

Constructing scientific & statistical graphs

Theresa A Scott, MS

Department of Biostatistics

theresa.scott@vanderbilt.edu

<http://biostat.mc.vanderbilt.edu/TheresaScott>

1

Graphics: A brief introduction

- Can be defined as visual translations of numbers & other data derived from tabulated observations or measurements.
- One of the most important aspects of analysis & presentation of your data.
- Help reveal the structure of your data & any patterns it may contain.
- Can be used to explain & illustrate known facts & relationships, but can also be used to explore & search for relevant questions & working hypotheses.
- Objective of this lecture: To discuss some common graphs you may choose to construct.

2

Sidebar I: Roles of variables

- Outcome: variable that is the focus of the study, whose variation or occurrence you are seeking to understand.
- Predictor: variable that may influence the size or occurrence of the outcome (aka exposure variable, risk factor).
 - In a clinical trial, includes the intervention.
- Confounder: variable that differs between values of the predictor variable and which also affects the outcome.
 - Needed to correctly estimate the relation between the outcome(s) and predictor(s).
- Often more than one outcome, predictor, & confounder in your data.

3

Sidebar II: 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).

4

Which graph do I choose?

- Think of the data (ie, variable(s), “column(s)” in your spreadsheet) you wish to graph:
 - How many variables? – One? Two? Three or more?
 - What types of variable(s)? – Continuous variables? Categorical variables? A combination of continuous & categorical variables?
 - What roles of variable(s)? – Outcome? Predictor? Confounder? Combination?
 - How many values per unit of measurement? - One record per patient? Multiple measurements/ records per patient?
 - What do I want to communicate? – Distribution of a variable? Association/relation between two (or more) variables? Something else?

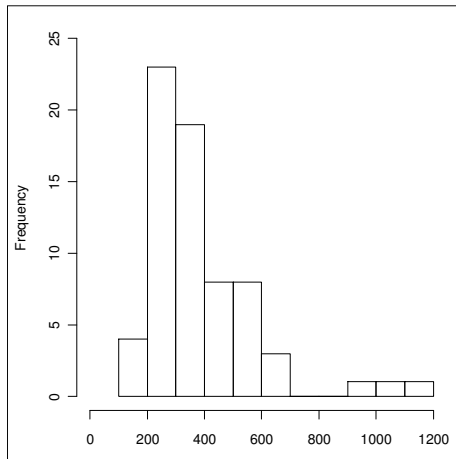
5

Graphs to illustrate a single variable's distribution

6

Histograms

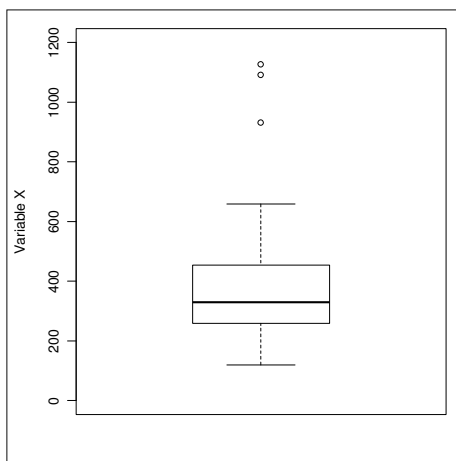
- Continuous variable.
- Values are divided into a series of intervals, usually of equal length.
- Data are displayed as a series of vertical bars whose heights indicate the number (count) or proportion (percentage) of values in each interval.
- Communicates
 - Overall shape
 - Symmetry / skewness
 - Variability
 - Concentration of points and/or outliers



