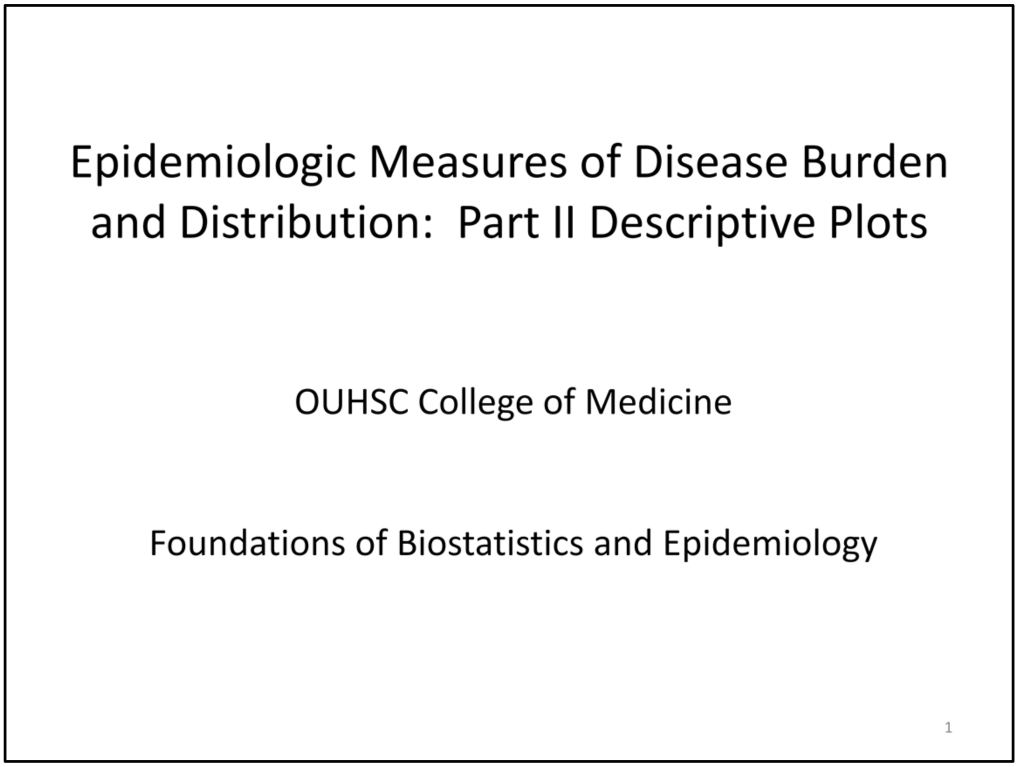


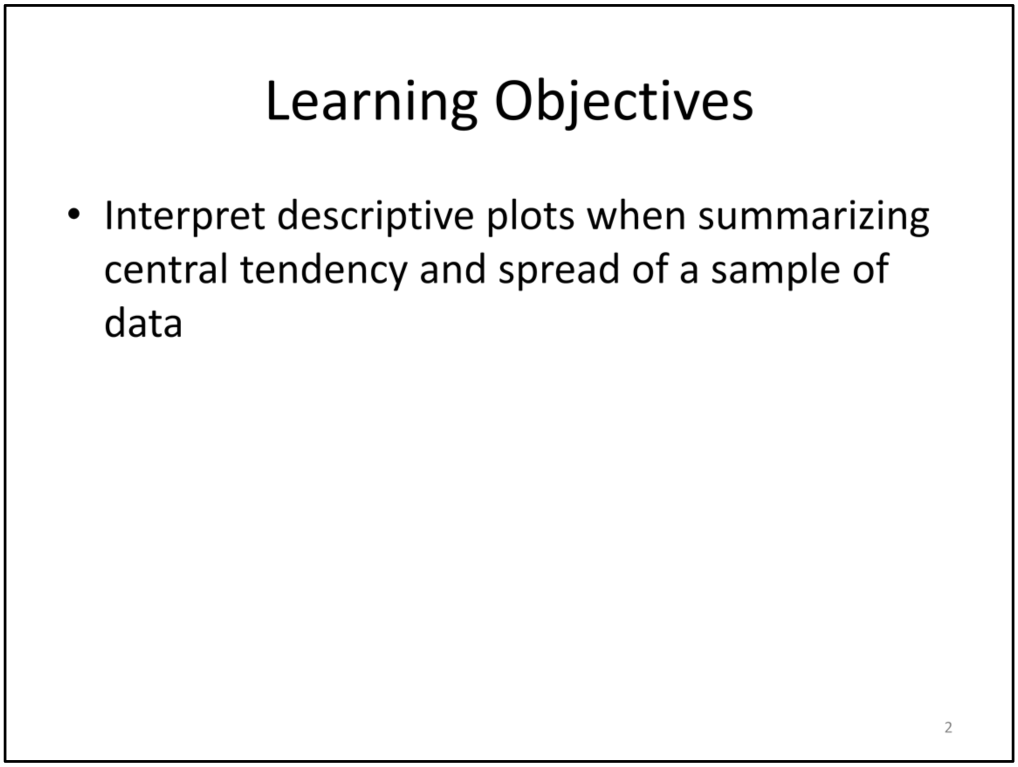
Epidemiologic Measures of Disease Burden and Distribution: Part II Descriptive Plots

OUHSC College of Medicine

Foundations of Biostatistics and Epidemiology

In this 4-part series, we will discuss epidemiologic measures of disease burden and distribution.

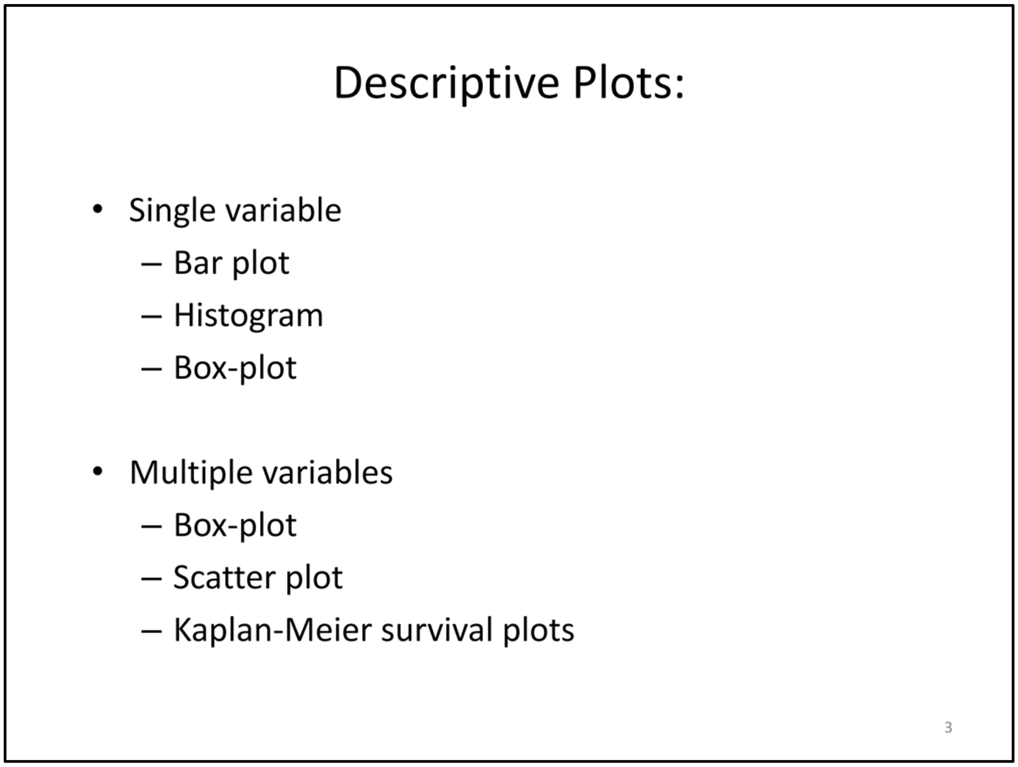
In this second module, I present information related to descriptive plots.



