University of California, Los Angeles
Department of Statistics

Nicolas Christou
Ivo Dinov

Measures of central tendency and variation
Data display

- **Measures of central tendency**

1. Sample mean:
   Let $x_1, x_2, \cdots, x_n$ be the n observations of a sample. The sample mean $\bar{x}$ is computed as follows:
   $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$
2. Median: It is the value that falls in the middle when the observations are sorted from smallest to largest.
   To compute the median, follow the next 2 steps:
   a. Sort the observations from smallest to largest.
   b. Compute the position of the median: $\frac{n+1}{2}$.

Examples:
A. Sample size n is odd:
7 annual incomes: 28, 60, 26, 32, 30, 26, 29. First sort these observations from smallest to largest:
26, 26, 28, 29, 30, 32, 60
Next compute $\frac{n+1}{2} = \frac{7+1}{2} = 4_{th}$. The median is the 4_{th} observation. Median=29.

B. Sample size n is even:
8 annual incomes: 26, 26, 28, 29, 30, 32, 60, 80
Again compute $\frac{n+1}{2} = \frac{8+1}{2} = 4.5_{th}$. The median is the average of the two middle observations. Median= $\frac{29+30}{2} = 29.5$.

Question: How do unusual observations affect the sample mean and the median? Example: 8 annual incomes:
26, 26, 28, 29, 30, 32, 60, 8000

- **Measures of non-central tendency**

  1. First quartile (Q_1) or 25_{th} percentile: Its position is $\frac{n+1}{4}$.
  2. Third quartile (Q_3) or 75_{th} percentile: Its position is $\frac{3(n+1)}{4}$.

  Example:
  Find Q_1 and Q_3 of the following 8 annual incomes:
  26, 26, 28, 29, 30, 32, 60, 80
  Position of Q_1: $\frac{n+1}{4} = \frac{8+1}{4} = 2.25_{th} \approx 2_{nd}$ (round to the nearest integer).
  Position of Q_3: $\frac{3(n+1)}{4} = \frac{3(8+1)}{4} = 6.75_{th} \approx 7_{th}$ (round to the nearest integer).
  Therefore, $Q_1 = 26, Q_3 = 60$.

  Five-number summary of a data set:
  $MIN \qquad Q_1 \qquad MEDIAN \qquad Q_3 \qquad MAX$

  Box plot:
  A popular way to display data and identify outliers. You are given 11 annual incomes in thousands of dollars: 26, 26, 28, 29, 30, 32, 60, 65, 70, 40, 44. Construct the boxplot of income using these 11 observations.

  Begin by sorting these incomes: 26, 26, 28, 29, 30, 32, 40, 44, 60, 65, 70

  Find the position of the first quartile, median, and third quartile:

Position of Q_1	$\frac{n+1}{4} = \frac{11+1}{4} = 3_{rd}$
Position of Median	$\frac{n+1}{2} = \frac{11+1}{2} = 6_{th}$
Position of Q_3	$3\frac{n+1}{4} = 3\frac{11+1}{4} = 9_{th}$

  Find the first quartile, median, and third quartile: $Q_1 = 28$, $Median = 32$, $Q_3 = 60$ and the interquartile range is $IQR = Q_3 - Q_1 = 60 - 28 = 32$.

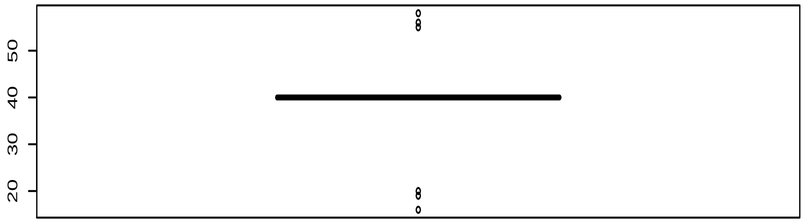
  Outliers are observations above $Q_3 + 1.5IQR$ or below $Q_1 - 1.5IQR$. Also, serious outliers are observations above $Q_3 + 3IQR$ or below $Q_1 - 3IQR$. In our example we do not have any outliers since $Q_3 + 1.5IQR = 60 + 1.5(32) = 108$ and $Q_1 - 1.5IQR = 28 - 1.5(32) = -20$. Now we can construct the box plot.

