Applied Statistics for Life Sciences

Dmitri D. Pervouchine

Centre de Regulació Genòmica

Module 1: Exploratory Data Analysis

Contents

1. Basic concepts
2. Descriptive statistics
    - Measures of center
    - Measures of spread
3. Summarizing Samples
    - Univariate data
    - Bivariate data
    - Skewness
4. Experiment Planning
5. Sampling
6. Sources of Bias

Sample vs. Population

- The entire group of objects about which information is wanted is called the **population**
- Individual members are called **units**
- A small part of the population is actually available for observation; it is called **sample**
- A **statistic** is a function of the sample

Problem 1.1

Fill in the missing words to the quote: "Statistical methods may be described as methods for drawing conclusions about . . . . . . based on . . . . . . computed from the . . . .

(A) *statistics, samples, populations*

(B) *populations, parameters, samples*

(C) *statistics, parameters, samples*

(D) *parameters, statistics, populations*

(E) *populations, statistics, samples*

Solution

(E) populations, statistics, samples

The types of variables

- **Numeric** variables have values that describe a measurable quantity as a number
  - A **continuous** variable can take any value between a certain set of real numbers
  - A **discrete** variable can take a value from a set of distinct (usually, integer) values
- **Categorical** variables have values that belong to categories
  - An **ordinal** variable has values that can be logically ordered or ranked
  - Values of a **nominal** cannot be logically ordered

Measures of center

- $\{X_1, X_2, \ldots, X_n\}$ is the sample
- Assume the sample is sorted, i.e., $X_1 \leq X_2 \leq X_3 \leq \ldots \leq X_n$
- Statistic $= f(X_1, X_2, \ldots, X_n)$
- Measures of center
  - **Median** is the $50\%^{\mathrm{th}}$ observation; MED $= X_{[\frac{n+1}{2}]}$
  - Note that if $[\frac{n+1}{2}]$ is not an integer, the definition of the median is a matter of convention
  - **Mean** is the center of mass; $\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$
  - **Mode** is the most frequent observation

Measures of center

Problem 2.1

Find the mean, median, and mode for the following sample: $1, 5, 4, 4, 3, 2, 4, 6, -2, 1$

Solution

- *First, sort the sample* $-2, 1, 1, 2, 3, 4, 4, 4, 5, 6$
- $\overline{X} = \frac{-2+1+1+2+3+4+4+4+5+6}{10} = 2.8$
- *The median is the average of* X_5 *and* X_6 *in the sorted sample*
- $MED = \frac{3+4}{2} = 3.5$
- *Mode=4*

Percentiles

- A **percentile** is a measure indicating the value below which a given percentage of observations in a group of observations fall
- $X_1 \leq X_2 \leq X_3 \leq \ldots \leq X_n$ is the sorted sample
- **Upper quartile** is the 75th percentile; $\mathrm{UQ} = X_{[0.75(n+1)]}$
- **Lower quartile** is the 25th percentile; $\mathrm{LQ} = X_{[0.25(n+1)]}$
- Median is the 50th percentile
- *Note that if* $[0.25(n+1)]$ *is not an integer, the definition of UQ and LQ is a matter of convention*

Measures of spread

- $X_1 \le X_2 \le X_3 \le \ldots \le X_n$ is the sample
- **Interquartile range** $IQR = UQ - LQ$
- RANGE $= X_{max} - X_{min}$
- **Variance** $= s_X^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$
- Variance is the moment of inertia
- **Standard deviation** $s_X = \sqrt{Variance}$