Learning Objectives

- Interpret descriptive plots when summarizing central tendency and spread of a sample of data

After completing this module, you will be able to interpret descriptive plots when summarizing the central tendency and spread of a sample of data.

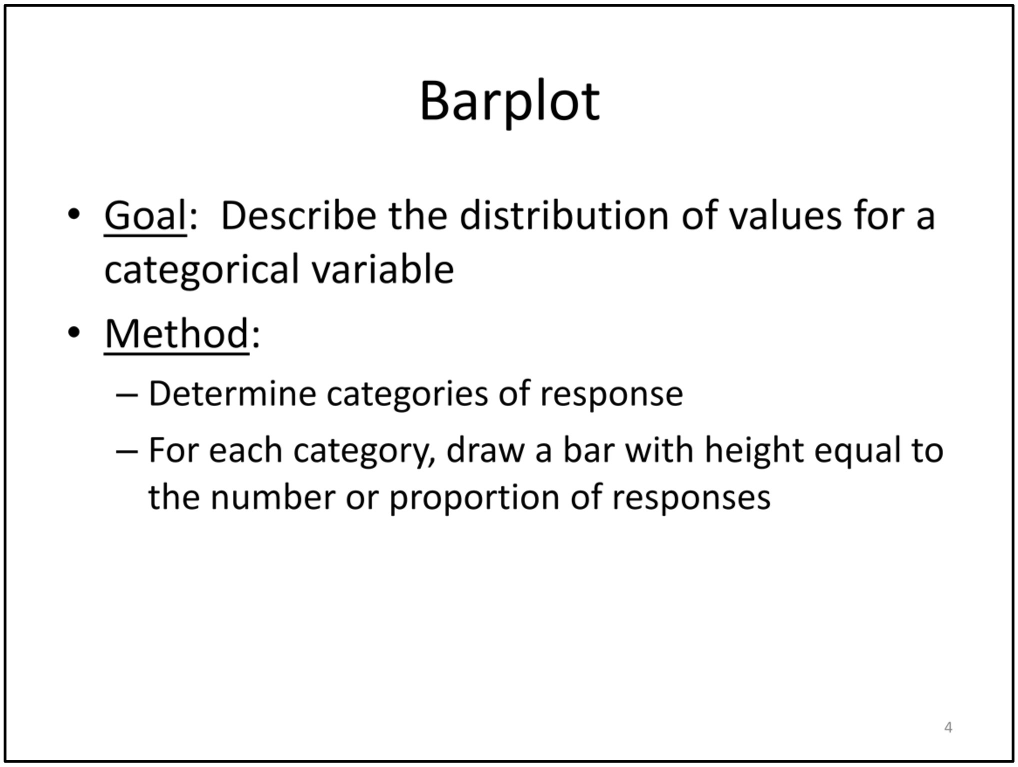


Descriptive Plots:

- Single variable
  - Bar plot
  - Histogram
  - Box-plot
- Multiple variables
  - Box-plot
  - Scatter plot
  - Kaplan-Meier survival plots

3

We will discuss two main types of plots, those that summarize single variables and those that can be used to summarize multiple variables.

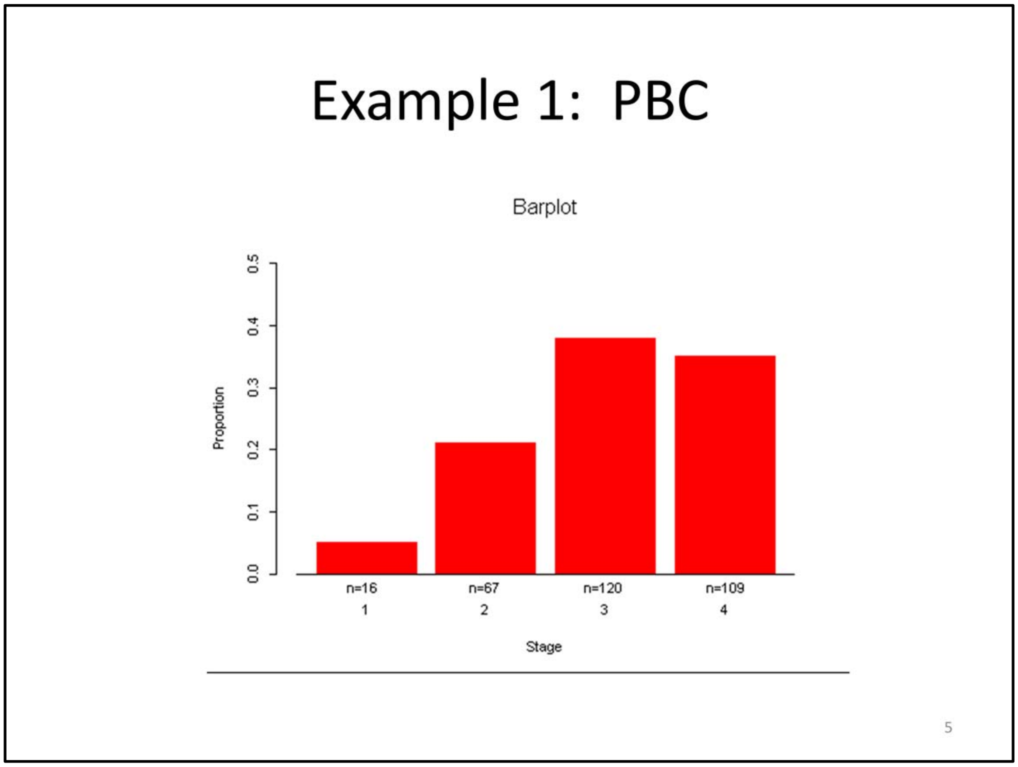


Barplot

- Goal: Describe the distribution of values for a categorical variable
- Method:
  - Determine categories of response
  - For each category, draw a bar with height equal to the number or proportion of responses

Let’s first begin with an overview of a barplot.

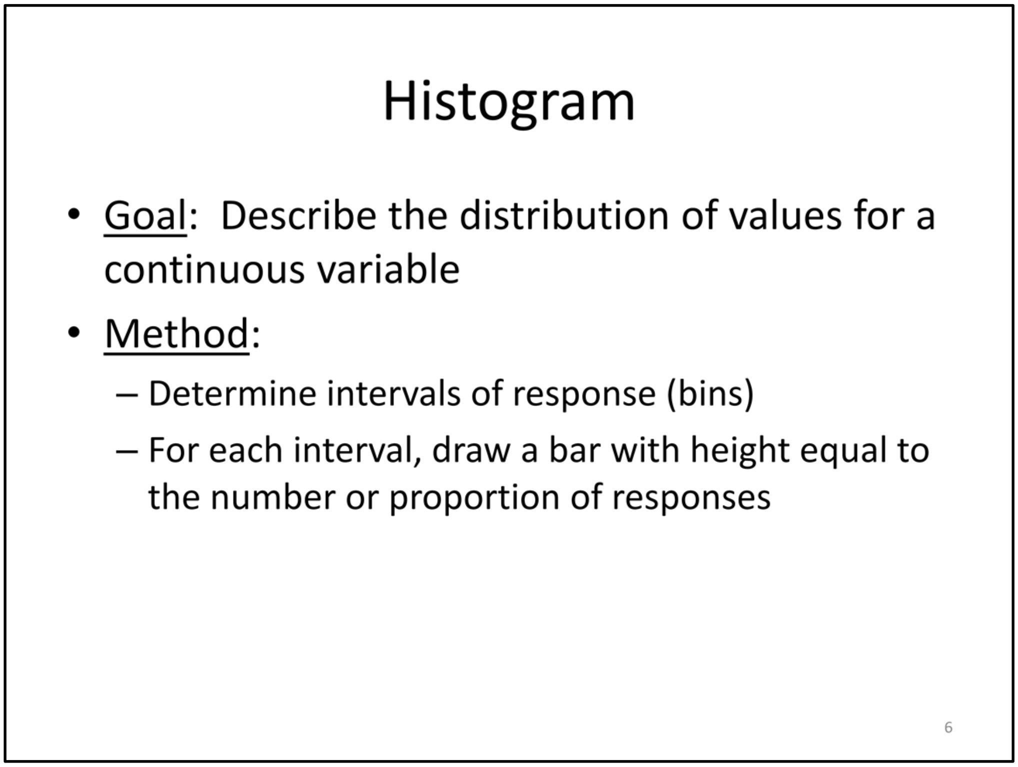
A barplot can be used to summarize the distribution of a categorical variable. It is created by first determining categories of response, such as male/female or age groups defined by 10-year increments. Then, a bar is drawn with a height that is equal to the number or percentage of participants in the given category.



