Guidelines for Good Graphics

Steve Simon

A good graph must convey statistical information quickly and efficiently. Recent research in perception has identified certain graphic principles that simplify the task of visual discrimination. This talk will discuss these principles as they apply to choices such as the size of a graph, the use of colour, and the coding of subgroups. Examples of a wide variety of graphs will be shown, based on real data, and using SPSS and two other statistical packages.

In this class, you will learn how to:

- avoid excessive and distracting gimmicks.
- replace graphs with effective data tables.
- use alternative to the ambiguous error bars.
- vary the size and shape of a graph to improve its clarity.

Outline

1. Introductory example: Bar or pie chart?

    - Visual process and mental processes.
    - Tables as an alternative to graphs.

2. Graphs with a continuous and a categorical variable.

    - Alternatives to error bars.

3. Graphs with two continuous variables.

    - Rectangular versus square graphs.
    - Colours *versus* symbols *versus* letters.

4. Graphs with 3+ continuous variables.

    - Problems with 3D perspectives.
    - Scatterplot matrix.

Definitions

Continuous variables can have a very large number of possible values, potentially any value in an interval.

Categorical variables can have only a limited number of values, each value corresponding to a specific category or level.

Conclusion

(for those who can't wait until the end).

The Minimum Ink Principle.

- Avoid gimmicks like pseudo 3-D effects or fancy crosshatching. Use the minimum amount of ink to get your point across.

The Small Table Principle.

- A small table is better than a large graph. If you graph contains 20 data points or less, use a table of numbers instead.

The Error of Error Bars Principle.

- Error bars are confusing and ambiguous. Plot all the data if possible, or use a box plot.

The Size and Shape Principle.

- Carefully consider the size and shape of your graph. Rectangular graphs are sometimes better than square graphs. Bigger is not always better.

Which is better, a bar or a pie chart?

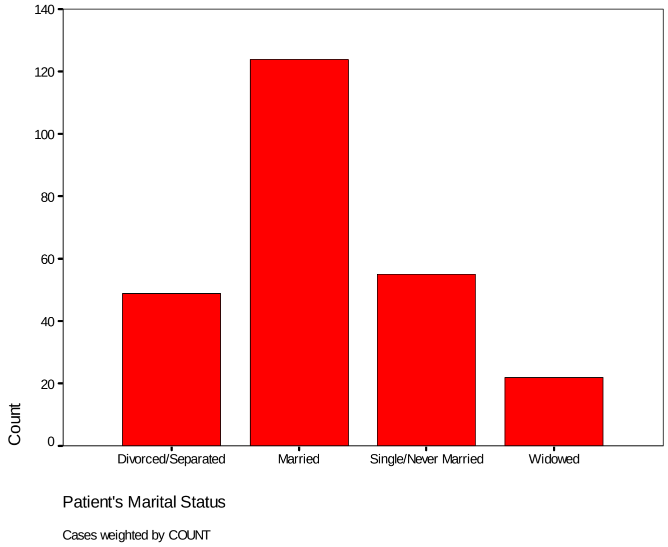


Figure 1. This is a bar chart showing the number of patients in a hospital in each marital status category.   Data is taken from Polit, Denise F. (1996) *Data Analysis and Statistics for Nursing Research*, Appleton & Lange, New York NY, page 25. To create this graph in SPSS, select GRAPHS BAR from the menu.

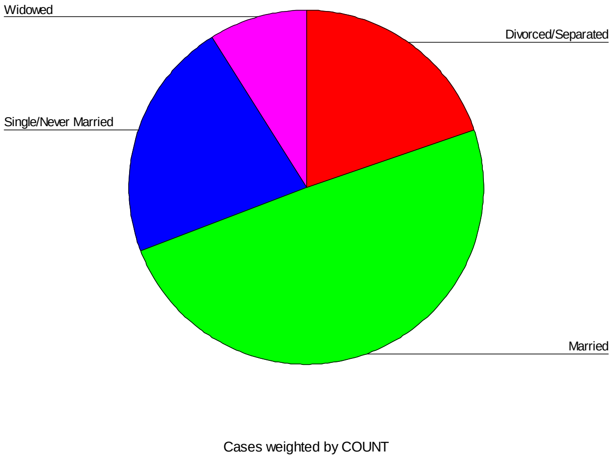


Figure 2. This is a pie chart of the same data. To create this graph in SPSS, select GRAPHS PIE from the menu.

To answer this question, we need to understand visual processing of graphs and mental processes during perception.

Visual processing of graphs.

Visually simple tasks

1. Position
2. Length
3. Angle/slope

Visually demanding tasks

4. Area
5. Volume
6. Density/Saturation/Hue

A good graph relies primarily on simple visual tasks.

Simkin, D., and Hastie, R. (1987) "An Information-Processing Analysis of Graph Perception," Journal of the American Statistical Association, 82, 454-465.

Mental processes during perception

1. Anchoring

- The implicit or explicit development of reference points.

2. Scanning

- Quantifying distance through the use of a mental tape measure.)

3. Projection

- Shifting an object in a horizontal or vertical direction in order to make a comparison.

4. Superimposition

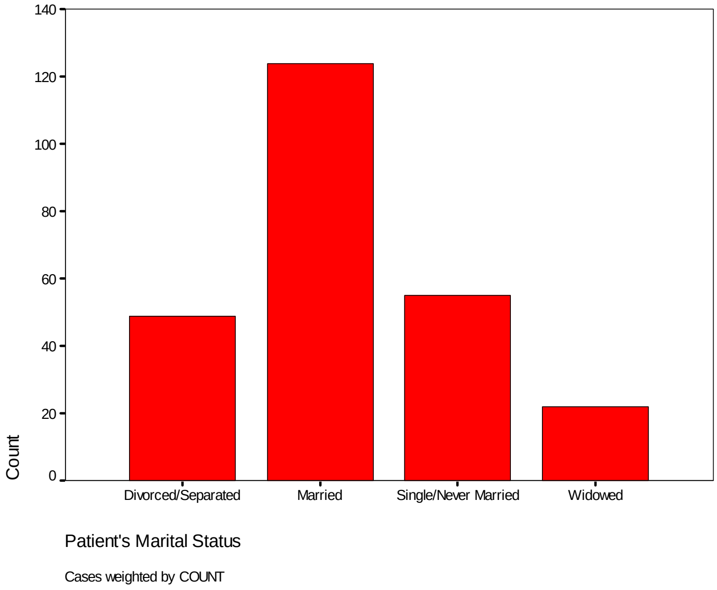
- Shifting in other directions (e.g., diagonal shifts or rotation) in order to make a comparison.

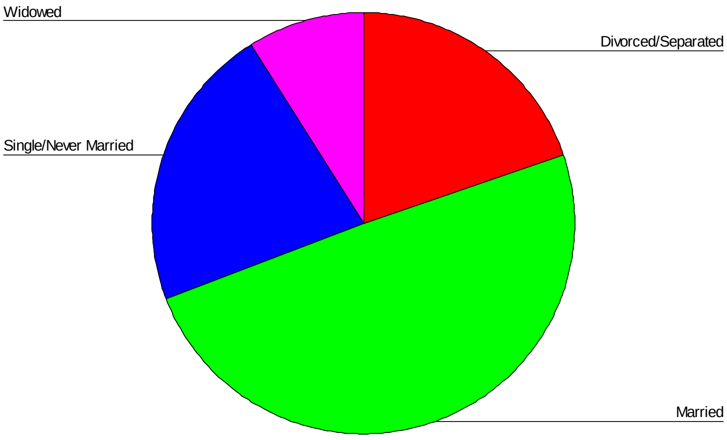
A good graph facilitates these mental processes.

Grid lines on a graph can help with anchoring and projection.