One important derivation for sample variance

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 =$$

$$\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} =$$

$$\frac{(X_1^2 - 2X_1\overline{X} + \overline{X}^2) + (X_2^2 - 2X_2\overline{X} + \overline{X}^2) + \cdots + (X_n^2 - 2X_n\overline{X} + \overline{X}^2)}{n-1}$$

$$\frac{(X_1^2 + X_2^2 + \cdots + X_n^2) - 2\overline{X}(X_1 + X_2 + \cdots + X_n) + n\overline{X}^2}{n-1} =$$

$$\frac{(X_1^2 + X_2^2 + \cdots + X_n^2) - 2n\overline{X}^2 + n\overline{X}^2}{n-1} =$$

$$\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$$

Outliers

- An **outlier** is an observation point that is distant from other observations
- An outlier is *defined to be* the value in the sample that differs from the nearest quartile by more than $1.5IQR$
- Susceptible to outliers: mean, variance, standard deviation, range
- Not susceptible to outliers: quartiles, median, interquartile range

Problem 2.2

Find interquartile range and standard deviation for the following sample:

$-2, 1, 1, 2, 3, 4, 4, 4, 5, 6$

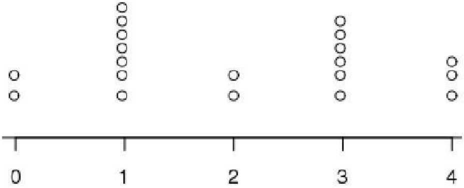
Solution

- $\overline{X} = \frac{-2+1+1+2+3+4+4+4+5+6}{10} = 2.8$
- $s_X^2 = \frac{(-2-2.8)^2+(1-2.8^2)+\cdots+(6-2.8)^2}{9} = 5.51$
- $s_X = 2.35$
- $\frac{n+1}{4} = \frac{11}{4} =$ *between 2nd and 3rd obs*
- $\frac{3(n+1)}{4} = \frac{33}{4} =$ *between 8th and 9th obs*
- $LQ = (1+1)/2 = 1$
- $UQ = (4+5)/2 = 4.5$
- $IQR = 4.5 - 1 = 3.5$

Dot plot

The **dot plot** consists of group of data points plotted on a simple scale

$$\{1, 3, 1, 3, 3, 3, 3, 4, 0, 1, 3, 4, 0, 4, 2, 1, 1, 1, 1, 2\}$$



This works only for a small number of dots

Stem-and-leaf plot

- **Leaf** contains the last digit of the number
- **Stem** contains all of the other digits
- Sample:

  27, 21, 33, 16, 17, 19, 24, 27, 30, 31, 29, 29, 25, 21, 26, 22, 35, 21, 26, 23
- Sort in ascending order:

  16, 17, 19, 21, 21, 21, 22, 23, 24, 25, 26, 26, 27, 27, 29, 29, 30, 31, 33, 35
-

Stem	Leaf
1	679
2	1112345667799
3	0135

Back-to-back stem-and-leaf plot

Two independent samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$

Sometimes it is convenient to plot stem-and-leaf plots back to back

99998887654	1	
97655333322211	2	1223566788999
	3	00112333668
	4	1

Problem 3.1

Forty students wrote a Statistics examination having a maximum of 50 marks. The distribution is given in the following stem plot. What is the third quartile of the mark distribution?

0	28
1	2245
2	01333358889
3	001356679
4	22444466788
5	000

(A) 75

(B) 44

(C) 32

(D) 37.5

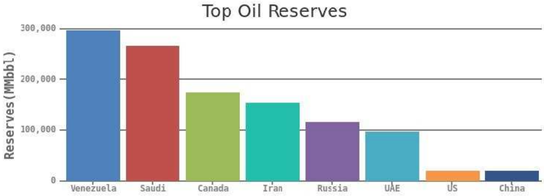
(E) 30

Solution

Correct: **(B)**

Barplot

A **bar chart** or **bar graph** or **bar plot** is a chart that presents grouped data with rectangular bars with lengths proportional to the values that they represent



Barplot

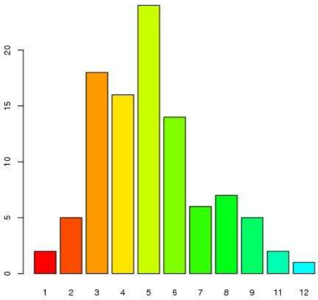
Very often bar plot represents frequencies of observations

