8.7	43.5
	33.00	1	4.0	4.3	47.8
	34.00	1	4.0	4.3	52.2
	35.00	1	4.0	4.3	56.5
	36.00	1	4.0	4.3	60.9
	37.00	2	8.0	8.7	69.6
	38.00	1	4.0	4.3	73.9
	42.00	3	12.0	13.0	87.0
	44.00	1	4.0	4.3	91.3
	47.00	1	4.0	4.3	95.7
	52.00	1	4.0	4.3	100.0
	Total	23	92.0	100.0	
Missing	System	2	8.0		
Total		25	100.0		

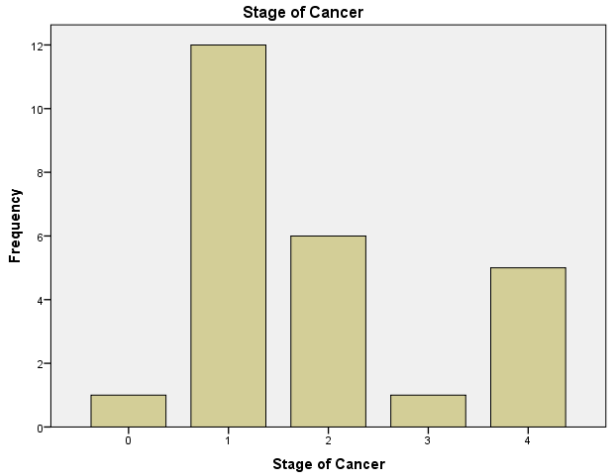
Histogram

```
FREQUENCIES total_add /HISTOGRAM.
```



Barchart

```
FREQUENCIES STAGE
  /BARCHART
  /FORMAT notable.
```



SPSS: ANOTHER "FREQUENCIES" USE

Quartiles

```
FREQUENCIES AGE
  /FORMAT NOTABLE
  /NTILES(4).
```

AGE Incoming Age

N	Valid	25
	Missing	0
Percentiles	25	51.50
	50	60.00
	75	67.50

n'tiles

```
FREQUENCIES AGE
  /FORMAT NOTABLE
  /NTILES(4)
  /NTILES(10).
```

AGE Incoming Age

N	Valid	25
	Missing	0
Percentiles	10	43.20
	20	47.00
	25	51.50
	30	53.60
	40	57.20
	50	60.00
	60	63.40
	70	67.00
	75	67.50
	80	68.80
	90	77.00

Percentiles

```
FREQUENCIES age
  /FORMAT NOTABLE
  /PERCENTILES(5, 95).
```

AGE Incoming Age

N	Valid	25
	Missing	0
Percentiles	5	31.50
	95	83.30

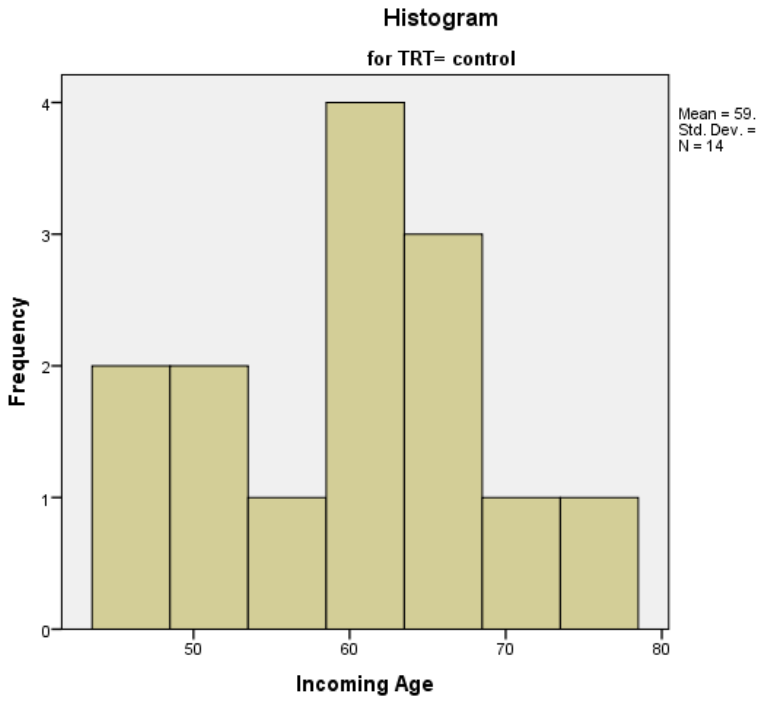
SPSS: COMPARE GROUPS W/"EXAMIME"