Example 1: PBC

Throughout this lecture we will use data from a clinical trial investigating the effect of D-penicillamine (DPA) in prolonging the overall survival of patients with primary biliary cirrhosis of the liver. The investigators collected detailed clinical, demographic, and biochemical information on patients at baseline to identify factors associated with poorer prognosis.

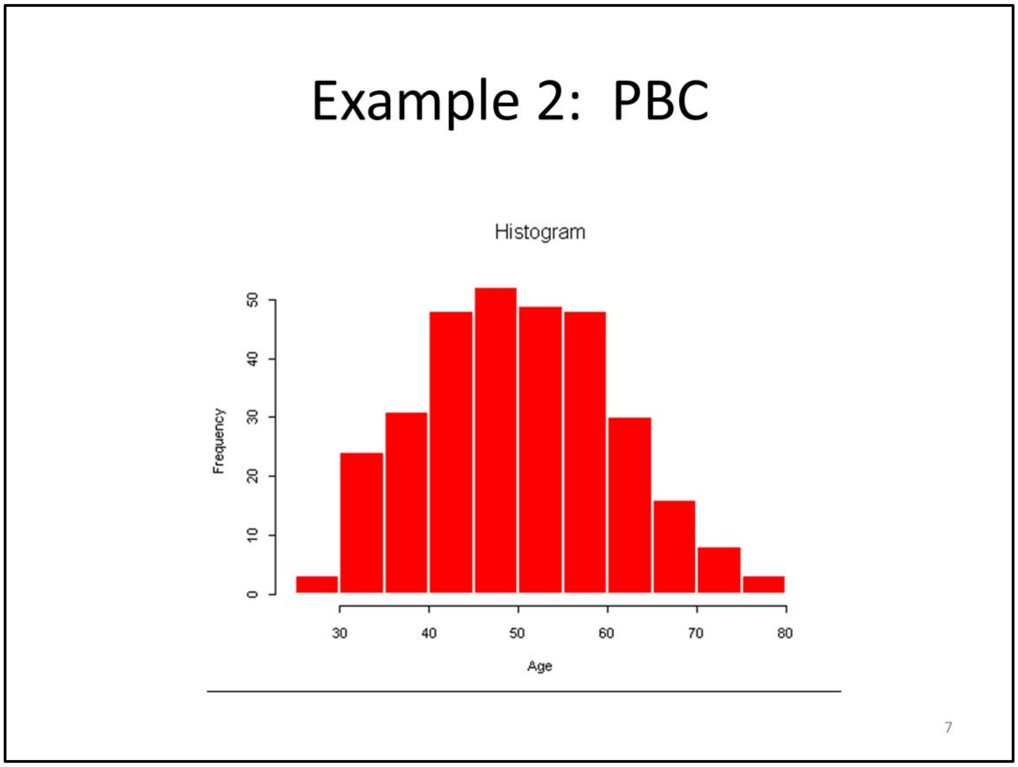
This is an example of a bar plot that is used to summarize the distribution of disease stage. The bars correspond to categories of Stage 1 through 4 and the height of the bar corresponds to the percentage of participants in each category. We see that very few participants were in the Stage 1 group while a higher percentage had Stage 3 or Stage 4 disease.



Histogram

- Goal: Describe the distribution of values for a continuous variable
- Method:
  - Determine intervals of response (bins)
  - For each interval, draw a bar with height equal to the number or proportion of responses

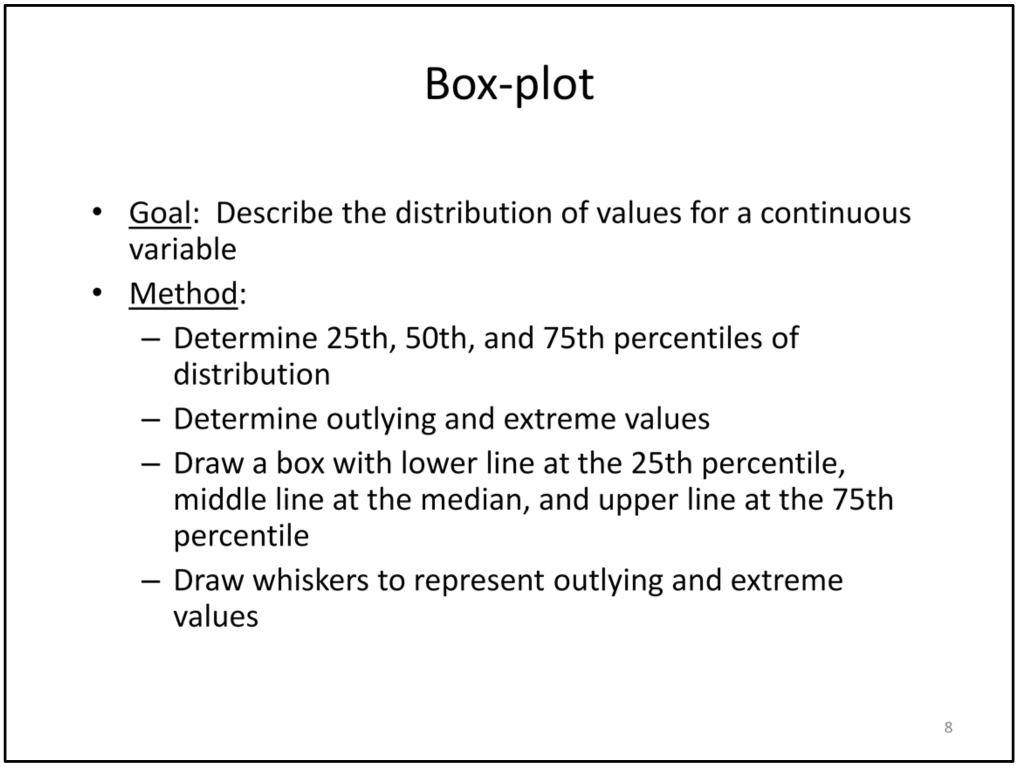
A histogram is similar to a bar graph; however, a histogram is used for summarizing the distribution of continuous measures. It is created by first determining categories of response, such as age groups defined by 10-year increments. Then, a bar is drawn with a height that is equal to the number or percentage of participants in the given category.



Example 2: PBC

7

In this figure, we see a histogram for the age of the participant. The x-axis corresponds to age categories, which are displayed in 5-year intervals and the y-axis corresponds to the number of participants in each category. We see that the age distribution is fairly symmetric and is centered around ages 45-65 years.