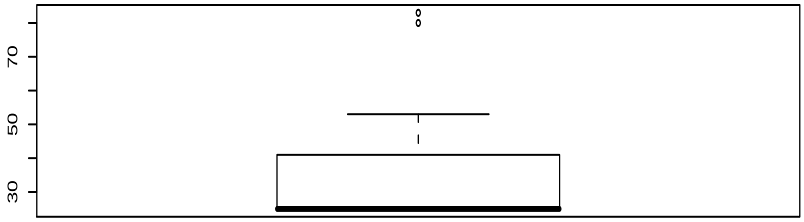
  See SOCR activities and applet on box plots:

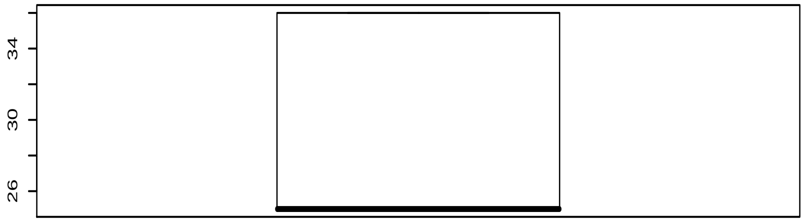
  `http://socr.ucla.edu/htmls/chart/BoxAndWhiskersChartDemo3_Chart.html`

  `http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ChartsActivities`

Box plot pathologies:
Here are some interesting box plots. Can you write down a set of observations that correspond to these box plots?







- **Measures of variation**

  1. Range:

  2. Interquartile range (IQR):

  3. Sample variance and sample standard deviation.
     Let $x_1, x_2, \cdots, x_n$ be the n values of a sample. The sample variance s^2 is the average of the squared deviations of each observation from the sample mean and it is computed as follows:
     $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$
     where $x_i - \bar{x}$ is the i_{th} deviation from the sample mean $\bar{x}$.
     It is easier for calculations to use:
     $$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]$$
     The standard deviation is simply the square root of the variance. Both $\bar{x}$ and s have the same units.
     $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$
     or easier for calculations
     $$s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]}$$
     Note:
     $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
     $\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i\right)^2$.

  Example:
  Find the sample mean $\bar{x}$, sample variance s^2, and sample standard deviation s of the following sample: 1, 1.1, 0.9, 1.3, 0.7 (weights of five oranges in ounces).

- **Adding and multiplying observations by a constant**

  Let $x_1, x_2, \cdots, x_n$ be the observations of a sample of size n, and let $\bar{x}$ and s^2 be the sample mean and sample variance respectively.

  a. Suppose that on each observation a constant a is added. Find the new sample mean and sample variance.

  b. Suppose that each observation is multiplied by a constant a. Find the new sample mean and sample variance.

  c. Examples:

     1. The weight of 5 water melons in pounds are: 3.3, 2.9, 2.5, 3.6, 3.0. Find the sample mean and sample standard deviation of these five water melons in pounds. Then find the sample mean and standard deviation in kilograms.

     2. In the the U.S. temperature is recorded in Farenheit degrees, while in most of the other countries it is recorded in Celcius degrees. Suppose a tourist from a country where temperature is recorded in Celcius degrees will visit Los Angeles in the summer. He was told that the July average in Los Angeles is 85 Farenheit degrees, with a standard deviation of 10 Farenheit degrees. Help this tourist understand the weather conditions in Los Angeles.

Data display

Three popular methods:

1. Dot plot
2. Frequency distribution
3. Histogram
4. Box plot

See SOCR activities:

`http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ChartsActivities`

- Dot plot:
  Simply place a dot for each observation in the data set.
- Frequency distribution:
  We can group data into classes (bins). The first step is to define the number of classes and the width of each class (define the number of bins). There many ways to do this.
- Histogram:
  The frequency distribution can be graphed. The graph is called histogram. To construct a histogram: On the horizontal axis place the class limits. Then construct a rectangle which has base the width of the class and height the frequency of that class. There is also a relative frequency histogram (the height of each rectangle is the the relative frequency of that class).
- Box plot:
  See page 2 of this handout.