11 6 3 5 5 4 5 7 6 5 8 6 6 6 3 3 7 7 7 3 12 9 5 5 4

8 6 8 4 4 5 9 3 4 6 2 3 2 4 3 3 5 5 4 5 7 5 4 3 6

5 4 6 6 3 3 3 9 8 5 3 2 2 4 5 3 2 7 4 5 5 8 8 5 3

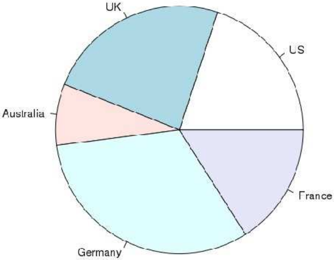
4 5 6 4 5 6 5 5 1 3 1 9 4 5 5 4 3 8 5 9 4 3 6 6 11



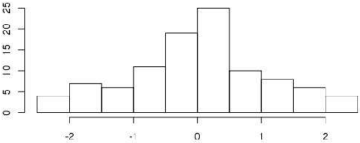
Pie chart

In a **pie chart**, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents.

US	UK	Australia	Germany	France
10	12	4	16	8



Histogram

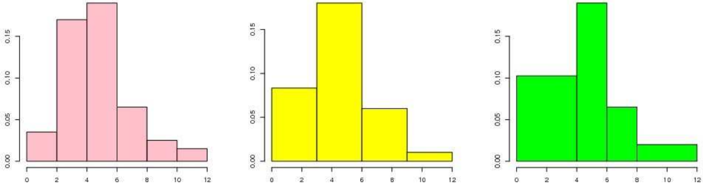


- The first step is to "bin" the range of values
- If the bins are of equal size, the *height* of the rectangle over the bin is proportional to the frequency
- If the bins are not of equal size, the *area* of the rectangle is proportional to the frequency
- The vertical axis is not frequency but density: the number of cases per unit of the variable on the horizontal axis

Histogram

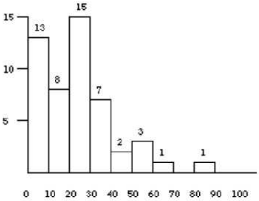
Histogram depends on the binning:

11 6 3 5 5 4 5 7 6 5 8 6 6 6 3 3 7 7 7 3 12 9 5 5 4
8 6 8 4 4 5 9 3 4 6 2 3 2 4 3 3 5 5 4 5 7 5 4 3 6
5 4 6 6 3 3 3 9 8 5 3 2 2 4 5 3 2 7 4 5 5 8 8 5 3
4 5 6 4 5 6 5 5 1 3 1 9 4 5 5 4 3 8 5 9 4 3 6 6 11



Problem 3.2

The following is a histogram showing the actual frequency of the closing prices on the New York exchange of a particular stock. Based on the above frequency histogram for New York Stock exchange, what is the class that contains the 80th percentile ?



(A) *20-30*

(B) *10-20*

(C) *40-50*

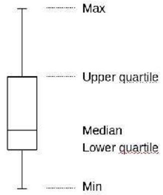
(D) *50-60*

(E) *30-40*

Solution

Correct: **(E)**

Boxplot



- A boxplot is a way of graphically depicting groups of numerical data through their quartiles
- Box plots also have lines extending the boxes (whiskers) indicating variability outside the upper and lower quartiles
- Outliers are plotted as individual points

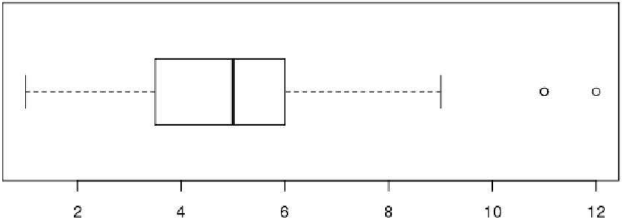
Problem 3.3

Draw the boxplot for the following sample

11 6 3 5 5 4 5 7 6 5 8 6 6 6 3 3 7 7 7 3 12 9 5 5 4
8 6 8 4 4 5 9 3 4 6 2 3 2 4 3 3 5 5 4 5 7 5 4 3 6
5 4 6 6 3 3 3 9 8 5 3 2 2 4 5 3 2 7 4 5 5 8 8 5 3
4 5 6 4 5 6 5 5 1 3 1 9 4 5 5 4 3 8 5 9 4 3 6 6 11