7

Boxplots

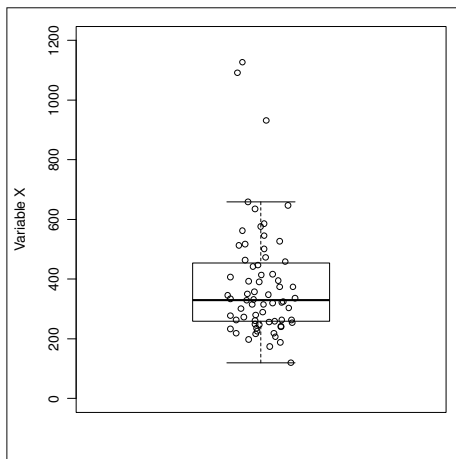
- Continuous variable.
- Displays a numerical summary of the distribution.
 - Most include the 25th, 50th (median), & 75th *percentiles*.
 - Optionally includes the mean (average).
 - May extend to the min & max or may use a rule to indicate outliers.
 - Graphed either horizontally or vertically.
- Communicates
 - Symmetry / skewness
 - Variability
 - Outliers



8

Boxplot with raw data

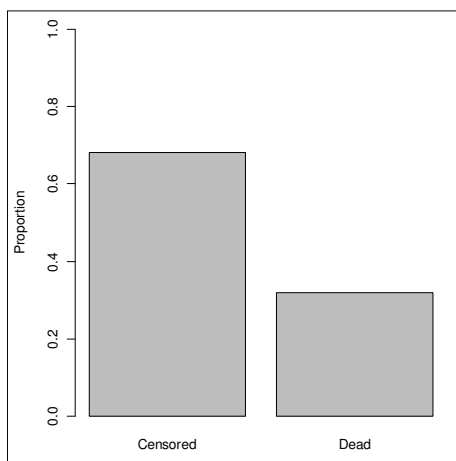
- Going a step beyond just a boxplot.
 - Boxplot is overlaid with the raw values of the continuous variable.
 - Therefore, displays both a numerical summary as well as the actual data.
 - Gives a better idea of the number of values the numerical summary (ie, boxplot) is based on & where they occur.
- 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).



9

Barplots (aka, bar charts)

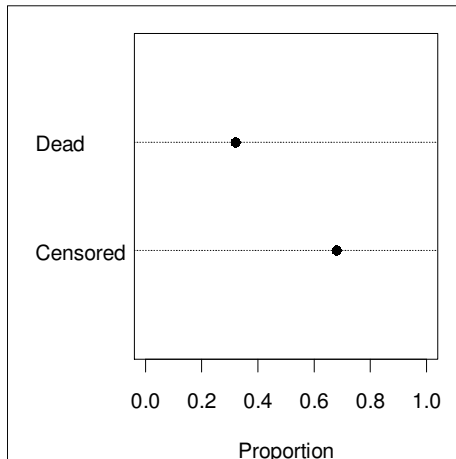
- Categorical variable.
- Data are displayed as a series of vertical (or horizontal) bars whose heights indicate the number (count) or proportion (percentage) of values in each category.
 - Visual representation of a table.
- Communicates data within & across categories.



10

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.



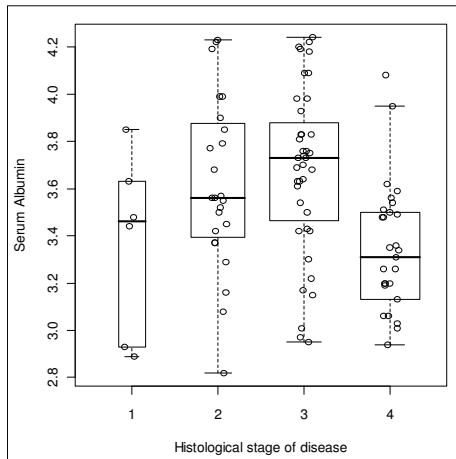
11

Graphs to illustrate the
association/relation
between two variables

12

Side-by-side boxplots

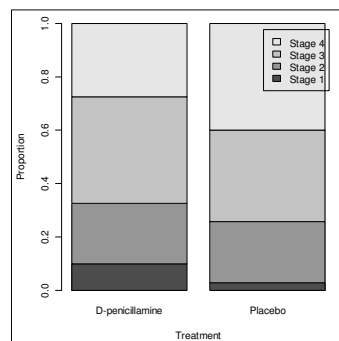
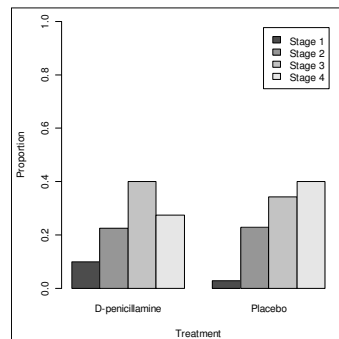
- A continuous variable & a categorical variable.
- Displays the distribution of the continuous variable within each category of the categorical variable.
- Width of the boxes can also be made proportional to the number of values in each category.
- Here, side-by-side boxplots are overlaid with the raw values.
- Communicates data within & across categories of the categorical variable.