Simkin and Hastie (1987)

Bar charts are better for comparing individual pieces to each other



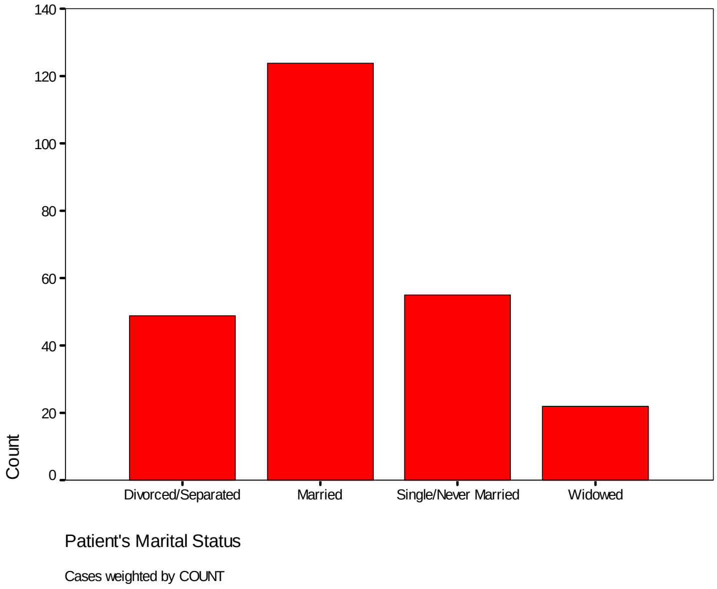


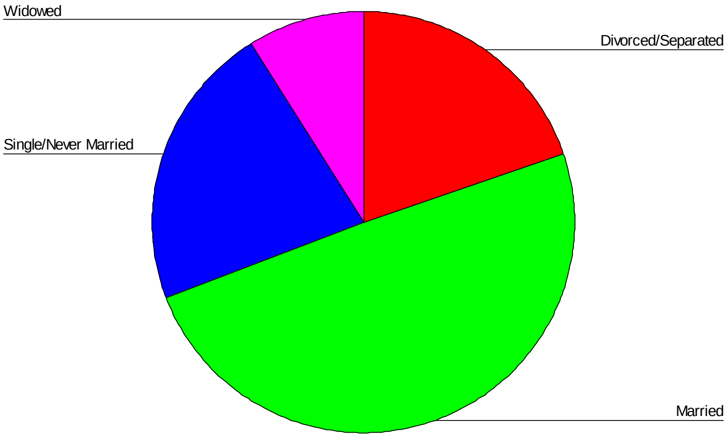
Cases weighted by COUNT

Are there more single patients than divorced patients?

- Bars require projection (shifting bars sideways).
- Pies require superimposition (rotating wedges).
- Projection is easier than superimposition.

Pie charts are better for comparing pieces to the whole.





Cases weighted by COUNT

What percentage of the patients are single?

- Bars have few anchors (it is difficult to visually split a bar into more than two pieces).
- Pies have many anchors (it is easy to visually split a pie chart into four pieces).

Simkin and Hastie (1987)

But, both bar and pie charts are inferior to tables

Be careful, however, in how you display a table!

Here is some raw output from SPSS

```
MARIT_ST  Patient’s Marital Status
                                                Valid     Cum
Value Label         Value  Frequency  Percent  Percent  Percent
Divorced/Separated      1        49     19.6     19.6     19.6
Married                 2       124     49.6     49.6     69.2
Single/Never Married    3        55     22.0     22.0     91.2
Widowed                 4        22      8.8      8.8    100.0
                             -------  -------  -------
                    Total       250    100.0    100.0

Valid cases     250      Missing cases      0
```

Here is a simplified table.

Patient’s Marital Status	
Married	124 (50%)
Single/Never Married	55 (22%)
Divorced/Separated	49 (20%)
Widowed	22 (9%)

Some guidelines for tables

- Show only 2 significant digits.
- Use smaller type than text.
- Sort rows with the largest numbers at the top.
- Put comparisons of interest in vertically.
- Use a table anytime you have 20 or fewer numbers.

Ehrenberg, A.S.C. (1981) “The Problem of Numeracy,” The American Statistician, 35(2), 67-71.

Avoid the use of artificial 3D effects in bar charts…

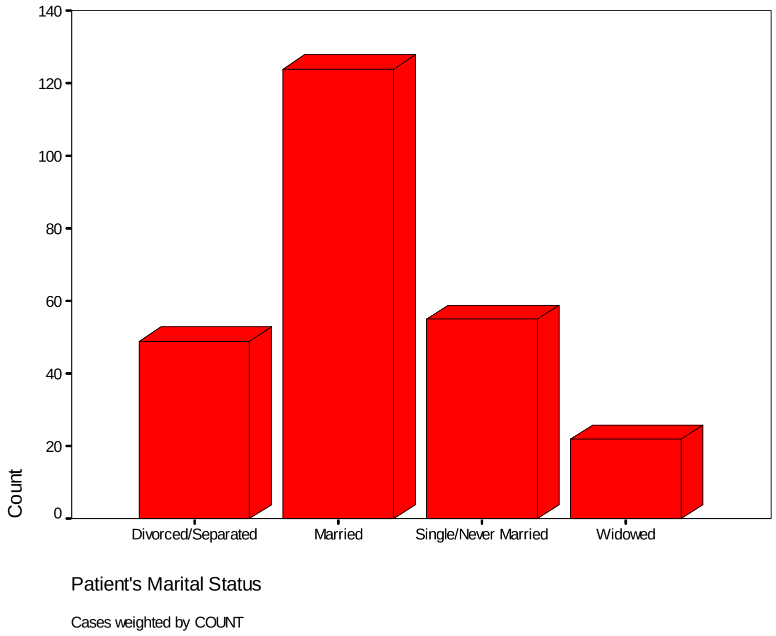


Figure 3. This is a bar chart of the same data using a 3D effect. This style of graph is NOT recommended.

This graph is confusing.

What represents the count, the height at the front of the bar or at the height at the back of the bar?

The best graph is the one which "gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space".

Tufte, E.R. (1983) The Visual Display of Quantitative Information, Graphics Press, Chesire, CT.

...and in pie charts.

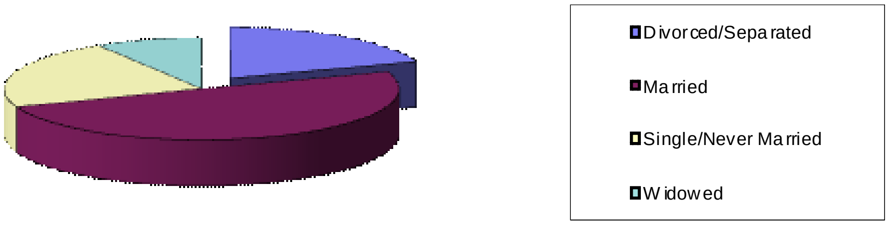


Figure 4. This is a 3D pie chart drawn in Excel using the same data. Thankfully, SPSS does not allow you To create 3D pie charts. This style of graph is NOT recommended.

There are two problems with this pie chart.

- Most people estimate the size of a pie slice or compare two pie slices by examining the angle at the center (a visually simple task) or sometimes the area (a slightly harder task). With this 3D effect, the angles are distorted and so is the area. To estimate or compare, this graph forces the reader to estimate volume, a very demanding task.
- To compare the exploded piece to any of the others, you have to mentally shove the piece back into the pie and then rotate it. There is no justification for making this task so difficult.