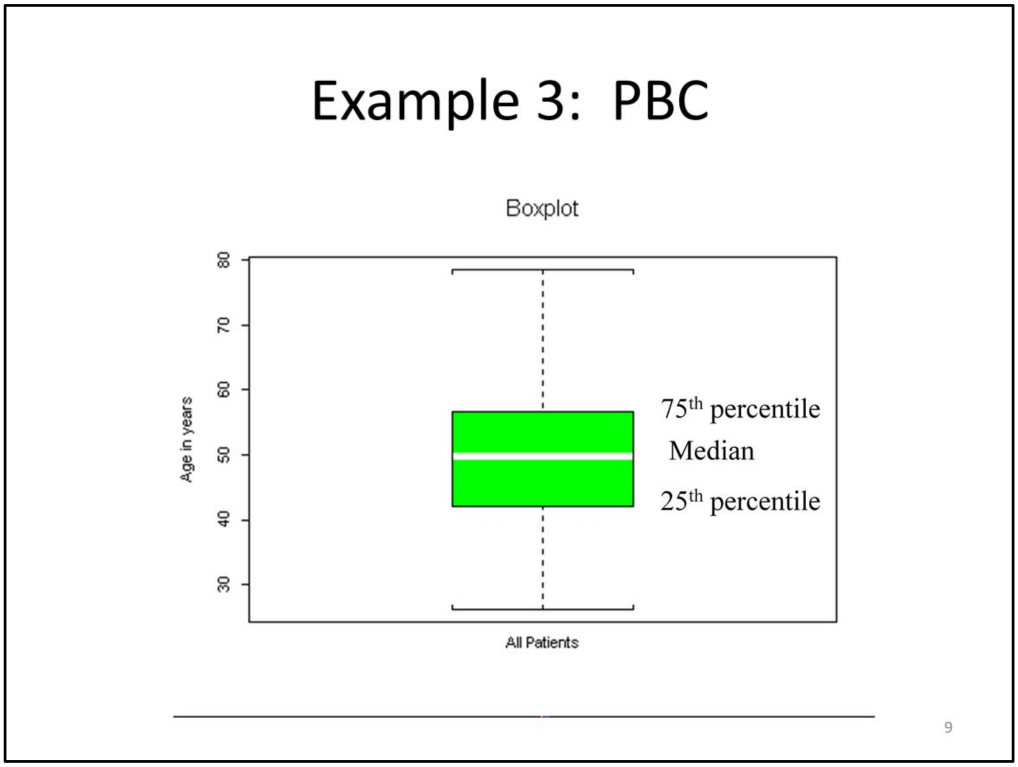
Box-plot

- Goal: Describe the distribution of values for a continuous variable
- Method:
  - Determine 25th, 50th, and 75th percentiles of distribution
  - Determine outlying and extreme values
  - Draw a box with lower line at the 25th percentile, middle line at the median, and upper line at the 75th percentile
  - Draw whiskers to represent outlying and extreme values

8

A box-plot is another plot that can be used to summarize the central tendency and spread of a continuous variable.

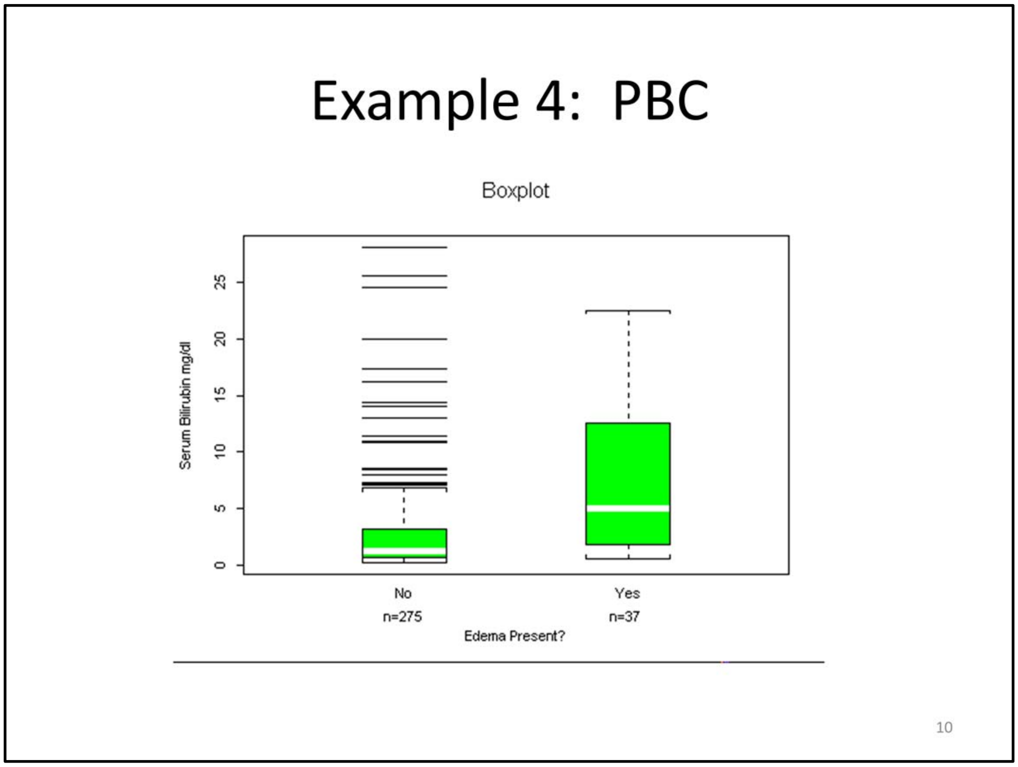
A box-plot is created by first determining the 25th, 50th, and 75th percentiles of the distribution. Then, you determine outlying and extreme values. Next, you draw a box with lower line at the 25th percentile, middle line at the median, and upper line at the 75th percentile. Finally, whiskers are drawn to represent outlying and extreme values.



Example 3: PBC

9

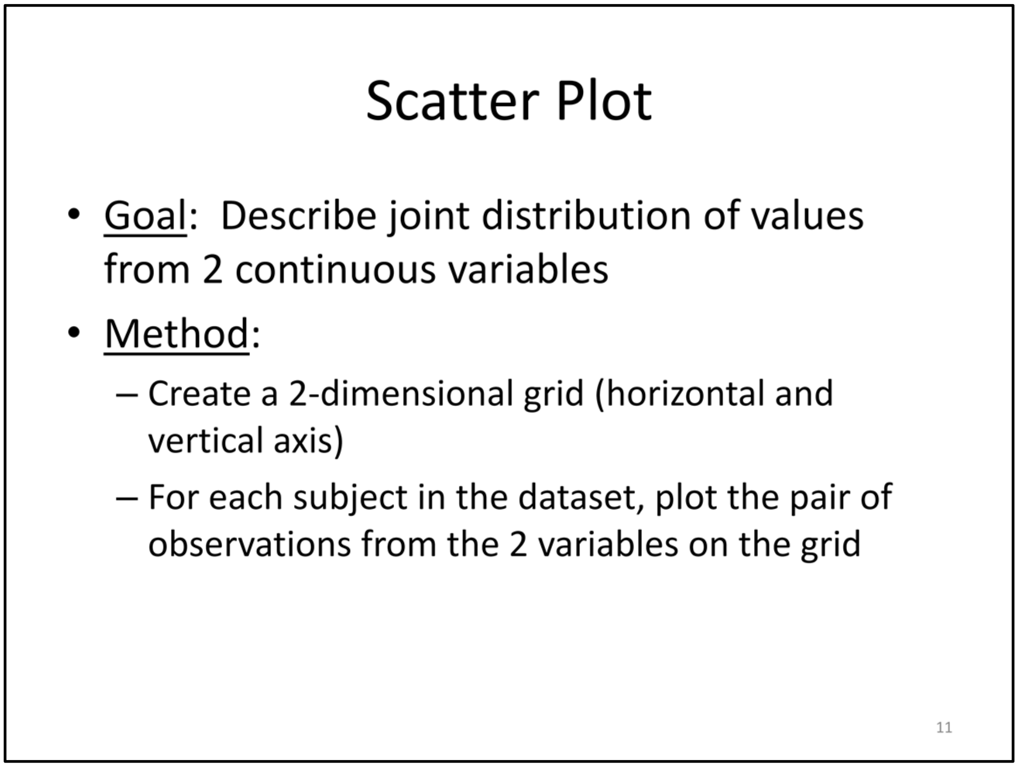
This slide displays a box-plot summarizing the age distribution of patients with primary biliary cirrhosis. We can see, similar to the histogram that we just discussed, the age distribution is fairly symmetric and is centered around a median of 50 years. The interquartile range (the difference between the 25th percentile and the 75th percentile) is from 45 to 55 years. This range encompasses the middle 50% of the distribution. Finally, the whiskers are drawn to +/- 1.5 times the interquartile range or to the maximum and minimum of the distribution if these values fall within +/- 1.5 times the interquartile range.



