How to interpret scientific & statistical graphs

Theresa A Scott, MS
Department of Biostatistics
theresa.scott@vanderbilt.edu
http://biostat.mc.vanderbilt.edu/TheresaScott

1

A brief introduction

- Graphics:
  - One of the most important aspects of presentation and analysis of data; help reveal structure and patterns.
- Graphical perception (ie, interpretation of a graph):
  - The visual decoding of the quantitative and qualitative information encoded on graphs.
- Objective:
  - To discuss how to interpret some common graphs.

2

Sidebar: Types of variables

- Continuous (*quantitative* data):
  - Have any number of possible values (eg, weight).
  - *Discrete numeric* – set of possible values is a finite (ordered) sequence of numbers (eg, a pain scale of 1, 2, …, 10).
- Categorical (*qualitative* data):
  - Have only certain possible values (eg, race); often not numeric.
  - *Binary (dichotomous)* – a categorical variable with only two possible value (eg, gender).
  - *Ordinal* – a categorical variable for which there is a definite ordering of the categories (eg, severity of lower back pain as none, mild, moderate, and severe).

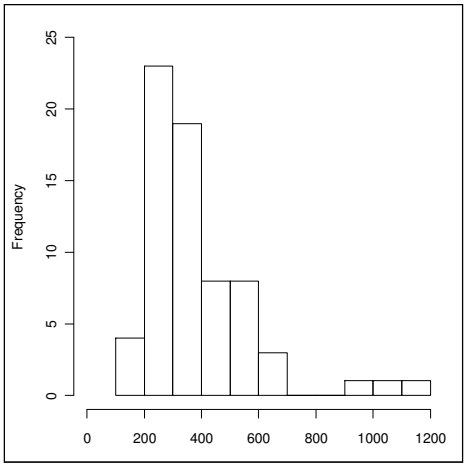
3

Graphs for a single variable's distribution

4

Histograms

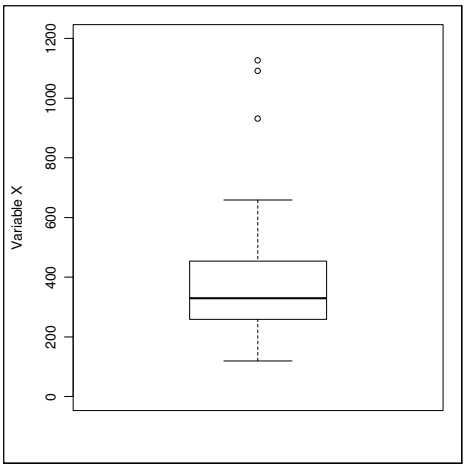
- Continuous variable.
- Values are divided into a series of intervals, usually of equal length.
- Data are displayed as a series of vertical bars whose heights indicate the number (count) or proportion (percentage) of values in each interval.
- What is the overall shape? Is it symmetric? Is it skewed?
  - Affected by the size of the interval.
- Is there more than one peak?
- What is the range of the intervals? Is the shape wide or tight (ie, what's the *variability*?)
- Look for concentration of points and/or *outliers*, which can distort the graph.



5

Boxplots

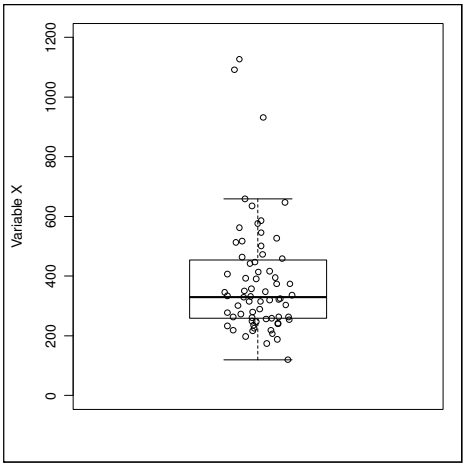
- Continuous variable.
- Displays a numerical summary of the distribution.
  - Most include the 25th, 50th (median), and 75th *percentiles*.
  - Optionally includes the mean (average).
  - May extend to the min & max or may use a rule to indicate outliers.
  - Graphed either horizontally or vertically.
- Interpretation:
  - What statistics are displayed?
  - Most often, the central box includes the middle 50% of the values.
  - Whiskers (& outliers) show the "range".
  - Symmetry is indicated by box & whiskers and by location of the median (and mean).