Avoid thick lines or crosshatching

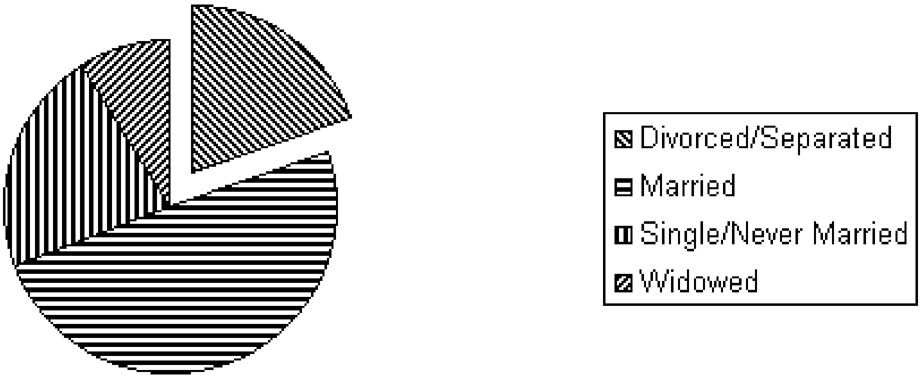


Figure 5. This is a pie chart of the same data, again drawn in Excel, using every garish cross hatching scheme that I could find. This style of graph is NOT recommended.

Thick lines are distracting and confusing. They create a vibratory effect that is unpleasant.

You should especially avoid lines that are roughly equal in thickness to the spaces between the lines.

Use colours and shades of gray instead, or if you have to, use pin stripes.

Place the bars in order of size

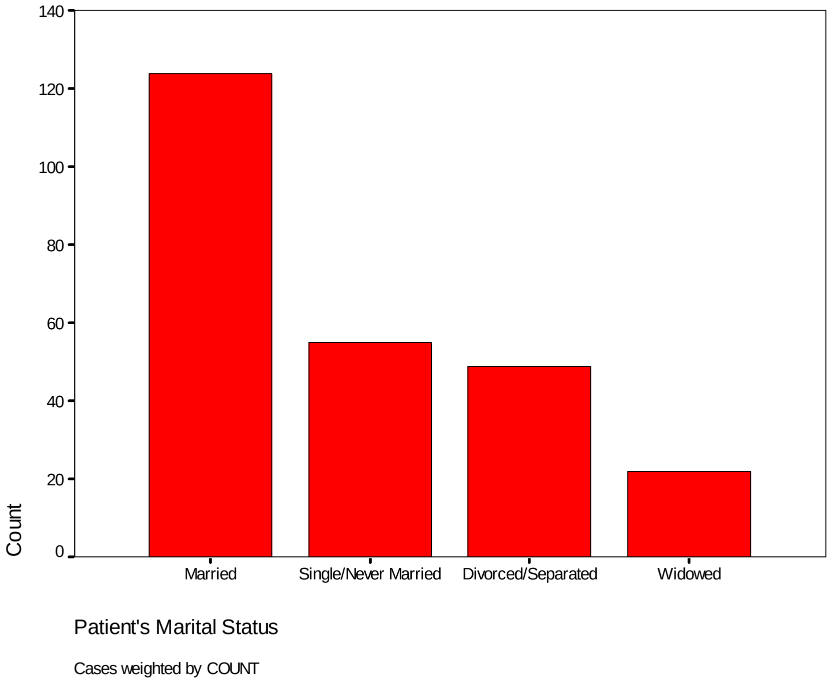


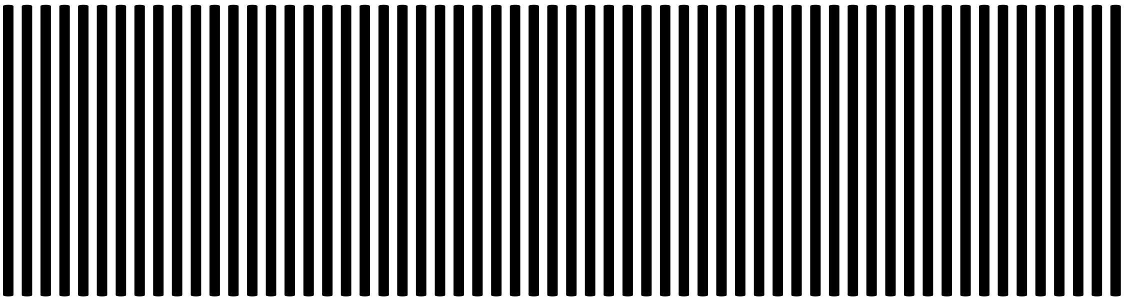
Figure 6. This is a bar chart of the same data where the bars are sorted by their height. This graph was drawn in SPSS, but required the tedious task of removing then re-inserting each bar in the order desired.

Placing the bars in order makes it easier to pick out the most frequent marital status, second most frequent marital status, etc.

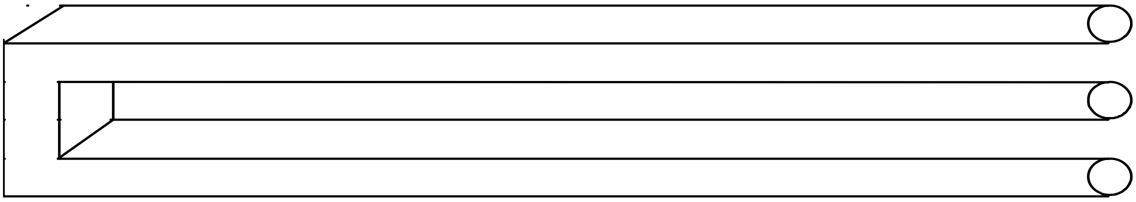
It also makes it easier to compare categories where the number of errors is roughly the same.

Make the gaps thinner than the bars

Evenly spaced regions of light and dark caused by solid bars can cause a vibration effect known as “moire.” Here is an extreme example.



Empty bars which are evenly spaced can causes a different type of optical illusion.



Both problems can be avoided if the gaps are 10-50% of the bar width. Thankfully, this is the default in SPSS.

Turn the bar chart sideways

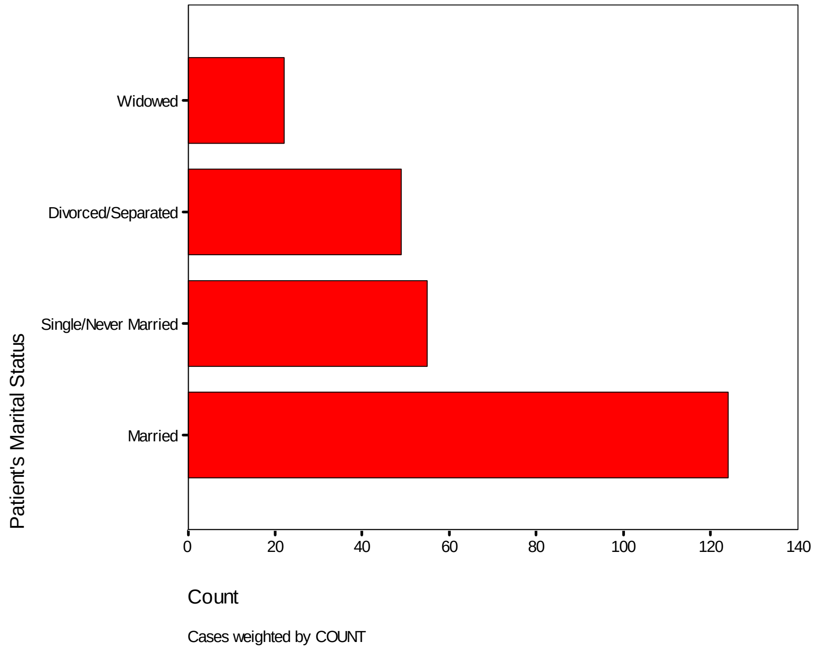


Figure 7. This is a bar chart of the same data where the bars are now horizontal. To make this change in SPSS, click on the EDIT button and select ATTRIBUTES SWAP AXES from the menu.