Example 4: PBC

Box-plots are particularly useful for summarizing the distribution of continuous measures for multiple groups of individuals. The box-plots can be compared side-by-side to understand how the distribution of a continuous measure differs between groups.

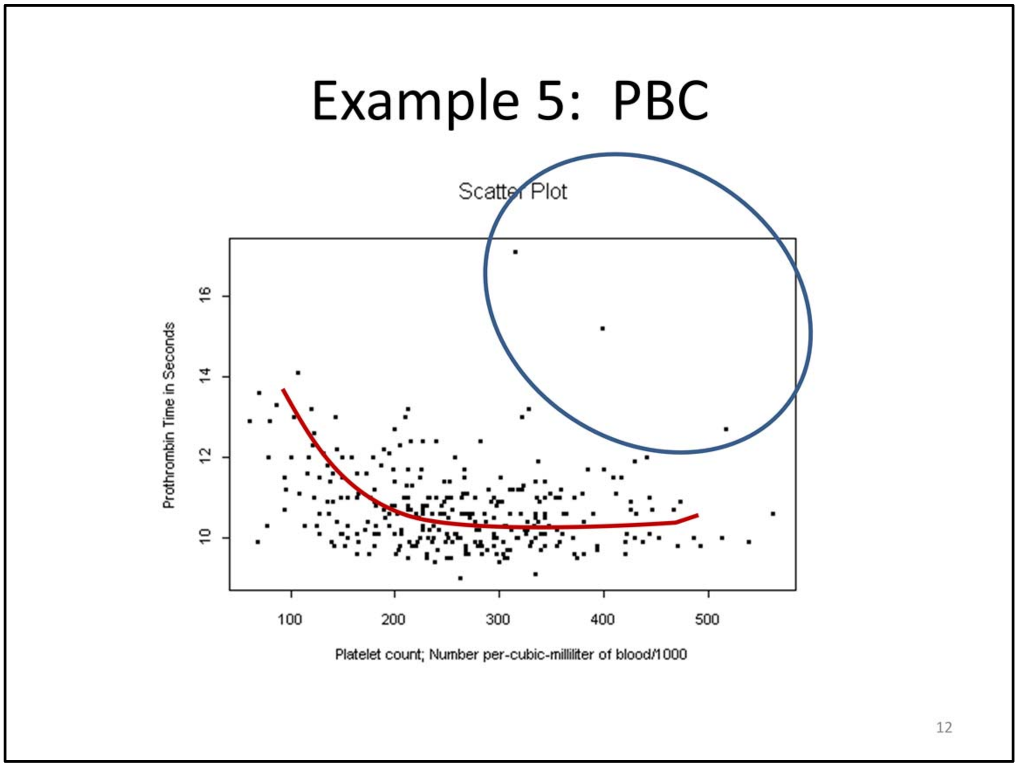
In this example, we are comparing the distribution of serum bilirubin levels (displayed on the y-axis) between patients with and without edema present. We see that the median and range of the 25th to 75th percentiles are lower for patients without edema present, while bilirubin levels are higher among patients with edema present. We can also see that there are patients without edema who have very high serum bilirubin levels.



Scatter Plot

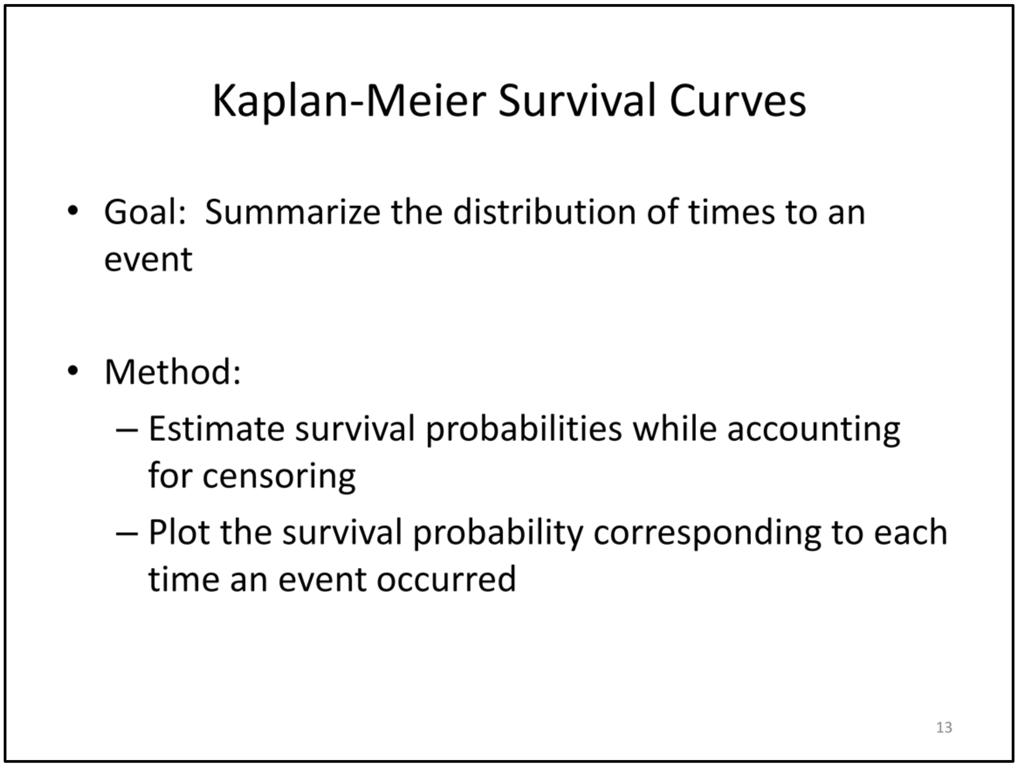
- Goal: Describe joint distribution of values from 2 continuous variables
- Method:
  - Create a 2-dimensional grid (horizontal and vertical axis)
  - For each subject in the dataset, plot the pair of observations from the 2 variables on the grid

The next plot that we will discuss is the scatter plot. A scatter plot is useful for investigating the association between two continuous measures. It is created by first drawing a 2-dimensional grid (horizontal and vertical axis) and then for each subject in the dataset, plot the pair of observations from the 2 variables on the grid. The scatter of points provides a summary of the joint distribution of the two variables and also allows us to identify outlying points that do not fall within the main scatter of points.