Examples:

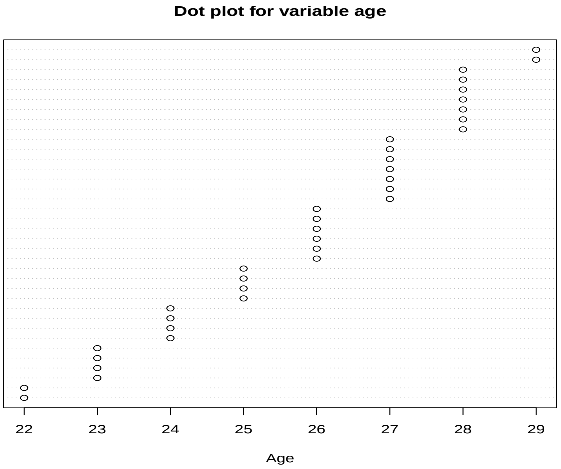
a. Body fat data. You can access the data at:

```
a1 <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/body_fat.txt", header=TRUE)
```

Here we only consider the variable age (`x3`) for ages less than 30:

```
 [1] 22 22 23 23 23 23 24 24 24 24 25 25 25 25 26 26 26 26 26
[20] 26 27 27 27 27 27 27 27 28 28 28 28 28 28 28 29 29
```

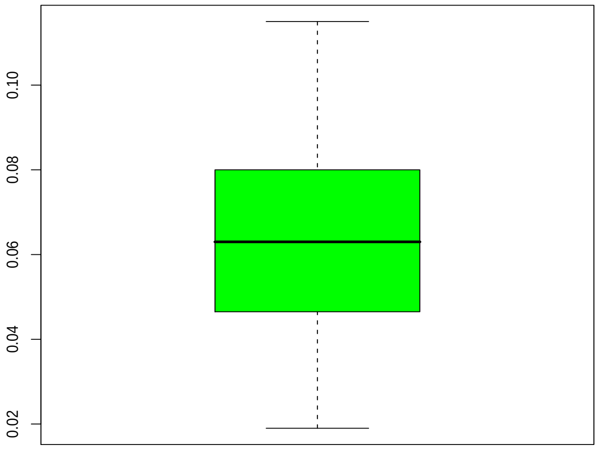
And the dot plot:



b. Box plot of variable ozone. Access the data here:

```
a2 <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/ozone.txt", header=TRUE)
```

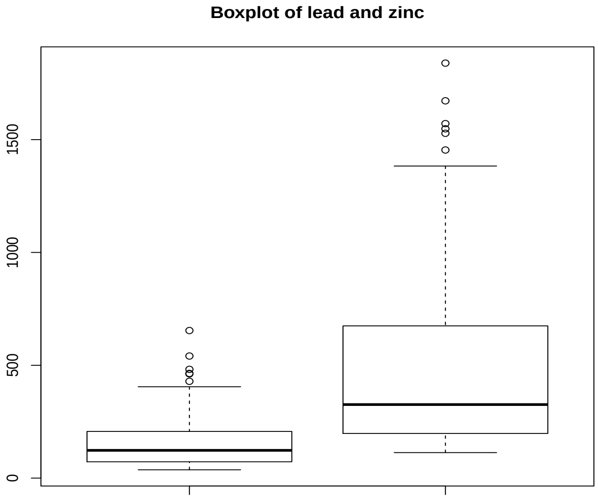
Box plot of ozone:



Side-by-side box plot of lead and zinc. Access the data here:

```
a22 <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/soil.txt", header=TRUE)
```