Labels fit better on a horizontal bar chart.

Also, there is a better sense of continuity when a horizontal label is combined with a horizontal bar.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983) Graphical Methods for Data Analysis, Wadsworth, Inc., Belmont, CA.

Schmid, C.F. (1983) Statistical Graphics: Design Principles and Practices, John Wiley and Sons, Inc., New York, NY.

Summary of Part 1.

The Minimum Ink Principle.

- Avoid gimmicks like pseudo 3-D effects or fancy crosshatching. Use the minimum amount of ink to get your point across.

The Small Table Principle.

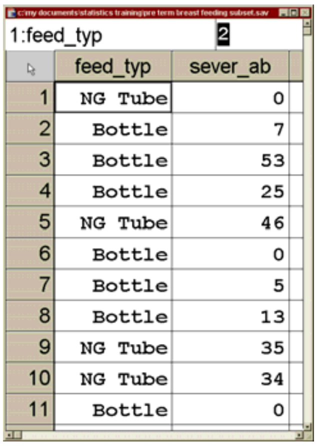
- A small table is better than a large graph. If you graph contains 20 data points or less, use a table of numbers instead.

Other things we learned.

- There are a variety of perceptual tasks involved in viewing a graph. We should try to facilitate these perceptual tasks.
- It is easier to compare the size of individual bars in a bar chart than individual slices in a pie chart.
- It is easier to estimate the percentage of a slice in a pie chart than a bar in a bar chart.

Graphs with a continuous and a categorical variable

Here is are the first 11 of 78 rows of a data set with a categorical variable (feed_typ) and a continuous variable (sever_ab).



c:\my documents\statistics training\pre term breast feeding subset.sav

1:feed_typ 2

	feed_typ	sever_ab
1	NG Tube	0
2	Bottle	7
3	Bottle	53
4	Bottle	25
5	NG Tube	46
6	Bottle	0
7	Bottle	5
8	Bottle	13
9	NG Tube	35
10	NG Tube	34
11	Bottle	0

What's the best way to plot this data?

One way to plot this data is by averaging it.

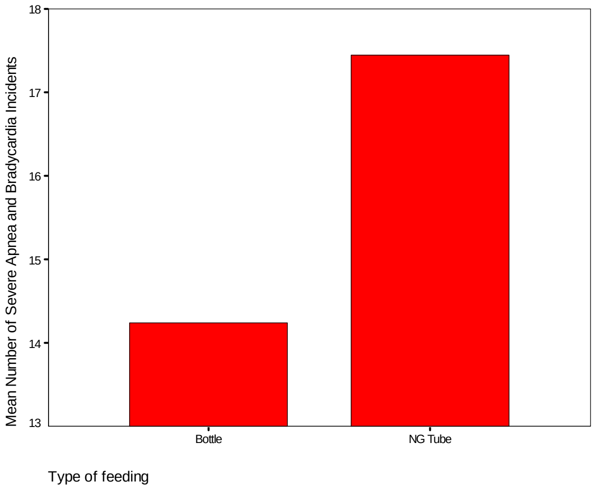


Figure 8. This is a bar chart showing the average number of severe Apnea and Bradycardia incidents for two groups of pre-term infants in a study of breast feeding. To create this graph in SPSS, select the mean in the DEFINE SIMPLE BAR dialog box.

This is the default scaling. Notice how the picture changes when the axis includes zero (see next page).

Here is a plot with alternate scaling.

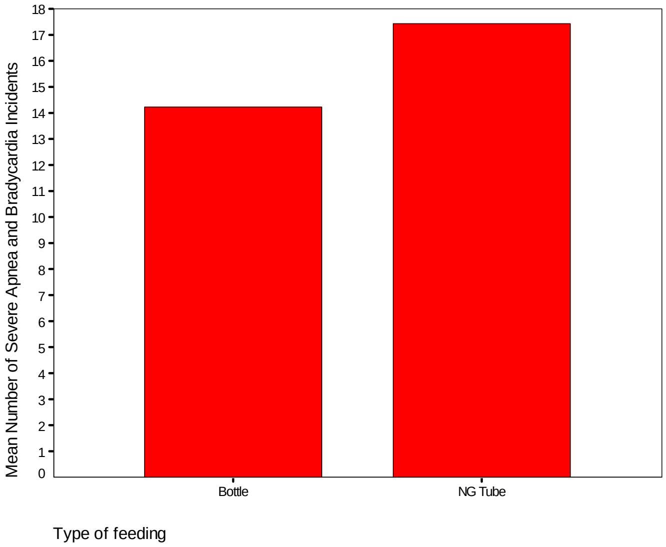


Figure 9. This is a bar chart of the same data, with the axis changed to include zero. To make this change in SPSS, click on the EDIT button and select CHART AXIS from the menu.

This graph makes it is easier to see that the NG Tube group has roughly 20% more incidents, on average. But there is a loss in resolution for this graph.

Suggestion: Draw the graph both ways and pick the graph that works best. Don't blindly accept the default settings.

How can we add information about the spread of the data?

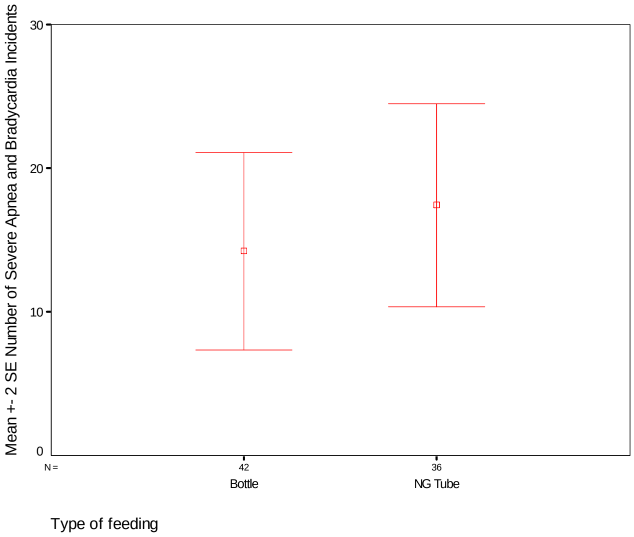


Figure 10. This is a graph of the same data using error bars. SPSS does not allow error bars on a bar chart, but this graph has a similar intent. To create this graph in SPSS, select GRAPHS ERROR BARS from the menu.

One solution is to use error bars. But they are problematic

Problems with error bars.

I recommend against the use of error bars.

- Sometimes error bars can represent a single standard deviation, or a single standard error, or a confidence interval, or a range. Often this is not documented and when it is, the documentation is often overlooked.
- Many people misinterpret an overlap of error bars as implying lack of statistical significance.
- Error bars fail to account for multiple comparisons.
- A two number summary (mean and standard deviation) fails to show information about skewness and outliers.

A good alternative is a five number summary.

Five number summary

- Maximum value
- 75th percentile
- 50th percentile (median)
- 25th percentile
- Minimum value

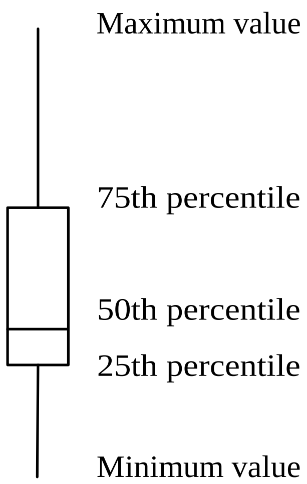
Definition: The p^{th} percentile is a value so that roughly *p* percent of the data is smaller and (100–*p*) percent of the data is larger.