Example 5: PBC

In this figure, we summarize the association between prothrombin time (y-axis) and platelet count (x-axis). We see that in general, prothrombin time decreases as platelet count increases. It is interesting to note that after a certain platelet count level, the prothrombin time is essentially stable. The figure can also be used to highlight several extreme measures corresponding to patients with high platelet counts and high prothrombin times. We would want to investigate these points to ensure that a data entry or data measurement error did not occur.

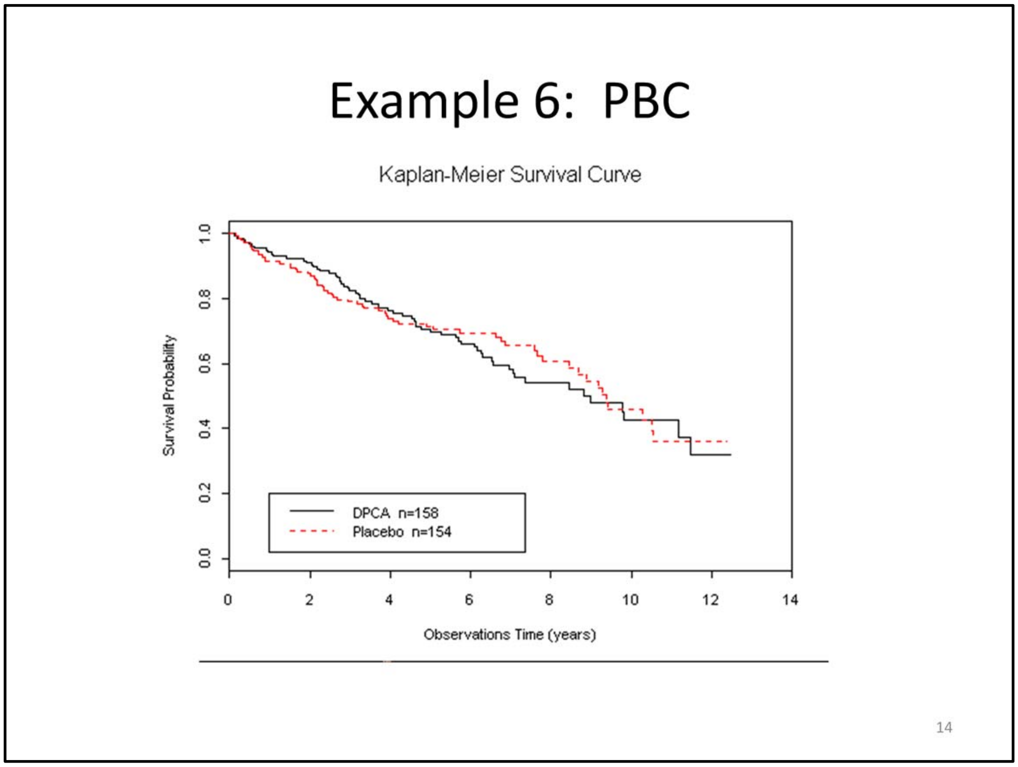


Kaplan-Meier Survival Curves

- Goal: Summarize the distribution of times to an event
- Method:
  - Estimate survival probabilities while accounting for censoring
  - Plot the survival probability corresponding to each time an event occurred

The next series of plots are Kaplan-Meier survival curves. They can be used to summarizing the distribution of times to an event, such as death or tumor recurrence.

To create a Kaplan-Meier curve, we first estimate survival probabilities while accounting for censoring (patients who do not develop the event of interest during the follow-up period) and then we plot the survival probability corresponding to each time an event occurred where the y-axis represents the survival probability and the x-axis represents the follow-up time.



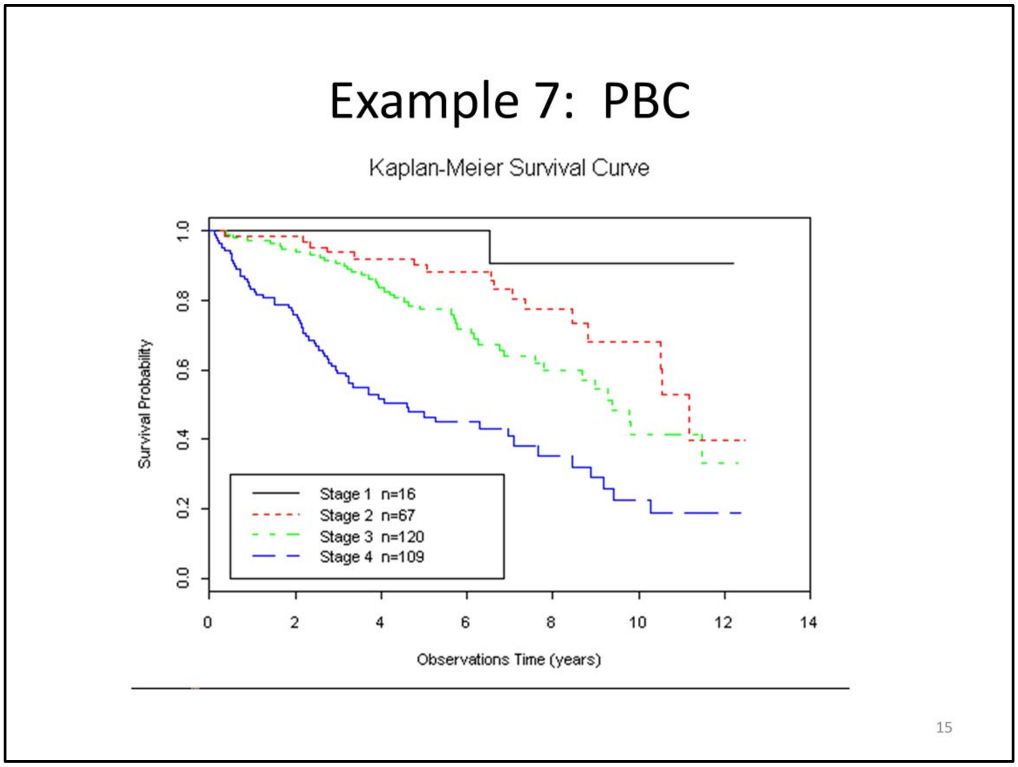
Example 6: PBC

14

In this figure, we have summarized the survival time among patients who received DPCA and patients who received placebo. The y-axis corresponds to survival probability and the x-axis corresponds to follow-up time, expressed in years.

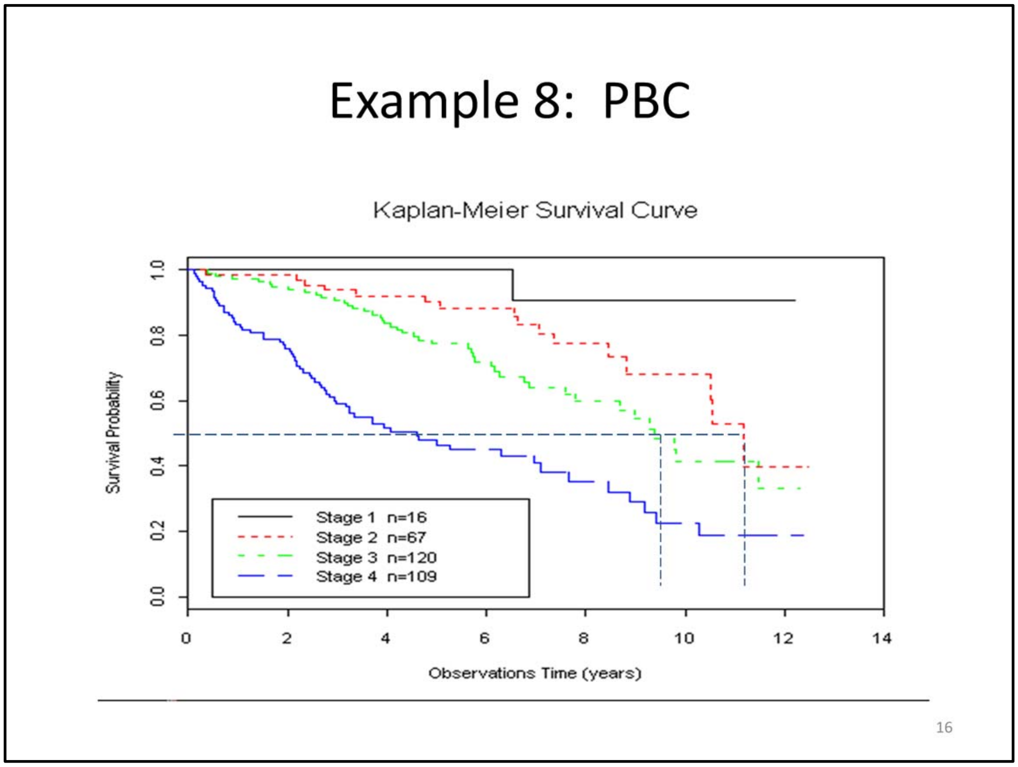
To orient you to the figure, we see that at time 0, the survival probabilities are 100%. Then, with each observed death, the estimated survival probability decreases. Curves that lie in the upper right quadrant correspond to better survival while curves that lie in the lower left quadrant correspond to poorer survival.

