Graphical representation

(STATS201.stat 202 10: Experimental design and descriptive statistics)

Anna Helga Jónsdóttir
Sigrún Helga Lund

December 16, 2012

Graphical representation

- The first step in any statistical analysis should be to look at the data visually.
- The essence of statistical analysis is to understand the nature of the measurements that are investigated.
- The variability of the data is a key issue - how are the measurements distributed?
  - How much difference do we notice in the outcomes of our subjects?
  - How are the outcomes distributed?
- Graphical representation is the best way to understand the nature of the distribution of the measurements.
- The graphs that we will see in this lecture will only show the measurements of one variable at a time and we make a distinction whether the variable is continuous or discrete.

Graphical representation of discrete variables

- The most common form of graphs for discrete variables are the **bar chart** and the **pie chart**.
- Pie charts are commonly used in business and media but are rare in journals and books in natural sciences.
- Bar charts are commonly seen and they are better suited then pie charts to show the measurements of discrete variables. than pie charts.

Bar chart

Bar chart

A bar chart consists of two or more bars. The number of bars is determined by the number of categories/values that the discrete variable takes. Every bar represents one category/value and they may not lie close to each other. The height of the bars shows the frequency of the corresponding category. The bars shall be ordered in an informative way, often by size.

Before one draws a bar chart it is often convenient to make a little table that shows the categories of the variable and how many subjects belong to each category.

Pie chart

Pie chart

When making a pie chart, it is important that all categories/values of the variable under investigation are pictured on the chart. The number of slices in the pie chart is determined by the number of categories/values of the variable. The size of the slice is determined by the proportional number of subjects in the corresponding category compared to the whole sample. Watch out that the ratios add up to 100 %.

Before one draws a bar chart it is often convenient to make a little table that shows the categories of the variable, how many subjects belong to each category and the corresponding percentage of subjects in that category as a proportion of the whole sample.

PIE CHARTS ARE NOT DRAWN BY HAND!

Graphical representation of continuous variables

- The most common method to visualize continuous variables are **histograms**.
- A **Box plot** is also a good method to visualize continuous variables and they will be introduced in the lecture about descriptive statistics (Lecture 40).
- **Scatter plots** will be introduced in the lecture about linear regression (Lecture170) but they are used to explore the relationship between two continuous variables.

Histograms

- Histograms are similar to bar charts but the main difference in their appearance is that there is no gap between the bars in a histogram.
- It is slightly more difficult to make a histogram then a bar chart as continuous variables do not contain real categories or groups.
- First one has to define groups (intervals) before one counts how many measurements belong to each group.
- When the groups have been made it is often useful to make a table that contains the groups and how many measurements lie within each group.

Histograms

Histogram

A histogram consists of bars that are lined continuously one by another. The number of bars is determined by the number of groups (intervals) that the continuous variable is split up into. When the groups are made is is often good to keep in mind the following criteria:

- Lower and upper limits should be simple and easily understood.
- The intervals may not overlap and must cover all measurements.
- The intervals should be equally wide.
- The number of intervals should be appropriate. A rule of thumb is that the number of intervals should be approx.
  5 times the logarithm of the total number of measurements.

When the intervals have been made, one bar is drawn for each interval and the height of the bar is determined by the number of measurements within the corresponding interval.

Shapes of distributions

Shapes of distributions

The following concepts are used to describe the distribution of measurements.

- The distribution of the smallest measurements are called the *left-tail* of the distribution. The distribution of the largest measurements is called the *right-tail* of the distribution.
- A distribution is *symmetric* if its right-tail is distributed as the mirror image of the left-tail.
- A distribution that is not symmetric is *skewed*. A distribution is *skewed to the right* if its right-tail is longer then the left-tail and *skewed to the left* if the left one is longer then the right one.
- If a distribution has one peak it is referred to as *unimodal*.
- If a distribution has two peaks it is referred to as *bimodal*.
- If a distribution has more then two peaks it is referred to as *multimodal*.

Outliers

Outliers

Outliers are measurements that differ greatly from other measurements in the sample. There can be various reasons for outliers and it is important to look at them specifically and consider their cause.