A five number summary shows:

- The middle of the data (50th percentile)
- The spread of the data (75th – 25th percentiles)
- Extremes of the data (minimum and maximum)
- Skewness of the data

A five number summary also splits the data into four regions, each of which contains 25% of the data.

A box plot is a graphic display of a five number summary.



Note: Values more than 1.5 box lengths away from either end of the box are considered outliers. Outliers are represented by individual points.

Box plots are an alternative to error bars.

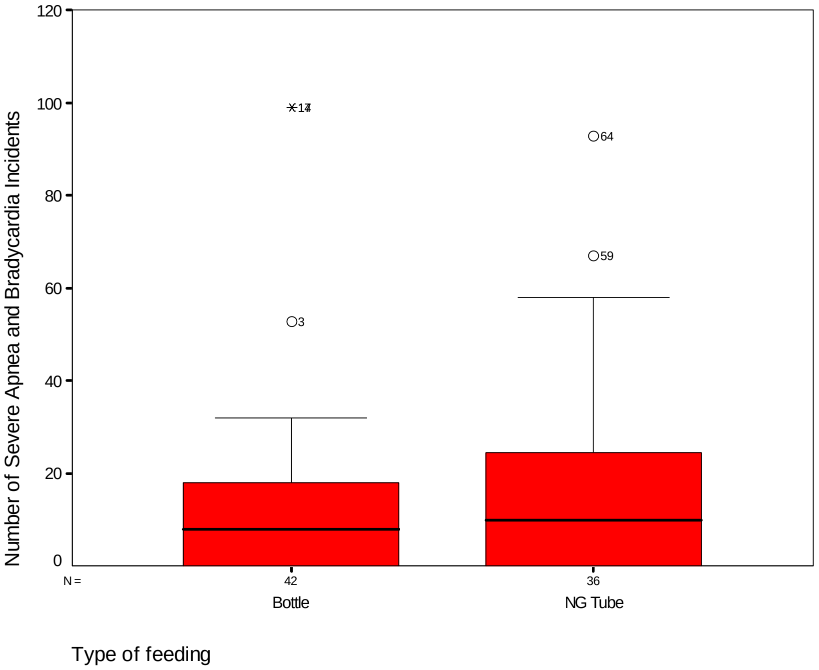


Figure 11. This is a box plot of the same data. To create this graph in SPSS, select GRAPS BOXPLOT from the menu.

In this box plot, there are no lines at the bottom because the minimum values and the 25th percentiles are identical. Notice also that there are several outliers in each group.

AnoPlotting each data point is another alternative to error bars

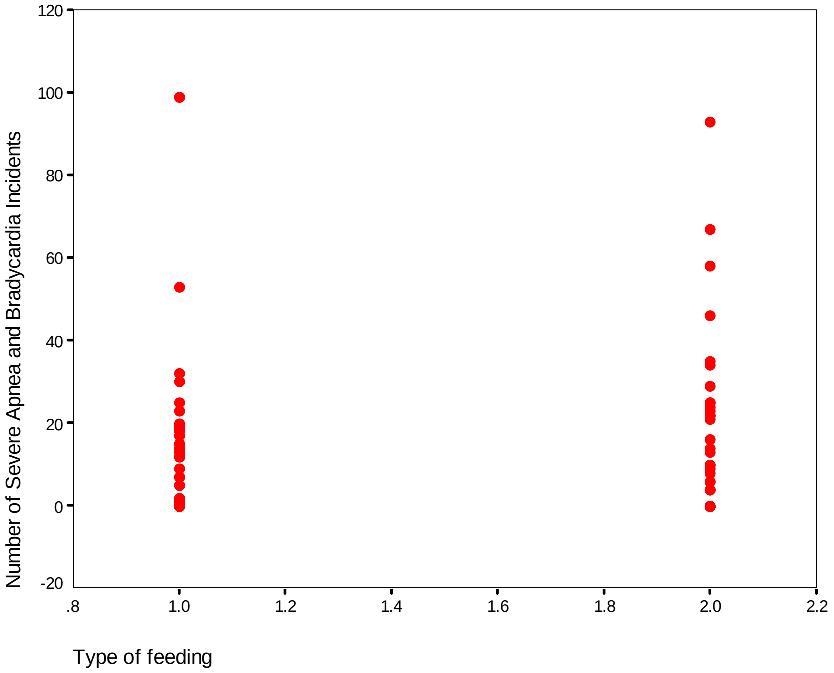


Figure 12. This is a scatterplot of the same data. To create this graph in SPSS, select GRAPHS SCATTER from the menu. Notice that SPSS does not use value labels on the graph, a serious limitation.

One problem with this graph, though, is that it’s hard to figure out how many points there are in certain regions of the graph.

Sunflowers are a good option for crowded graphs

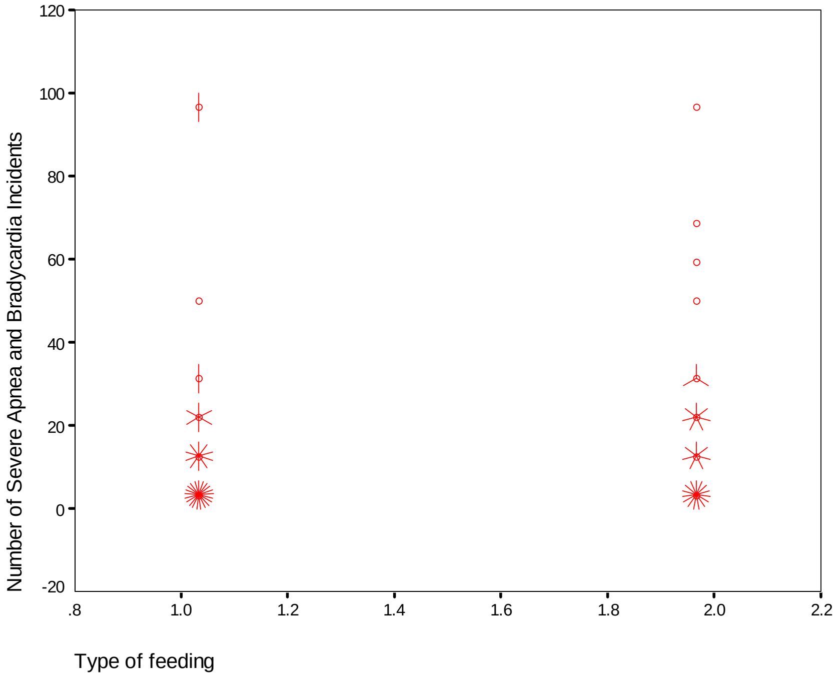


Figure 13. This is a scatterplot of the same data, using sunflowers to show how many data points fall into certain regions of the graph. To create this graph in SPSS, click on the EDIT button, and select CHART OPTIONS from the menu.

The number of “petals” emanating from a point is an indication of the number of data points in that region. An absence of petals indicates a single data point.