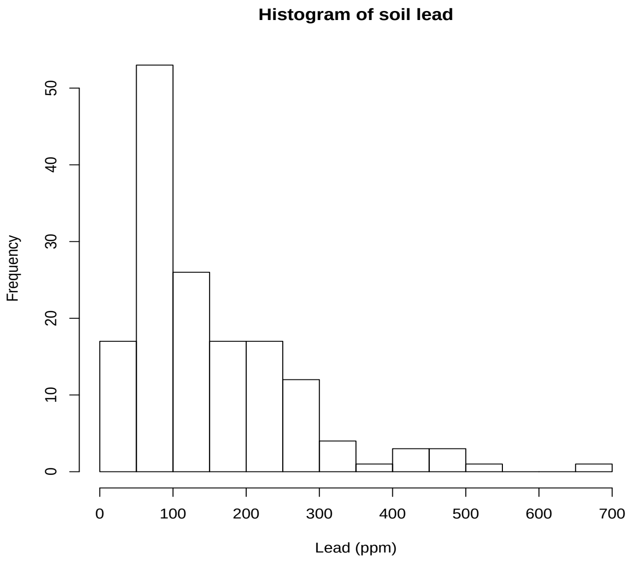
Box plot of ozone:



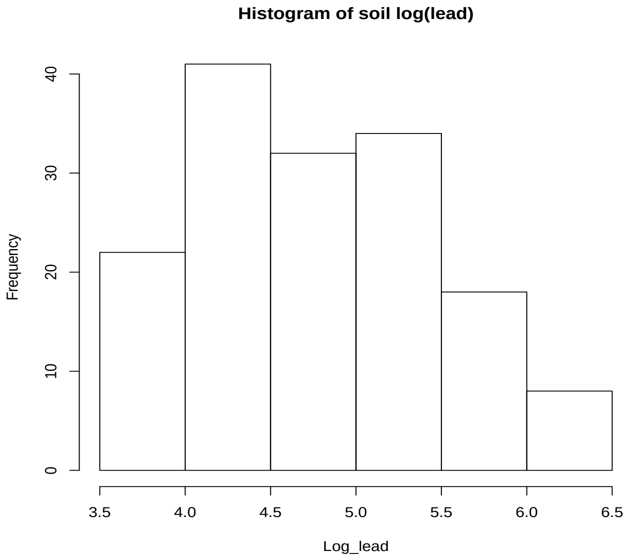
c. Soil lead and zinc data (area of interest in the Netherlands). You can access these data at:

```
a3 <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/soil.txt", header=TRUE)
```