In this case, both groups of patients demonstrate similar survival outcomes as demonstrated by the overlapping curves.



Example 7: PBC

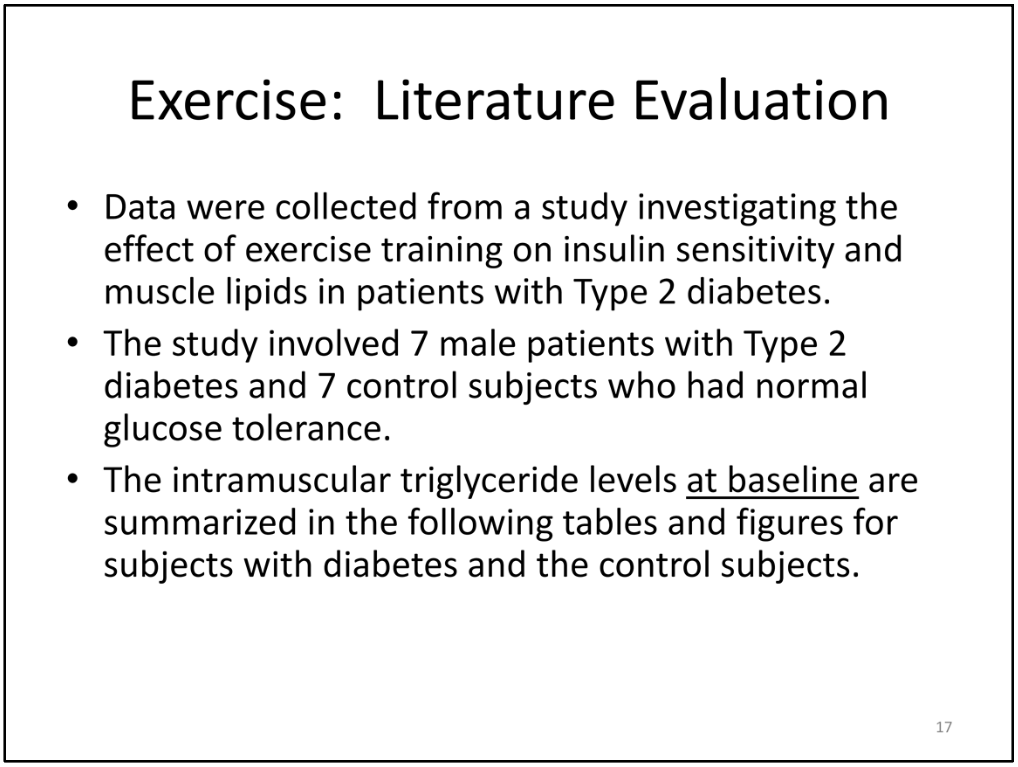
In this figure, we summarize the survival distributions among groups of patients categorized by stage of disease. We see that survival is poorer for patients with a higher stage of disease. This is demonstrated by the survival curves that drop off to the lower left as time increases, meaning, more patients are dying and fewer are surviving.



Example 8: PBC

16

The Kaplan-Meier curve can be used to read off summary statistics for the survival of patients. For example, we can use the figure to determine the median survival time, the time at which half of the patients are still alive, for patients in each stage group. The median survival time is 4 years for patients with Stage 4 disease, 9.5 years for Patients with Stage 3 disease and 11 years for patients with Stage 2 disease. The median survival could not be estimated for patients with Stage 1 disease because 90% are still alive at the end of the follow-up period.



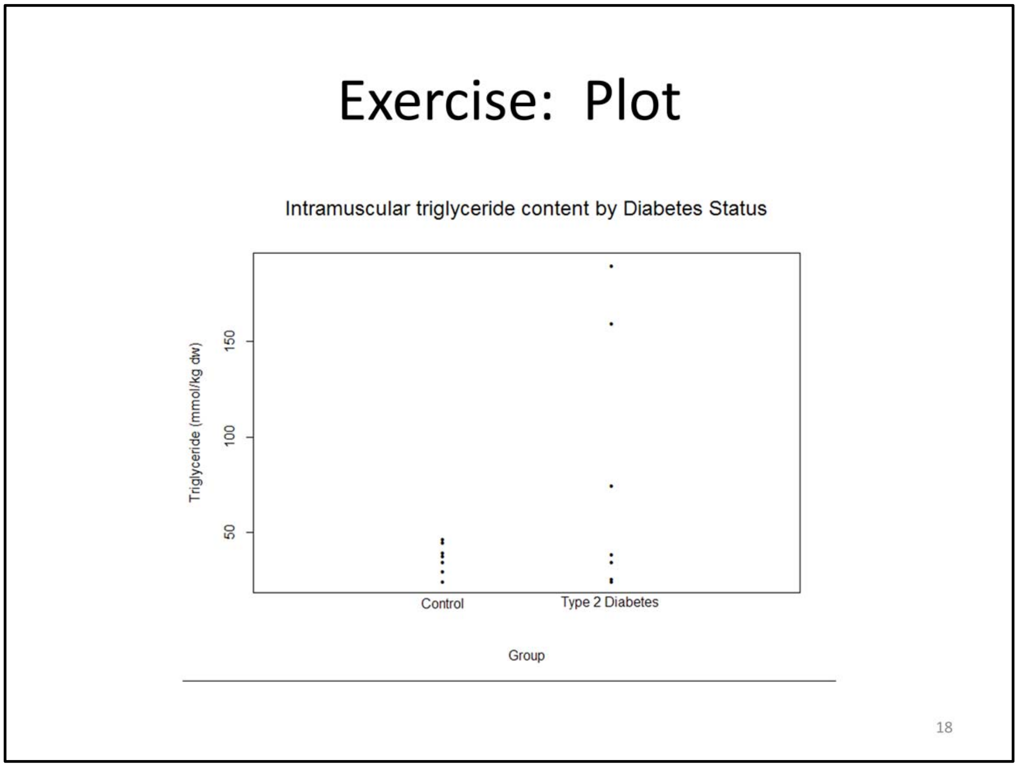
Exercise: Literature Evaluation

- Data were collected from a study investigating the effect of exercise training on insulin sensitivity and muscle lipids in patients with Type 2 diabetes.
- The study involved 7 male patients with Type 2 diabetes and 7 control subjects who had normal glucose tolerance.
- The intramuscular triglyceride levels <u>at baseline</u> are summarized in the following tables and figures for subjects with diabetes and the control subjects.