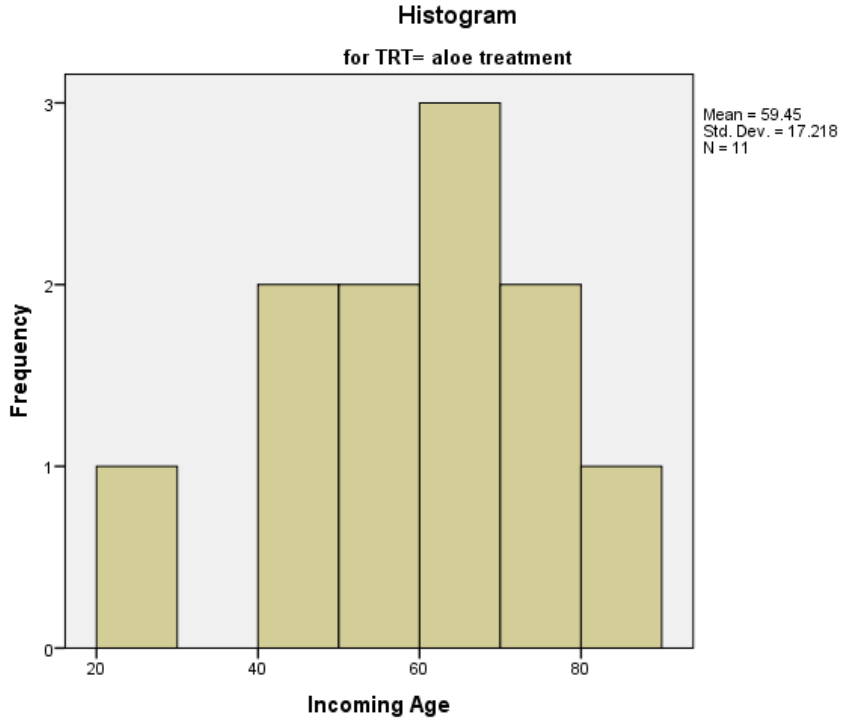
What you get out of the menu window

```
EXAMINE VARIABLES=AGE BY TRT
  /PLOT BOXPLOT STEMLEAF HISTOGRAM
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

Minimal code with defaults

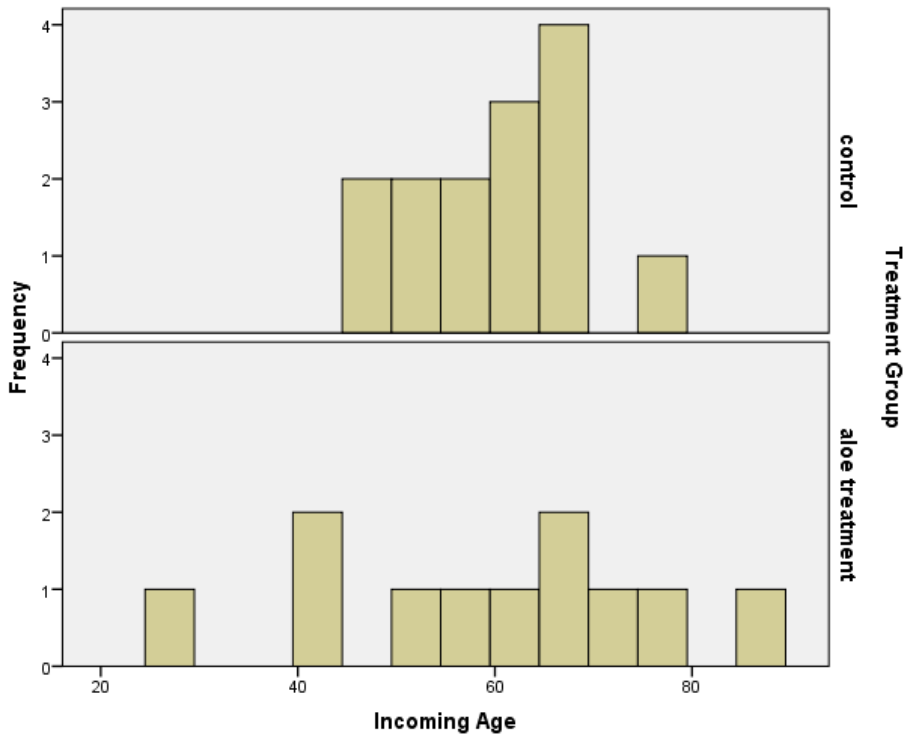
```
EXAMINE AGE BY TRT
  /PLOT HISTOGRAM
  /COMPARE GROUPS.
```





SPSS: COMPARE GROUPS W/"GRAPH" & "PANEL" COMBO

```
GRAPH
  /HISTOGRAM AGE
  /PANEL ROWVAR = TRT.
```




```
GRAPH
  /HISTOGRAM AGE
  /PANEL COLVAR = TRT.
```