Jittering also helps

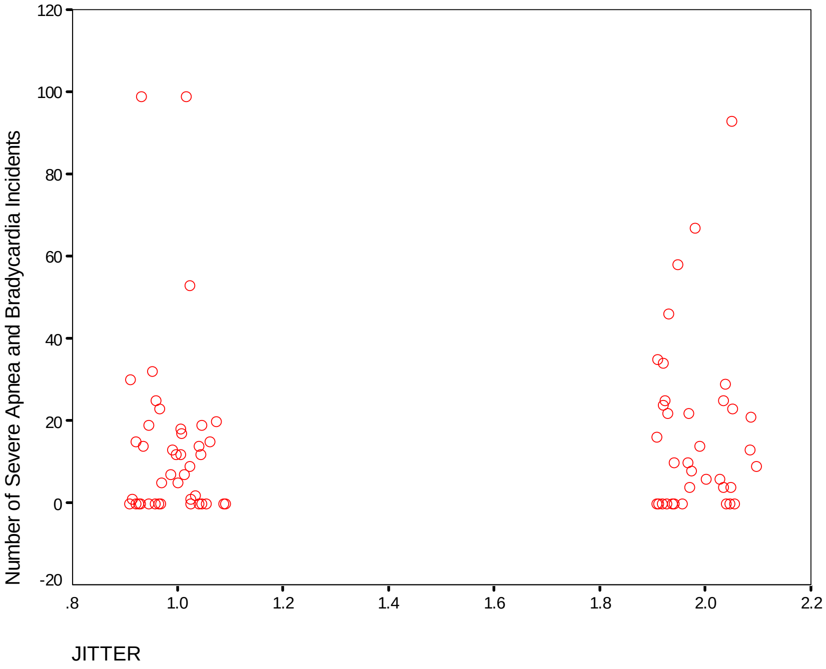


Figure 14. This is a scatterplot of the same data, using jittering. To create the jittered value in SPSS, use the formula JITTER = FEED_TYP + 0.2 * (UNIFORM(1) - 0.5). The use of open instead of solid circles also helps.

Definition: *Jittering* is a slight random shifting of data to avoid overprinting.

Chambers et al (1983)

Wilkinson (1988)

Summary of Part 2

The Error of Error Bars Principle.

- Error bars are confusing and ambiguous. Plot all the data if possible, or use a box plot.

Other things we learned.

- If your plot is crowded, use jittering or sunflowers.

What is the best shape, square or rectangular?

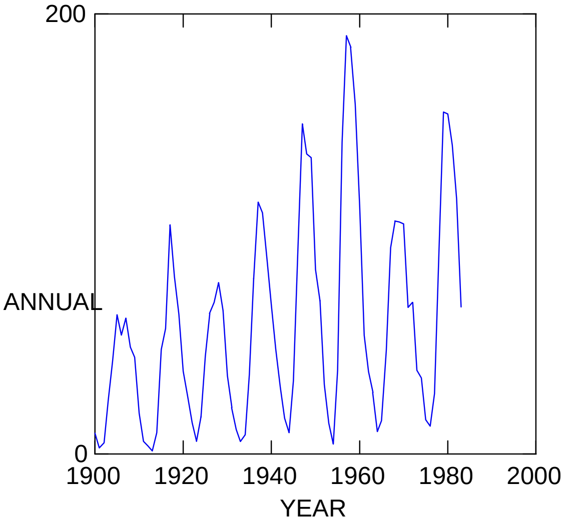


Figure 15. This is a square plot of the annual sun spot numbers across the past decade. This graph was created by SYSTAT.

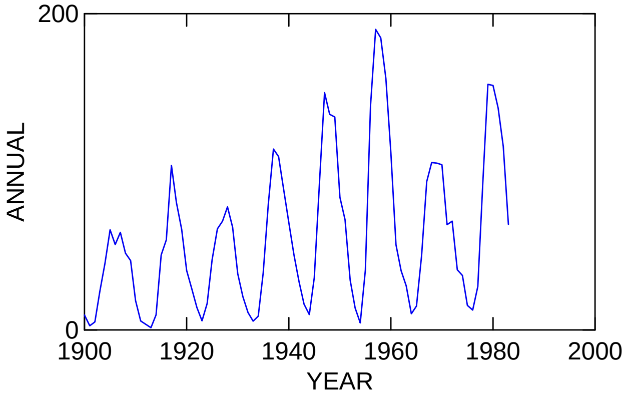


Figure 16. This is a rectangular plot of the same data, again using SYSTAT.

Answer: it depends.

In many scatter plots, the key information is in the slopes of the lines.

The eye can best discriminate slopes when they are close to plus or minus 45 degrees.

This graph is adjusted so that the median absolute slope is 45 degrees.

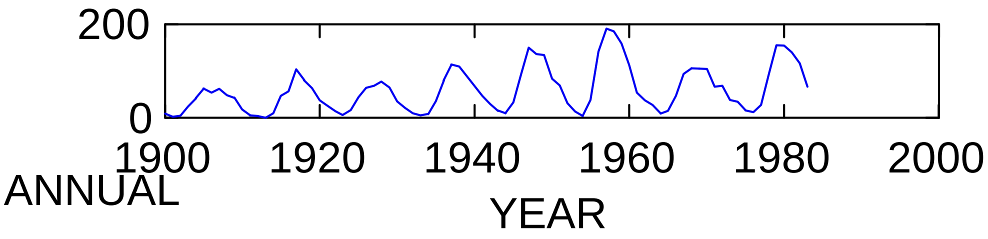


Figure 17. This is a plot of the same data, using a special SYSTAT feature to adjust the shape of the plot so that the median absolute slope is 45 degrees.

Cleveland, W.S., McGill, M.E., and McGill, R. (1988) "The Shape Parameter of a Two-Variable Graph," Journal of the American Statistical Association, 83, 289-300.

Other considerations for the shape of a scatter plot

Labels fit better on rectangular plots.

The human eye seems to be trained to look for horizon effects. It tends to scan side-to-side better than up-and-down.

Don't settle for the default shape. Try different shapes and choose the one that works best.

Tufte (1983)

Should you use grid lines?

Advantage of grid lines

- Assist in projection and/or anchoring.

Disadvantage of grid lines

- Distract from the main picture.

Avoid grid lines if the main focus of the graph is on angles/slopes.

Never use grid lines which are darker than the graph itself.

Chambers et al (1983)

Tufte (1983)

What if most of your data is clustered in a corner of the graph?

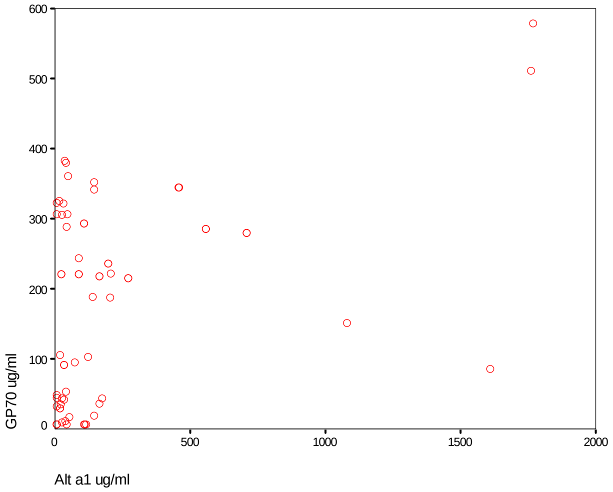


Figure 18. A scatterplot of two mold allergens, GP70 and Alt a1.

A log transformation can sometimes help.

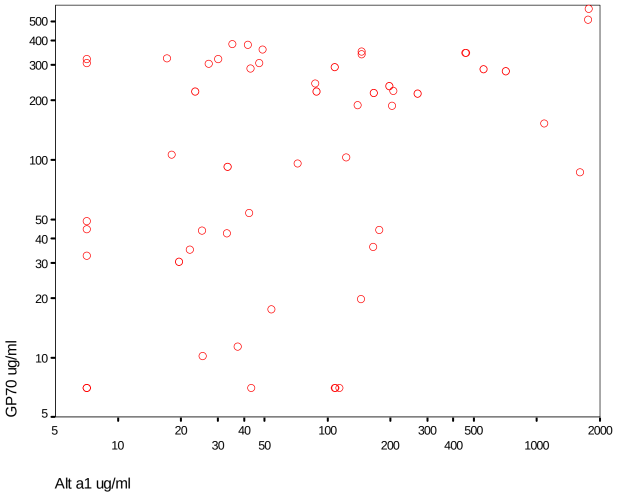
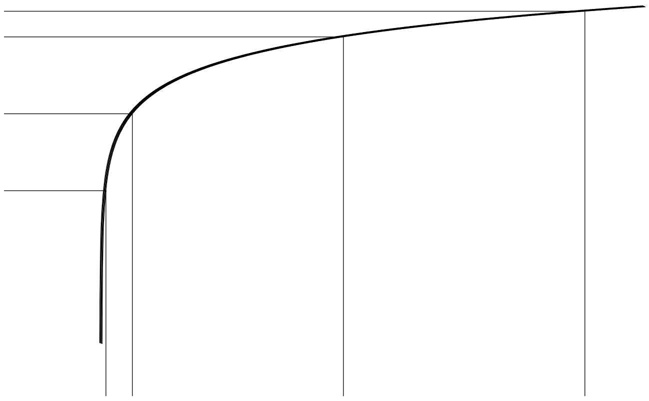


Figure 19. A scatterplot of the same data using log scaling. To make this change in SPSS, click on the EDIT button and select CHART AXIS from the menu.