6

Boxplot with raw data

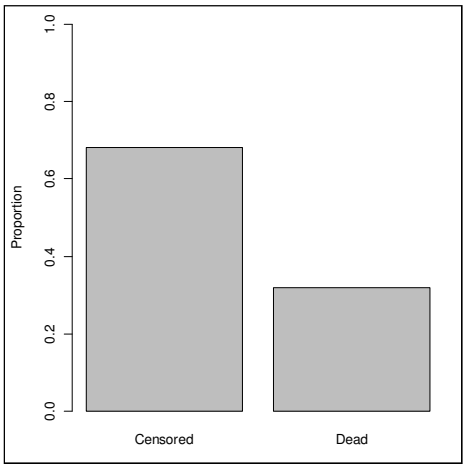
- Going one step beyond just a boxplot.
  - Boxplot is overlaid with the raw values of the continuous variable.
  - Therefore, displays both a numerical summary as well as the actual data.
  - Gives a better idea the number of values the numerical summary (ie, boxplot) is based on and where they occur.
- Raw values are often "jittered" – that is, in order to visually depict multiple occurrences of the same value, a random amount of noise is added in the horizontal direction (if boxplot is vertical; in the vertical direction if the boxplot is horizontal).
- Look for concentration of points and (as before) outliers.



7

Barplots (aka, bar charts)

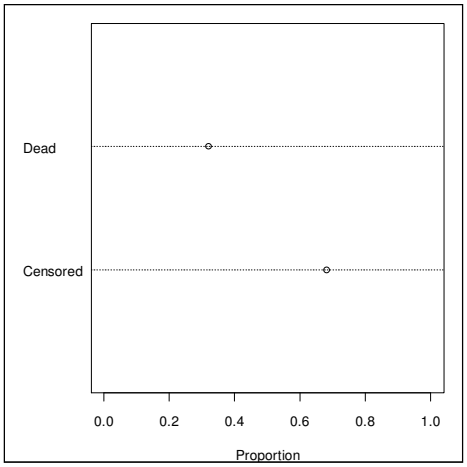
- Categorical variable.
- Data are displayed as a series of vertical (or horizontal) bars whose heights indicate the number (count) or proportion (percentage) of values in each category.
  - Visual representation of a table.
  - How do the heights of the bars compare? Which is largest? Smallest?



8

Dot plots (aka, dotcharts)

- Categorical variable.
- Alternative to a barplot (bar chart).
- Height of the (vertical) bars are indicated with a dot (or some other character) on a (often horizontal dotted) line.
  - Line represents the counts or percentages.
- Same interpretation as barplot (bar chart).



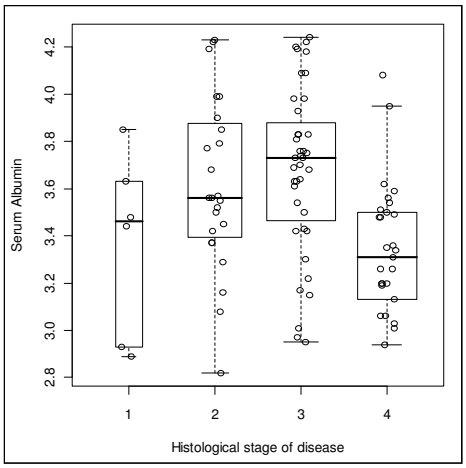
9

Graphs for the association/relation between two variables

10

Side-by-side boxplots

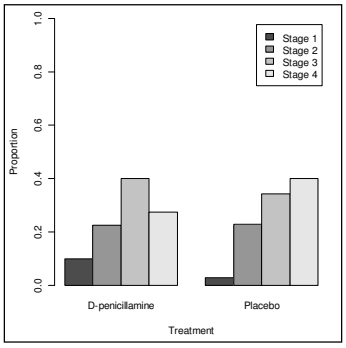
- A continuous variable and a categorical variable.
- Displays the distribution of the continuous variable within each category of the categorical variable.
- Width of the boxes can also be made proportional to the number of values in each category.
- Here, side-by-side boxplots are overlaid with the raw values.
- How does the symmetry of each boxplot differ across categories? How do they compare to the boxplot of the continuous variable ignoring the categorical variable? Is there a concentration of points and/or outliers in one particular category? Is the number of values in each category fairly consistent?

