University of California, Los Angeles
Department of Statistics

Statistics 13 **Instructor: Nicolas Christou**

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 Q_1 $MEDIAN$ Q_3 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

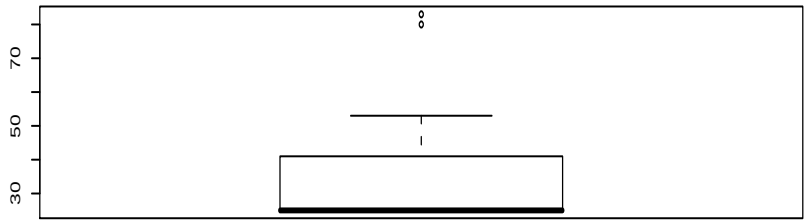
  Find the position of the first quartile, median, and third quartile:

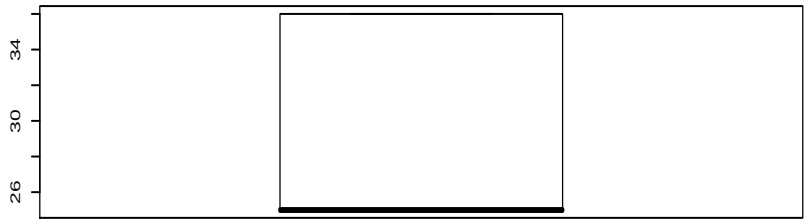
Position of Q_1	$\frac{n+1}{4} = \frac{11+1}{4} = 3_{rd}$
Position of Median	$\frac{n+1}{2} = \frac{11+1}{2} = 6_{th}$
Position of Q_3	$3\frac{n+1}{4} = 3\frac{11+1}{4} = 9_{th}$

  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.

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:
     $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
     $\sum_{i=1}^{n} x_i^2 \neq (\sum_{i=1}^{n} x_i)^2$.

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.

Data display

Three popular methods:

1. Stem-and-leaf display
2. Frequency distribution
3. Histogram

- Stem-and-leaf display:
  Split each observation into a "stem" and "leaf". Then place the stems in a column from smallest to largest. Next to each stem place the leaves from smallest to largest.
- 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).

Construct by hand the stem and leaf plot of the following observations (ozone data `ppm`):

```
 [1] 0.044 0.081 0.035 0.080 0.053 0.077 0.051 0.059 0.041 0.027 0.090 0.069 0.057
[14] 0.029 0.052 0.083 0.068 0.078 0.096 0.019 0.065 0.061 0.094 0.035 0.097 0.057
[27] 0.036 0.060 0.032 0.036
```

See more examples on the next pages.

a. California ozone data. You can access the data at:

`http://www.stat.ucla.edu/~nchristo/statistics13/ozone.txt`

Here are the data:

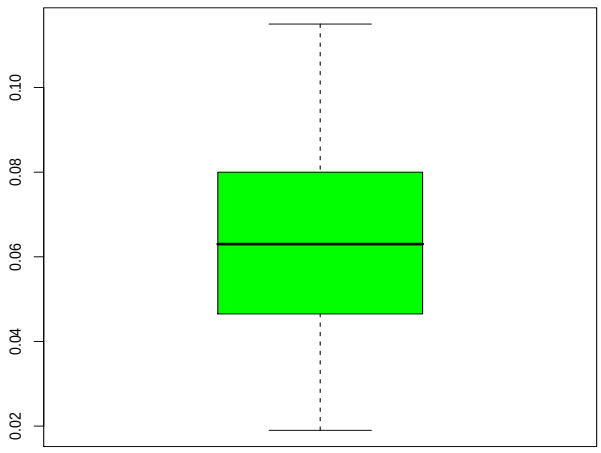
```
  [1] 0.044 0.081 0.035 0.080 0.053 0.077 0.051 0.059 0.041 0.027 0.090 0.069
 [13] 0.057 0.029 0.052 0.083 0.068 0.078 0.096 0.019 0.065 0.061 0.094 0.035
 [25] 0.097 0.057 0.036 0.060 0.032 0.036 0.051 0.029 0.030 0.105 0.047 0.078
 [37] 0.084 0.095 0.079 0.067 0.094 0.081 0.077 0.048 0.052 0.059 0.101 0.038
 [49] 0.028 0.046 0.089 0.033 0.036 0.034 0.078 0.062 0.056 0.085 0.041 0.029
 [61] 0.059 0.115 0.043 0.082 0.094 0.099 0.059 0.089 0.093 0.038 0.099 0.064
 [73] 0.050 0.068 0.079 0.041 0.056 0.094 0.082 0.051 0.071 0.077 0.063 0.063
 [85] 0.061 0.068 0.039 0.061 0.024 0.054 0.082 0.061 0.065 0.036 0.054 0.046
 [97] 0.067 0.073 0.050 0.105 0.029 0.102 0.055 0.053 0.090 0.063 0.055 0.082
[109] 0.041 0.097 0.079 0.097 0.056 0.036 0.078 0.061 0.066 0.092 0.070 0.039
[121] 0.096 0.065 0.043 0.067 0.049 0.086 0.079 0.073 0.081 0.080 0.073 0.043
[133] 0.083 0.080 0.068 0.077 0.077 0.048 0.046 0.066 0.102 0.111 0.079 0.047
[145] 0.037 0.067 0.071 0.072 0.100 0.071 0.038 0.074 0.075 0.035 0.100 0.036
[157] 0.058 0.035 0.049 0.079 0.084 0.112 0.082 0.028 0.111 0.037 0.051 0.044
[169] 0.027 0.053 0.080 0.044 0.059 0.055 0.054
```

And the stem and leaf plot:

```
The decimal point is 2 digit(s) to the left of the |

 1 | 9
 2 | 477889999
 3 | 02345555666667788899
 4 | 1111333444666778899
 5 | 00111122333444555666778999999
 6 | 011111233345566777788889
 7 | 01112333457777788888999999
 8 | 00001112222233445699
 9 | 0023444456677799
10 | 0012255
11 | 1125
```

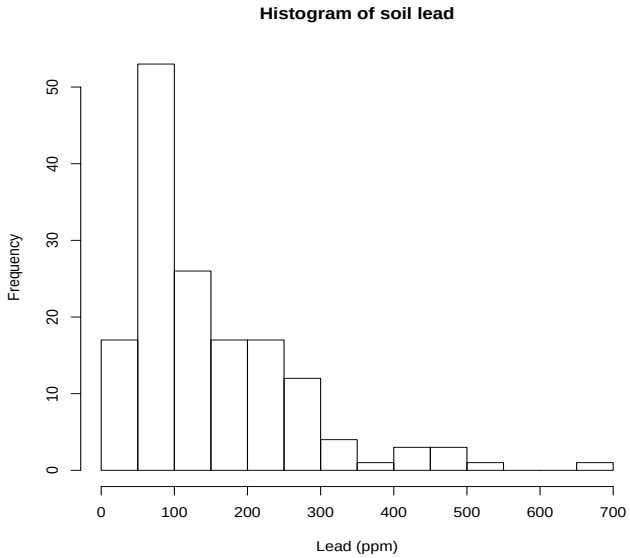
Box plot of ozone:



b. Soil lead and zinc data (area of interest in the Netherlands - see next handout in `R`). You can access these data at:

`http://www.stat.ucla.edu/~nchristo/statistics13/soil.txt`

Histogram of `lead`



Histogram of `log(lead)`