13

Barplots

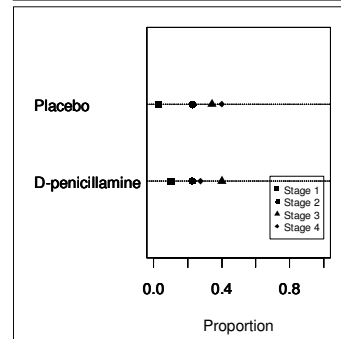
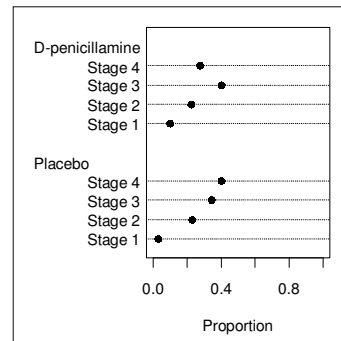
- Two categorical variables.
 - Visual representation of a two-way table.
- Bars are most often “nested”.
 - The count/proportion of the 2nd variable's categories is displayed within each of the 1st variable's categories.
 - Allows you to compare the 2nd variable's categories (1) *within* each of the 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.



14

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.
- Communication same data as barplot (bar chart; within and across categories), but often a lot cleaner.



15

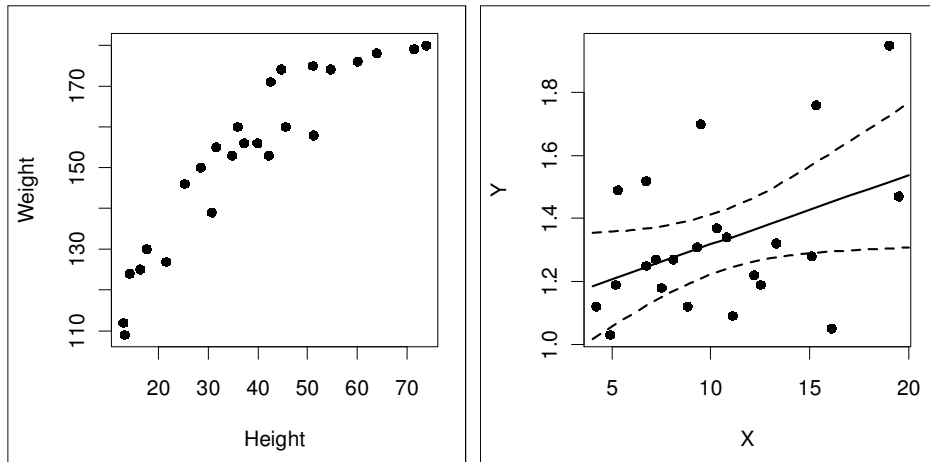
Scatterplots

- 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”.
- Communicates:
 - Overall pattern (positive, negative, or no association/ relation).
 - Form of the association/relation (linear or non-linear (ie, a curve)).
 - Strength of the relation/association (ie, how tightly clustered are the points; how variable is the relation/ association)
 - Outliers
 - Possible “lurking” variables: A 3rd (continuous or categorical) variable that is related to both continuous variables & may *confound* the association/relation.
 - Often incorporated into graph – see “Graphs for multivariate data” slides.