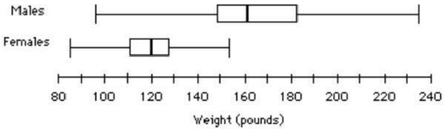
Solution

Min	*LQ*	*Median*	*UQ*	*Max*
1.00	*3.75*	*5.00*	*6.00*	*12.00*



Problem 3.4

The weights of the male and female students are summarized in the following boxplots



Which of the following is NOT correct?

(A) *About 50% of the male students have weights between 150 and 185 lbs*

(B) *About 25% of female students have weights more than 130 lbs*

(C) *The median weight of male students is about 162 lbs*

(D) *The mean weight of female students is about 120 because of symmetry*

(E) *The male students have less variability than the female students*

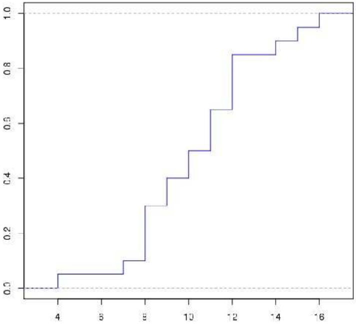
Solution

Correct **(E)**

CDF: Cumulative Density function

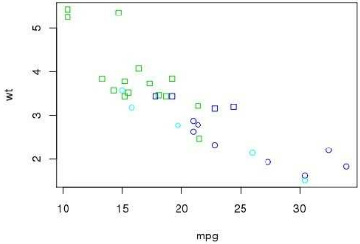
- (Empirical) cumulative distribution function (eCDF)
  $F(x) =$ The proportion of observations that are $\leq x$
- It is a step function that jumps up by $\frac{1}{n}$ at each of the n data points
-

$$16, 10, 12, 4, 12, 11, 8, 9, 8, 7, 12, 11, 8, 14, 9, 12, 8, 15, 10, 11$$



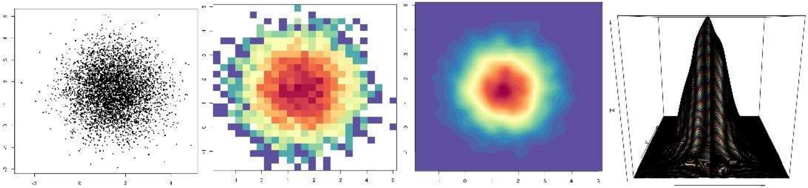
Scatterplot

Scatterplot is used to display values for bivariate data, i.e., for two variables for a set of observations



- Attributes can be represented by
  - Color
  - Shape
  - Size

2D-density



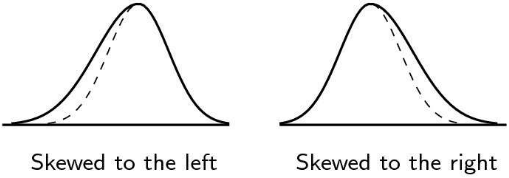
- A **heatmap** is a graphical representation of data where the values contained in a matrix are represented as colors
- Often the values to represent are frequencies

Contingency table

	5	6	7	8	9
(56,72]	24	3	0	1	10
(72,79]	5	15	2	9	10
(79,85]	1	7	19	7	5
(85,97]	0	5	10	14	5

- **Contingency table** is a matrix that displays the multivariate frequency distribution of the variables
- The levels of the factors are discrete
  - Nominal (unordered) variables
  - Ordinal (ordered) variables
  - Discrete interval variables with only a few values
  - Continuous variables grouped into small number of categories