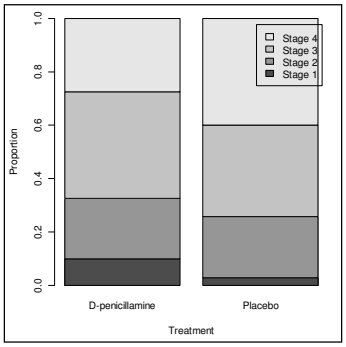


11

Barplots

- Two categorical variables.
  - Visual representation of a two-way table.
- Bars are most often "nested".
  - The count/proportion of the 2nd variable's categories is displayed within each of the 1st variable's categories.
  - Allows you to compare the 2nd variable's categories (1) *within* each of the 1st variable's categories, and (2) *across* the 1st variable's categories.
- Bars can also be "stacked".
  - A single bar is constructed for each category of the 1st variable & divided into segments, which are proportional to the count/ percentage of values in each category of the 2nd variable.
  - Counts should sum to the no. of values in the dataset; percentages should sum to 100%.
  - Unlike "side-by-side", segments do not have a common axis – makes difficult to compare segment sizes across bars.

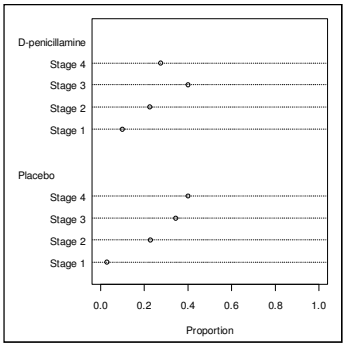


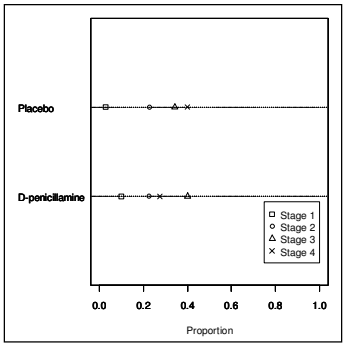


12

Dot plots

- Two categorical variables.
  - Alternative visual representation of a two-way table.
- Like barplots, can be "nested".
  - Have different lines for each category of the 2nd variable grouped for each category of the 1st variable.
- Can also be "stacked".
  - Categories of the 2nd variable are shown on a *single* line; one line for each category of the 2nd variable; 1st variable's categories are distinguished with different symbols.
  - Unlike "stacked" barplots, do have a common axis for comparisons.
- Same interpretation as barplot (bar chart).
  - Same comparisons – within and across categories.





13

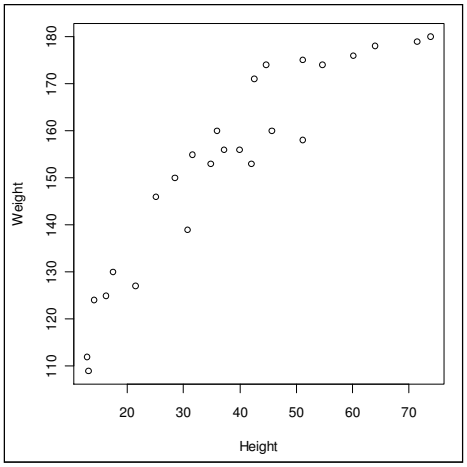
Scatterplots

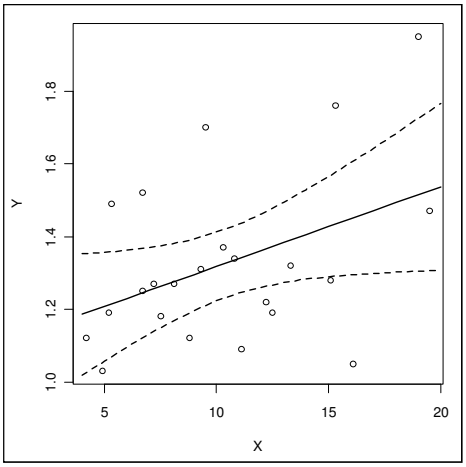
- Two continuous variables.
- Usually, the "response" variable (ie, outcome) is plotted along the vertical (y) axis and the explanatory variable (ie, predictor; risk factor) is plotted along the horizontal (x) axis.
  - Doesn't matter if there is no distinction between the two variables.
- Each "subject" is represented by a point.
- Often include *lines* depicting an estimate of the linear/non-linear relation/ association, and/or confidence "bands".
- What to look for :
  - Overall pattern: Positive association/ relation? Negative association/ relation? No association/relation?
  - Form of the association/relation: Linear? Non-linear (ie, a curve)?
  - Strength of the relation/association: How tightly clustered are the points (ie, how variable is the relation/ association)?
  - Outliers
  - "Lurking" variables: A 3rd (continuous or categorical) variable that is related to both continuous variables and may *confound* the association/relation.
    - Often incorporated into graph – see "Graphs for mutlivariate data" slides.