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

16

Example Scatterplots



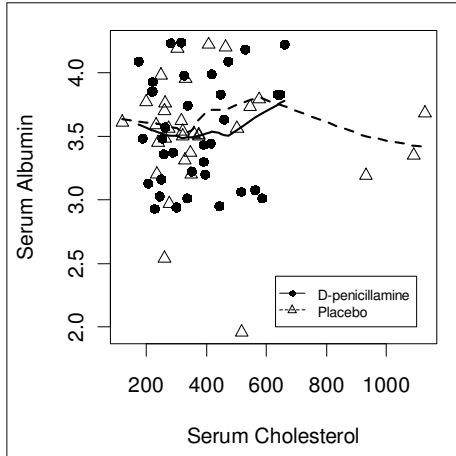
17

Graphs to illustrate
multivariate data
(ie, more than two variables)

18

(More complex) Scatterplots

- Two continuous variables & 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.
- Communicates data within & across categories of the categorical variable.

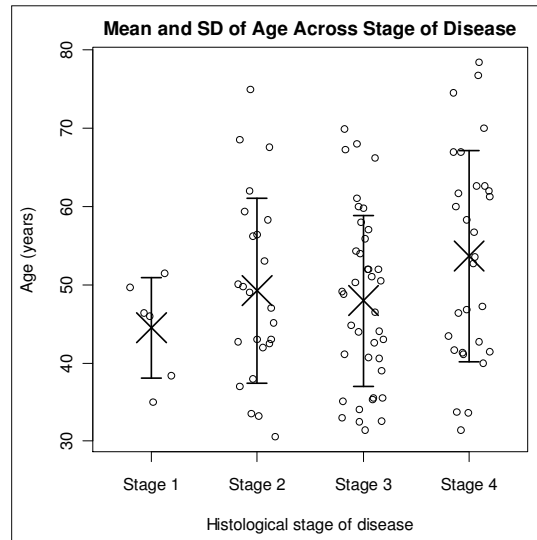


19

Examples of other graphs
you might construct

20

Modified “side-by-side boxplot” (great alternative to a “dynamite plot” –next slide)



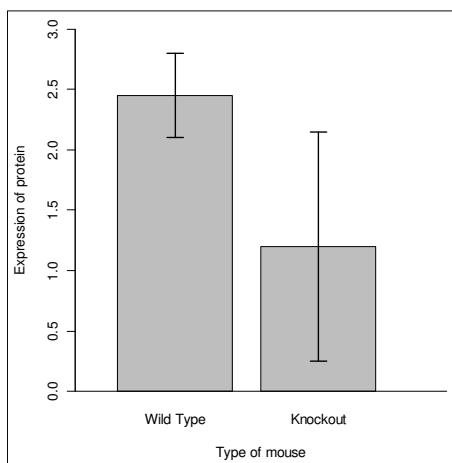
21

“Dynamite plot”

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

IMPORTANT

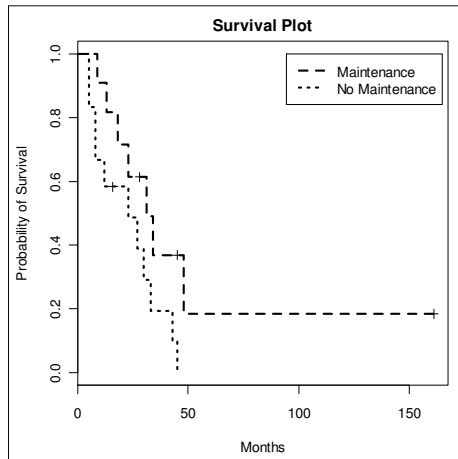
- Even though commonly seen, not a good graph to generate.
 - Interested in the height of the bar (rest of the bar is just unnecessary ink).
 - Have no idea how many values the mean and standard deviation are based on (often quite small) or how the raw values are distributed.
 - Both affect the values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



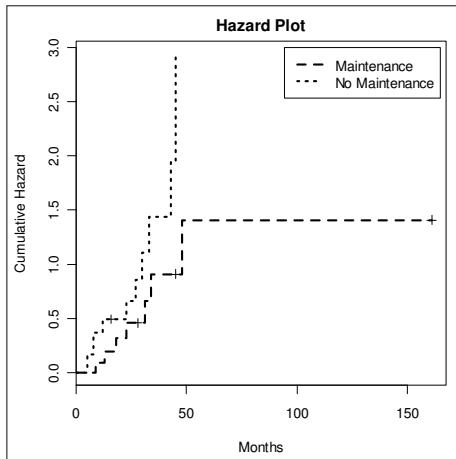
22

Survival & Hazard plots

(Kaplan Meier curves & Cox Proportional Hazards regression)