Histogram of `lead`

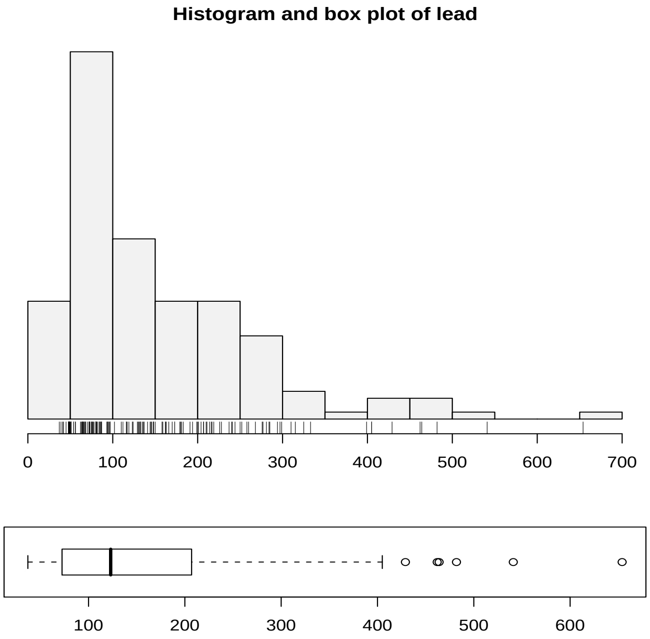


Histogram of `log(lead)`

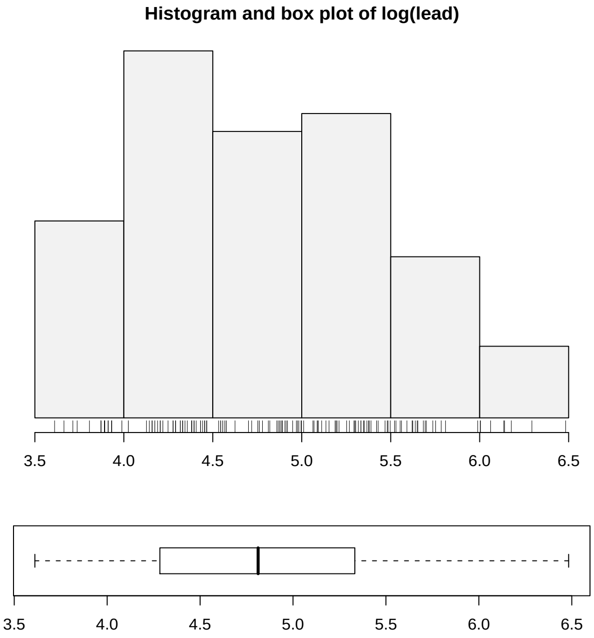


We can plot the histogram and the box plot on the same graph as shown below. The first graph shows a skewed to the right histogram. The second graph shows a more symmetrical histogram. We observe that the corresponding box plots agree with the histograms.

Histogram and box plot of `lead`



Histogram and box plot of `log(lead)`



Categorical data

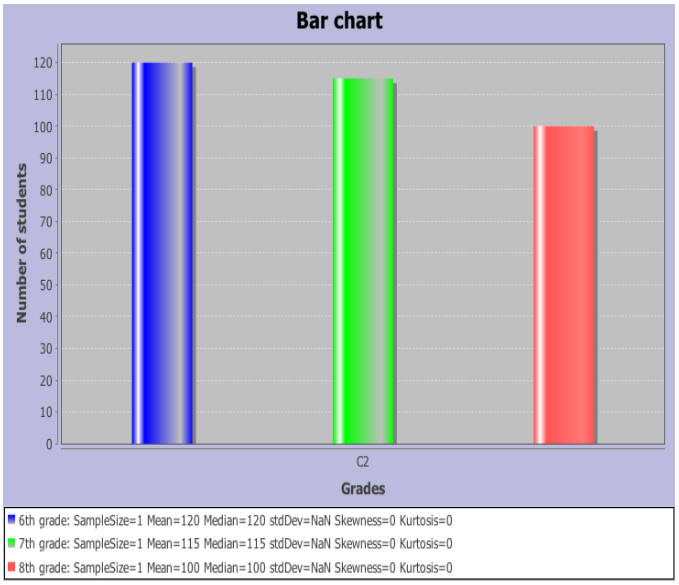
Many experiments result in count data. These data are counted and placed into categories (categorical data), also called qualitative data. A bar chart, or a pie chart will be appropriate for categorical data. For example, see tables below:

SOCR activities:

`http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_PieChart`

Number of students per grade:

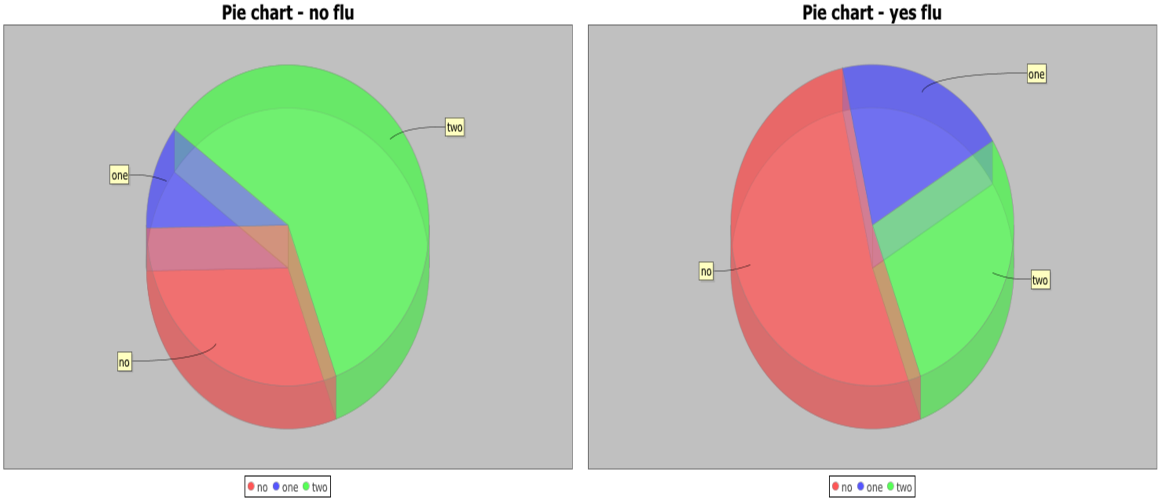
```
6th 7th 8th
120 115 100
```



There may be two characteristics (contingency table). Number of flu vaccinations and whether person had the flu. We want to explore if having the flu is independent of the number of flu vaccinations.

```
   Yes No
no  24 289
one  9 100
two 13 565
```