http://www.stat.sfu.ca/~cschwarz/Stat-201/Handouts/node41.html

14

Example Scatterplots





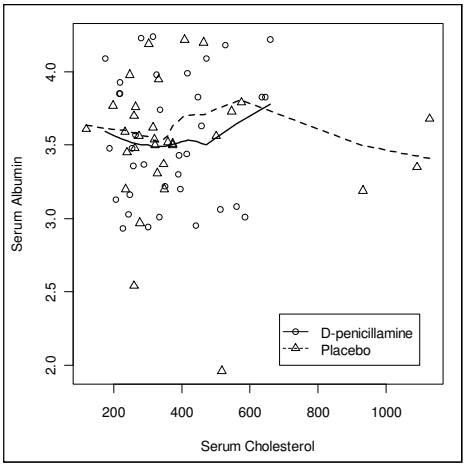
15

Graphs for multivariate data (ie, more than two variables)

16

(More complex) Scatterplots

- Two continuous variables and a categorical variable.
- Often, categorical variable is a confounder – the association/relation between the two continuous variables is (possibly) different between the categories of the categorical variable.
- Categorical variable incorporated using different symbols and/or line types for each category.
- What to look for:
  - Same as mentioned for general scatterplot.
  - Does the association/relation between the two continuous variables differ between the categories of the categorical variable? If so, how?



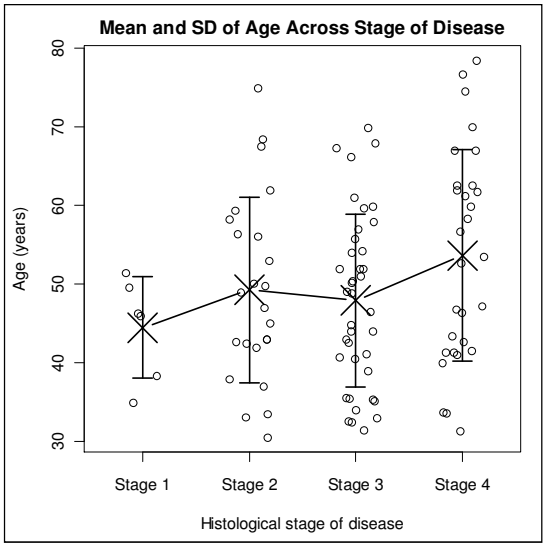
17

Examples of other graphs you might encounter

18

Modified “side-by-side boxplot”

(great alternative to a “dynamite plot” –next slide)



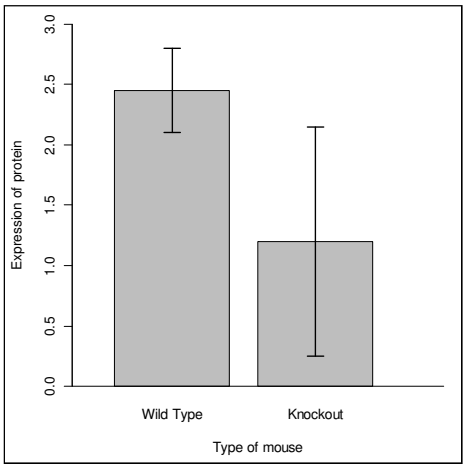
19

“Dynamite plot”

(often, height of bar = mean; error bar = standard deviation)