Why do logarithms help?

Logarithms will tend to stretch the small data values and squeeze the large data values.



Which base to use?

- 2+ orders of magnitude: consider base 10.
- Otherwise, consider base 2.

Cleveland, W.S. (1984) "Graphical Methods for Data Presentation: Full Scale Breaks, Dot Charts, and Multibased Logging," Journal of the American Statistical Association, 38, 270-280.

What is the best way to designate groups?

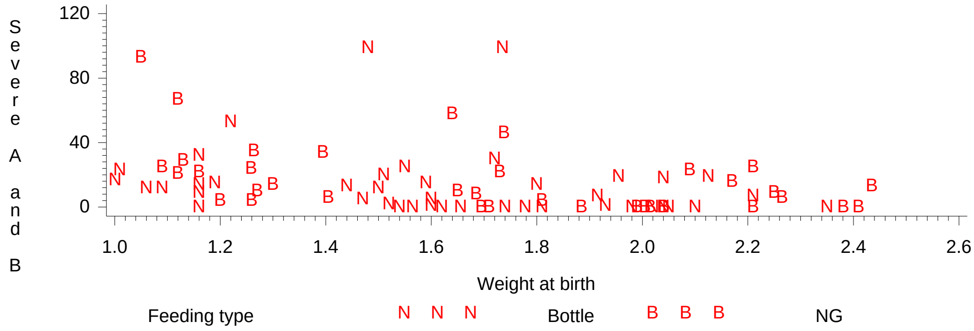


Figure 20. Graph of severe apnea and bradycardia events versus birth weight, using letters to distinguish the feeding type group.

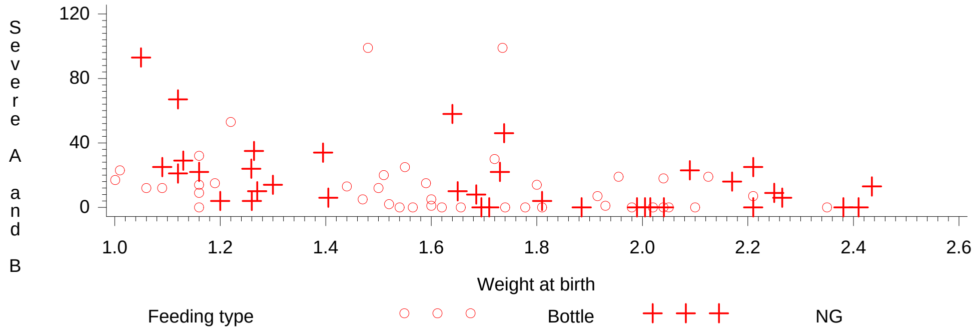


Figure 21. Graph of severe apnea and bradycardia events versus birth weight, using symbols to distinguish the feeding type group.

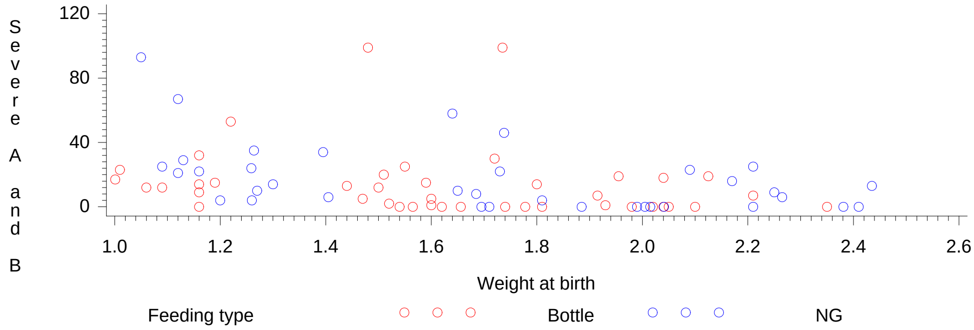


Figure 22. Graph of severe apnea and bradycardia events versus birth weight, using colours to distinguish the feeding type group. Because of the expense and trouble of colour printing/xeroxing, please use your imagination on this graph. All of these graphs created using SAS.

Answer: it depends.

Colour gives the fastest and most accurate discrimination.

Some researchers claim that symbols are better than letters; others claim the reverse. My personal preference is for letters.

Warning: Whether you choose letters or symbols, make sure that your graph can withstand reproduction and/or reduction.

Cleveland, W.S. (1984) "Graphs in Scientific Publications," Journal of the American Statistical Association, 38, 261-269.

Cleveland, W.S., and McGill, R. (1984) "The Many Faces of a Scatterplot," Journal of the American Statistical Association, 79, 807-822.

Lewandowsky, S., and Spence, I. (1989) "Discriminating Strata in Scatterplots," Journal of the American Statistical Association, 84, 682-688.

Danger in the use of letters.

Some letter combinations are poorly discriminable.

Examples of easily confused letters:

E and F

P and R

O and Q

Easily distinguished letters:

M and F

C and E

B and W

Letter have an advantage over symbols and colours when they can make use of a legend unnecessary.

Lewandowsky and Spence (1989)

Danger in the use of symbols

Symbols require use of a legend, which is often distracting.

More than four symbols will tend to overload short term memory.

Certain symbols, such as circles and squares, are easily confused, especially if the symbol size is small.

Danger in the use of colour

Colour should never be used for ordinal data. Shades of gray work better with ordinal data.

Bright colours can lead to optical illusions. For example, areas in bright red sometimes appear larger than areas in bright green.

Certain colour combinations are difficult to distinguish.

- Blue against a black background.
- Yellow against a white background.

More than 8% of all males and more than 1% of all females are colour-blind. A red-green deficiency is most common.

Beatty, J.C. (1983) "Raster Graphics and Color," The American Statistician, 37(1), 60-75.

Cleveland, W.S., and McGill, R. (1983) "A Color-Caused Optical Illusion on a Statistical Graph," The American Statistician, 37(2), 101-105.

Lewandowsky and Spence (1989)

Summary of Part 3

The Size and Shape Principle.

- Carefully consider the size and shape of your graph. Rectangular graphs are sometimes better than square graphs. Bigger is not always better.

Other things we learned.

- If most of your data is clustered in a corner, consider a log transformation.
- Make sure that your choice of colour, symbols, or letters does not deteriorate when reproduced or reduced.