Skewness



- skewed to the left: the left tail is longer
- skewed to the right: the right tail is longer
- Relationship between mean and median
  - skewed to the left: mean $<$ median
  - skewed to the right: mean $>$ median
- Outliers from the right: skewed to the right

Experiment vs. Observational study

- Experiment
- Observational Study
- Sample Survey
- **Experiment involves treatment**
- **Experiment is the only way to establish causal relationship**

Cause and effect relationship

- Every time the trees move the wind starts blowing
- Blow the wind, do the trees move?
- Move the trees, does the wind blow?
- **Association doesn't imply causation**

Types of variables

$$y = f(x, \alpha_1, \alpha_2, \ldots, \alpha_n)$$

- x is **explanatory variable**; the experimenter sets its values
- y is **response variable**; it is the measured response
- $\alpha_2, \alpha_2, \ldots, \alpha_n$ are **confounding variables**
- **Confounding variables also affect the response variable, but their values cannot be controlled**

Control over confounding variables

- Blocking = divide units into blocks
    - within block elements are similar
    - between blocks elements are different
- Control group
    - Subtract baseline levels
    - measure the change instead of the absolute value
- Randomization
    - assign elements randomly into the treatment and control group
    - hope that the effects of confounding factors will “average out”

Simple Random Sample

Simple Random Sample (SRS)

- All elements of the population have equal chances of being selected
- All groups of n elements of the population have equal chances of being selected

Problem 5.1

A big production plant consists of 5 units, and each unit employs 200 workers. You have 10 free tickets to a theater, and you want to distribute them in a fair way among the workers of the plant. One way would be to take 1000 pieces of paper, write worker names on each of them, put the pieces in a hat, and randomly draw ten of them. Another way would be to give two tickets to the heads of each of the five units and to ask each of them to give them to their 200 workers (for instance, by using 200 pieces of paper in a hat). Which procedure will result in a simple random sample?

Solution

(I) Simple random sample

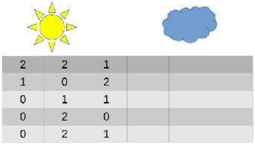
(II) Stratified sample

Problem 5.2

A new fertilizer is to be tested at two concentrations, single and double dose. Thirty plots are available for testing, and half of them are in shade. Describe an experiment, including sampling technique, to test the efficacy of the new fertilizer.

Solution

- *Response variable: yield (in kg)*
- *Explanatory variable: dose of the fertilizer, 3 levels*
- *Confounding factors:*
  - *Shade – blocking*
  - *Crop fertility – control*
  - *Soil fertility – randomization*



2	2	1		
1	0	2		
0	1	1		
0	2	0		
0	2	1		

Sampling Techniques

- Convenience sample
- Simple Random Sample
- Quota sample: pre-defined chances of being selected
- Systematic sample: every nth element
- Stratified sample: union of SRS taken in each stratum
- Multi-level sampling: SRS of SRS
- There is room for you to invent your own sampling technique

Sources of Bias

- Selection bias
- Non-response bias
- Response bias
- Voluntary response bias
- Undercoverage bias
- . . .

Blinding

- **Blinding** is used to avoid placebo effect
- **Placebo effect** is a spontaneous reaction of the brain to the fact of treatment
- **Blind trial** – the subject doesn't know which group he/she belongs
- **Double blind trial** – both the subject and the doctor don't know which group the patient belongs
- **Triple blind trial** – even the statistician who analyses the data doesn't know to which group the patient belongs
- **This is a joke, of course!! Triple blind trial doesn't exist!!!**

Summary

- Statistics draws conclusions about the population based on a sample
- A statistic is a (numeric) function of a sample
- The vertical axis of a histogram is not frequency, but density
- Unlike observational study, experiment involves treatment
- Experiment is the only way to establish causal relationship
- Fight confounding factors: control, blocking, randomization
- In a simple random sample, all groups of n elements have equal chances of being selected