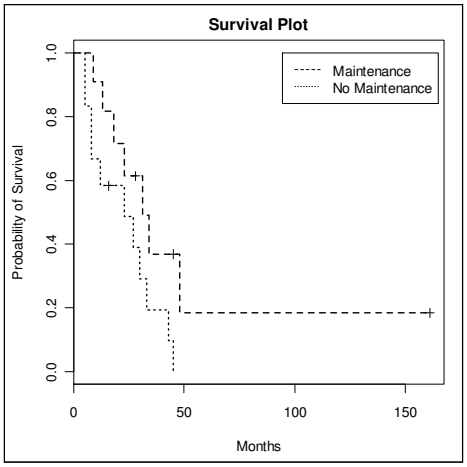
IMPORTANT

- Even though commonly seen, not a good graph to generate.
  - Interested in the height of the bar (rest of the bar is just unnecessary ink).
  - Have no idea how many values the mean and standard deviation are based on (often quite small) or how the raw values are distributed.
  - Both affect the values of the mean and standard deviation.
  - Bars can also be “hanging”, which may represent *negative* values – very confusing.

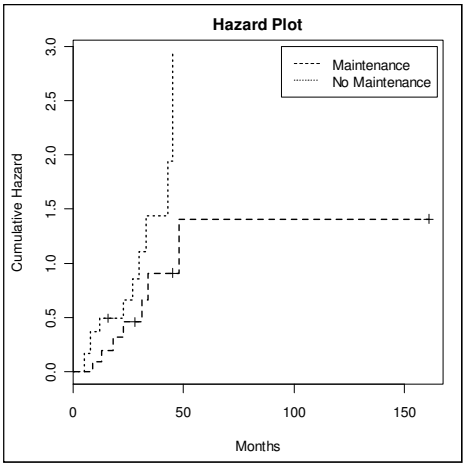


20

Survival & Hazard plots



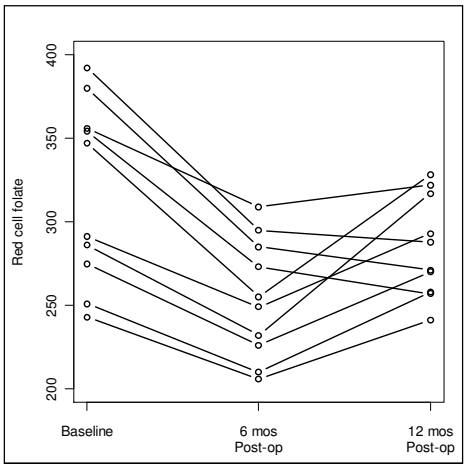
Each step down represents one or more "deaths"; "+" signs represent *censoring*.



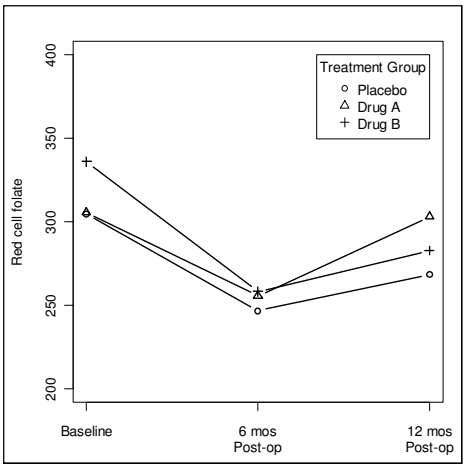
Each step up represents one or more "deaths"; "+" signs represent *censoring*.

21

"Spaghetti" & Line plots



Each line plots the raw data points of a single "subject".



Each line plots summary measures (eg, mean) from a group of subjects.

22

WARNING: Very easy for a graph to lie

- What are the limits of the axis/axes? Is the scale consistent?
- How do the height and width of the graph compare to each other? Is the graph a square? A rectangle (ie, short & wide; tall & skinny)?
- If two or more graphs are shown together (eg, side-by-side, or in a 2x2 matrix), do all of the axes have the same limits? Same scale? Do they have the same relative dimensions?
- Are there two x- or y-axes in the same graph? If so, do they have the same scale?
- Can you get a feel for the raw data? The number of data points?
- Does a graph of a continuous variable show outliers? Does the data look too "pretty"?

23

General steps

- Do I understand this graph?
  - If NO: (1) it might be a really bad graph; or (2) it might be a type of graph you don't know about.
- Carefully examine the axes and legends, noting any oddities.
- Scan over the whole graph, to see what it is saying, generally.
- If necessary, look at each portion of the graph.
- Re-ask "Do I understand this graph?"
  - If YES, what is it saying?
  - If NO, why not?

"Overview of Statistical Graphs", Peter Flom 24