Three-dimensional plots have problems with perspective

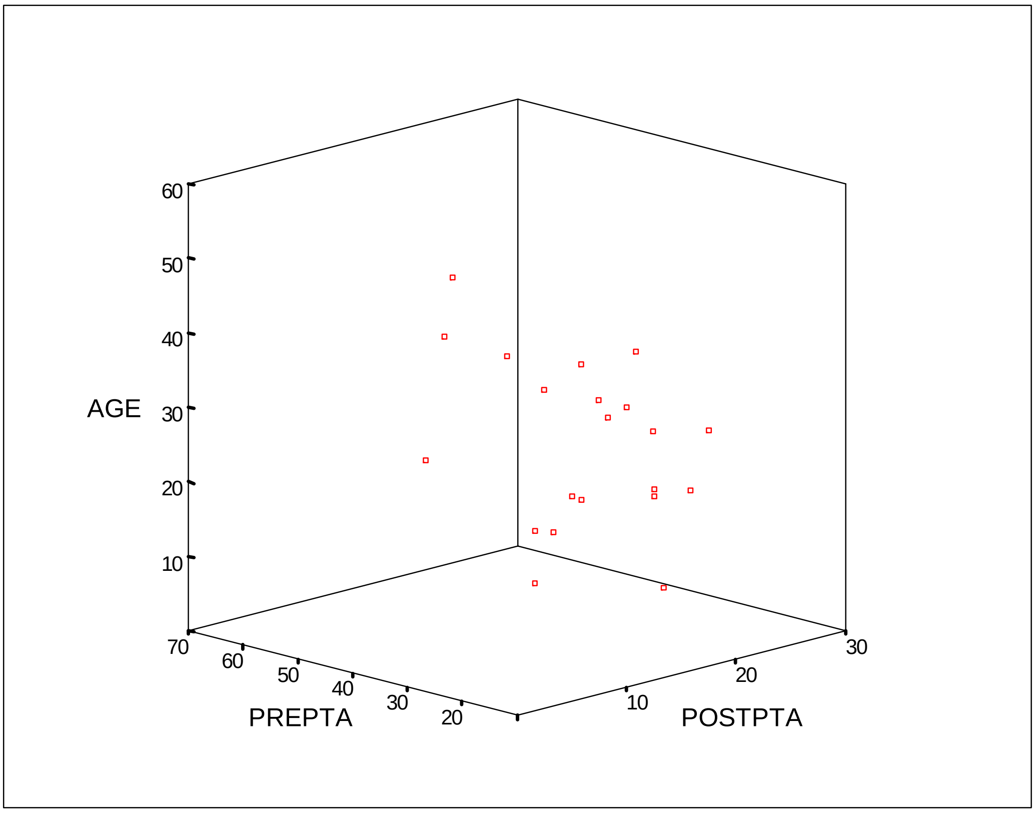


Figure 23. A three dimensional scatterplot of hearing tests prior to and following surgery, along with the patient's age in months. To create this graph in SPSS, select GRAPHS SCATTER from the menu, and 3-D from the dialog box.

Even simple tasks like selecting the largest and smallest values are difficult to do for a three-dimensional graph.

Rotating the plot can sometimes give a better sense of perspective

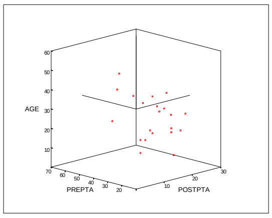


Figure 24

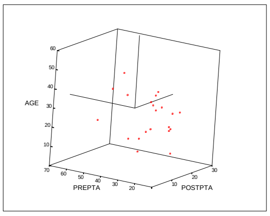


Figure 25

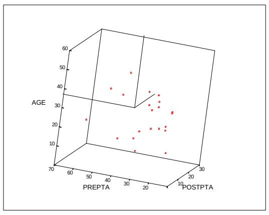


Figure 26

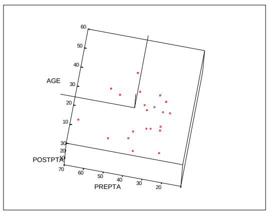


Figure 27

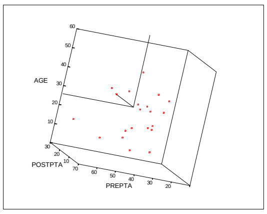


Figure 28

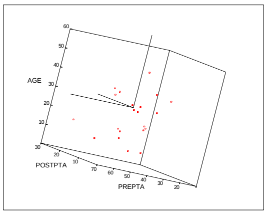


Figure 29

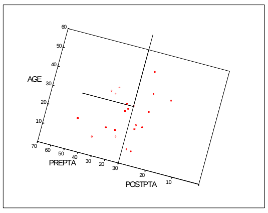


Figure 30

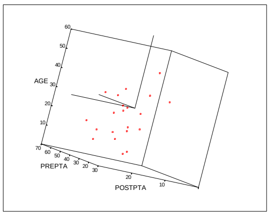


Figure 31

Figures 24-31. Various rotations of the same data. To rotate a three dimensional graph in SPSS, click on the EDIT button. A series of buttons on the toolbar will let you rotate the graph in real time.

An alternative is the scatterplot matrix

The scatterplot matrix is a convenient way of displaying relationships among many variables simultaneously.

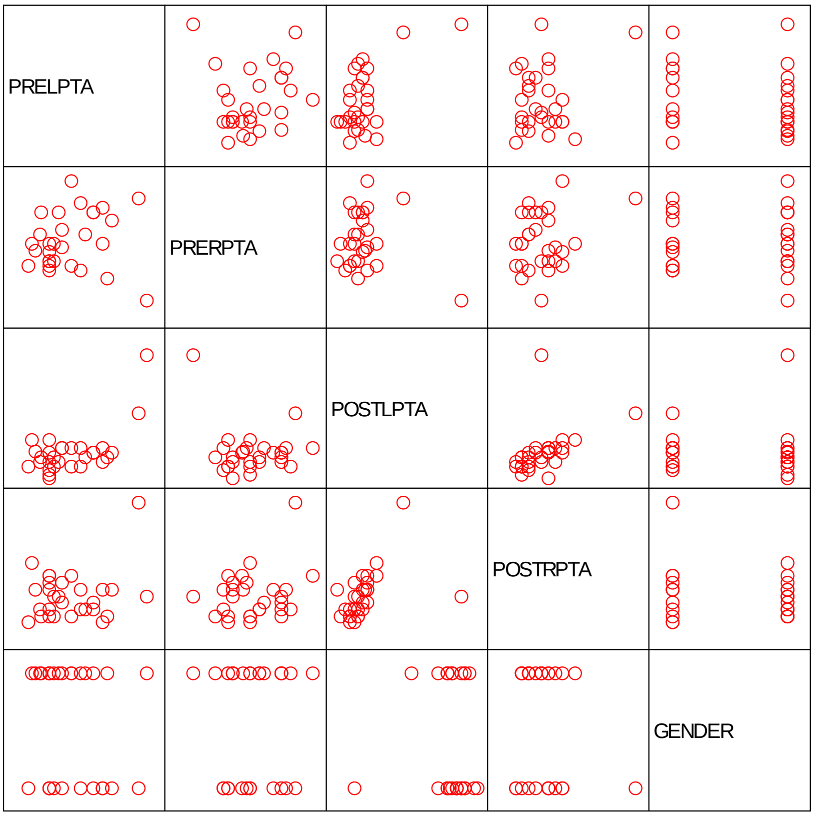


Figure 32. A scatterplot matrix of left and right ear hearing tests, both prior to and following surgery, along with the gender of the patient. To create this graph in SPSS, select GRAPHS SCATTER from the menu, and then select MATRIX from the dialog box.

Conclusion

The Minimum Ink Principle.

The Small Table Principle.

The Error of Error Bars Principle.

The Size and Shape Principle.

…and two bonus principles

The R & R (Reproduction and Reduction) principle.

- Colours, shades of gray, and small symbols may get lost when a journal prints your graph. Make sure your graphs can withstand reproduction and reduction.

The Fault of Default Principle.

- Graphing is an iterative process. Don't rely on the default options provided by your graphics package. Try everything. Re-draw your graphs as often as you rewrite your text.