17

Let's apply information that we have learned regarding descriptive statistics and plots to an example from the literature.

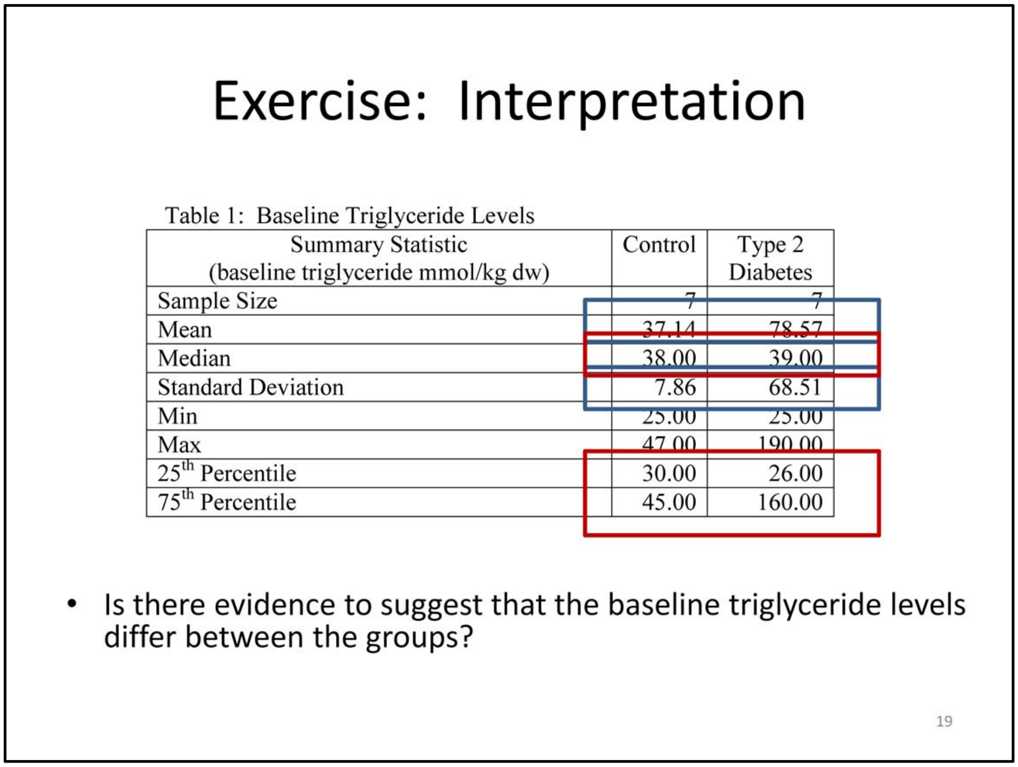
In this study, data were collected to investigate the effect of exercise training on insulin sensitivity and muscle lipids in patients with Type 2 diabetes. The study involved 7 male patients with Type 2 diabetes and 7 control subjects who had normal glucose tolerance. The intramuscular triglyceride levels <u>at baseline</u> are summarized in the following tables and figures for subjects with diabetes and the control subjects.



Exercise: Plot

18

Given that we have very few points, instead of drawing a box-plot of the distribution, we can indicate each individual point on the figure. We see that triglyceride levels (summarized on the y-axis) are lower for control patients without diabetes while triglyceride levels are much more variable, and higher in some cases, for patients with Type 2 diabetes. We need to keep in mind both the location and spread of the distribution.



Exercise: Interpretation

Table 1: Baseline Triglyceride Levels

Summary Statistic (baseline triglyceride mmol/kg dw)	Control	Type 2 Diabetes
Sample Size	7	7
Mean	37.14	78.57
Median	38.00	39.00
Standard Deviation	7.86	68.51
Min	25.00	25.00
Max	47.00	190.00
25th Percentile	30.00	26.00
75th Percentile	45.00	160.00

- Is there evidence to suggest that the baseline triglyceride levels differ between the groups?

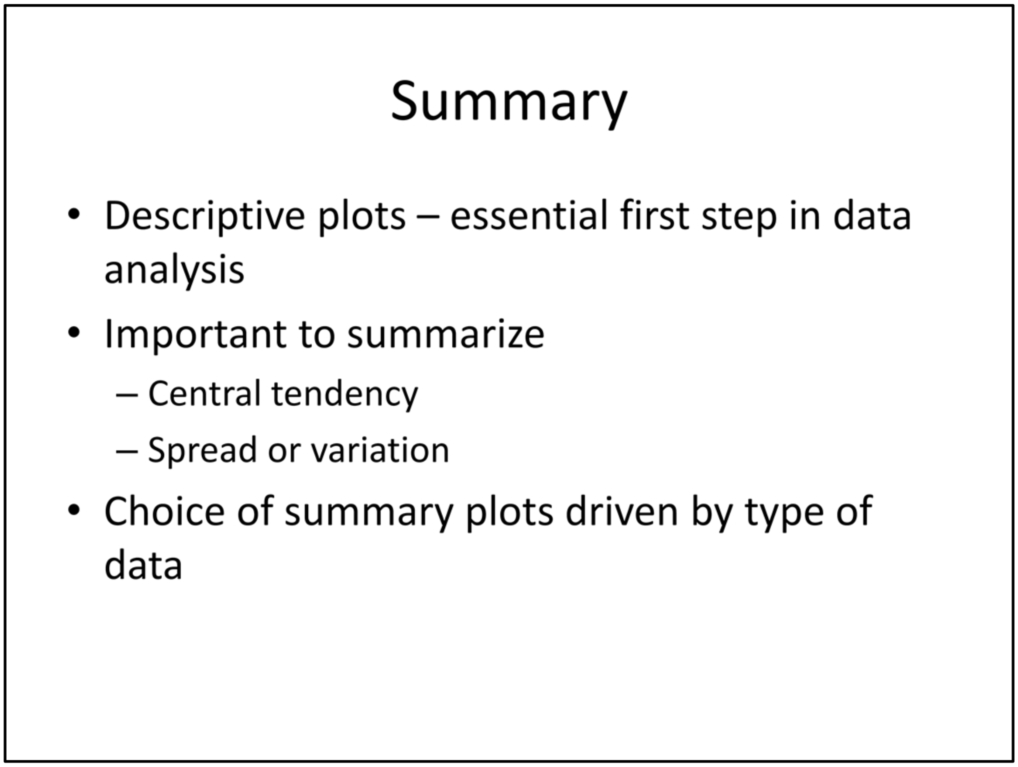
In this slide, we see the descriptive statistics for triglyceride levels for control and patients with Type 2 diabetes.

Based on the figure and the summary statistics, is there evidence to suggest that the baseline triglyceride levels differ between the groups?

If we focus on the mean as a measure of central tendency, we see that the mean triglyceride levels is higher among patients with Type 2 diabetes; however, we note that the standard deviation is also much higher. Given the small sample size in each group, we are concerned that the extreme triglyceride measures in the patients with Type 2 diabetes may have undue influence on the mean and standard deviation.

When we focus on the median and interquartile range, we see that the summary measures are more similar between groups, with the exception of the 75th percentile, which is much higher among patients with Type 2 diabetes.

In this case, while there is a suggestion of increased triglyceride levels among patients with Type 2 diabetes, we would need to collect data on more patients in order to draw conclusions regarding this association.



Summary

- Descriptive plots – essential first step in data analysis
- Important to summarize
  - Central tendency
  - Spread or variation
- Choice of summary plots driven by type of data

In summary, we see that descriptive plots are an essential first step in data analysis.

It is important to summarize both the central tendency and the spread of the data.

Finally, the choice of summary plot is driven by the type of data.

In the next module, we will learn more about measures of disease morbidity.