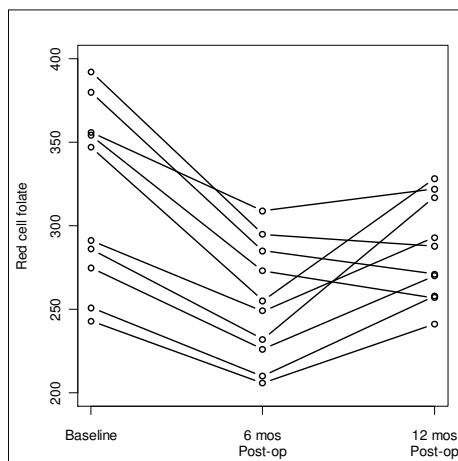
Each step down represents one or more "deaths"; "+" signs represent *censoring*.



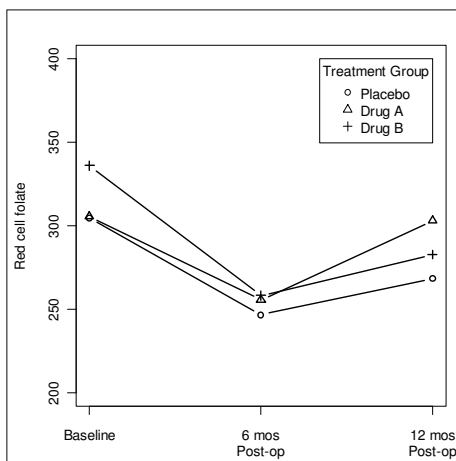
Each step up represents one or more "deaths"; "+" signs represent *censoring*.

23

"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.

24

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”?

25

Constructing effective graphics: *Graphical perception*

- From Ross Ihaka's lecture on “Perception” from his Statistical Graphics course:¹
 - When we draw a graph we encode a numerical value as a graphical attribute.
 - When we look at a graph the aim is to decode the graphical attributes & extract information about the numbers which were encoded.
 - To design effective graphs we must know which graphical attributes are most easily decoded.
 - We need a selection of possible graphical attributes & an ordering of their ease of decoding.
- Most notable references regarding “graphical perception”: Cleveland¹ & McGill
 - (1) “*Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods*” (Journal of the American Statistical Association, Sept 1984).
 - (2) “*Graphical Perception and Graphical Methods for Analyzing Scientific Data*” (Science, Aug 1985).

¹ - <http://www.stat.auckland.ac.nz/~ihaka/courses/787/>

26

Constructing effective graphics: *Recommendations*

- Use the highest possible encoding on the “Cleveland-McGill scale”.
 - The preferred encodings are:
 - Position on a common scale (eg, scatter plot, dotplot).
 - Position on identical, unaligned scales (eg, a group of scatter plots).
 - Length (eg, barplot).
- Be careful when using angles & slopes when encoding numerical values (eg, pie chart).
- Don’t use area or volume to encode numerical values (eg, bubble chart).
- Don’t use color to encode numerical values (eg, heatmap).
- IMPORTANT: scale isn’t meant to be a definitive answer on what to use/not use.
 - Cleveland & McGill note: it presents a “framework within which to work.”
 - Always remember the context, the audience, & what you’re trying to accomplish.
 - ALSO REMEMBER: sometimes a table is “better”.

1 – Has authored several books (eg, “The Elements of Graphic Data”, 1994). E.R. Tufte is another recommended author.

27

Software options & Getting help

- Software options:
 - Excel – will find it very limiting
 - SPSS
 - Stata
 - SAS
 - R (what most of us use)
- Getting help:
 - Your Biostats collaboration
 - Daily Biostatistics Clinics
 - VICTR (via StarBRITE) – Biostats “Voucher”

28