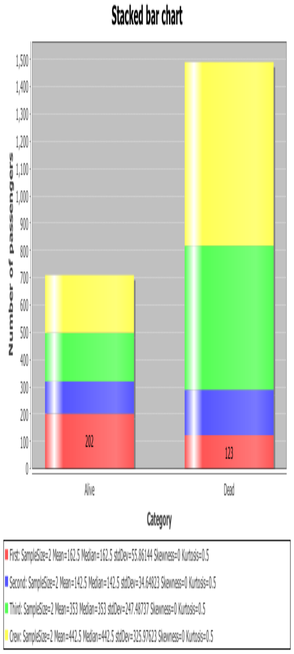
Pie charts using the flu data:



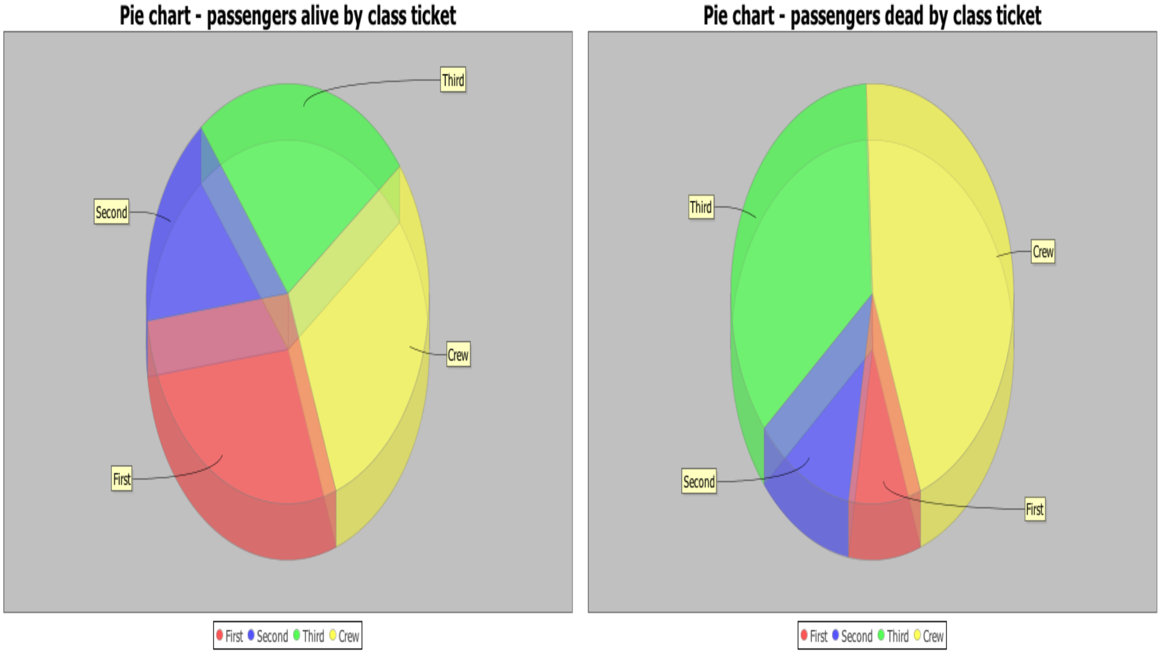
Contingency table using Titanic data (surviving/class):

	First	Second	Third	Crew	Total
Alive	202	118	178	212	710
Dead	123	167	528	673	1491
Total	325	285	706	885	2201

Stacked bar chart using the Titanic data:



Pie charts using the Titanic data:


Pie chart - passengers alive by class ticket
Third
Second
Crew
First
● First ● Second ● Third ● Crew
Pie chart - passengers dead by class ticket
Third
Crew
Second
First
● First ● Second ● Third ● Crew