MATH 205: Statistical methods

Lecture 3: Interquartile range and box plot

Tentative schedule

Date	Theme/Topic	Labs	Assignments
Aug 31	Syllabus		
Sep 2—9	Chapter 1: Describing dataset	Section 2: Handling data	
Sep 12—16	Chapter 2: Looking at Relationships	Section 3: Univariate data	
Sep 19—23	Chapter 3: Basic Ideas in Probability	Section 4: Bivariate Data	Homework 1 (due 09/23)
Sep 26—30	Chapters 3-4	Section 4: Correlation	
Oct 3—7	Chapter 4: Random variables and expectations	Section 6: Random data	Homework 2 (due 10/07)
Oct 10—14	Chapter 5: Useful distributions	Section 7: The central limit theorem	
Oct 17—21	Chapter 6: Samples and populations	Section 9: Confidence interval estimation	Homework 3 (due 10/21)
Oct 24—28	Review Midterm exam		Midterm: Oct 28 (lecture) Oct 24-26 (labs)
Oct 31—Nov 4	Chapter 7: The significance of evidence	Section 10: Hypothesis testing	
Nov 7—11	Goodness of Fit	Section 12: Goodness of Fit	Homework 4 (due 11/11)
Nov 14—18	Linear Regression	Section 13: Linear regression	
Nov 21–25	Thanksgiving break		
Nov 28 —Dec 2	One-Way Analysis of Variance	Section 15: Analysis of variance	Homework 5 (due 12/02)
Dec 5—7	Selected topics + Review		
Exam week			

Chapter 1: Describing dataset

- Categorical vs. continuous data
- Datasets as d-tuples

Filter

	eruptions	waiting
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55
7	4.700	88
8	3.600	85
9	1.950	51
10	4.350	85
11	1.833	54
12	3.917	84
13	4.200	78

Summarizing univariate data:

- Mean
- Median
- Standard deviation
- Interquartile Range

Visualizing univariate data:

- Bar chart
- Pie chart
- Histogram
- Box plot

Definition 1.1 (Mean) Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. Their mean is

$$\mathsf{mean}\left(\{x\}\right) = \frac{1}{N}\sum_{i=1}^{i=N} x_i.$$

Properties of the Mean

Useful Facts 1.1 (Properties of the Mean)

- Scaling data scales the mean: or

$$\text{mean}\left(\{kx_i\}\right) = k\text{mean}\left(\{x_i\}\right).$$

- Translating data translates the mean: or

$$\text{mean}\left(\{x_i + c\}\right) = \text{mean}\left(\{x_i\}\right) + c.$$

- The sum of signed differences from the mean is zero: or,

$$\sum_{i=1}^{N}(x_i - \text{mean}\left(\{x_i\}\right)) = 0.$$

Median

Definition 1.4 (Median) The median of a set of data points is obtained by sorting the data points, and finding the point halfway along the list. If the list is of even length, it's usual to average the two numbers on either side of the middle. We write

$$\text{median}\left(\{x\}\right)$$

for the operator that returns the median.

Median is not affected by outliers

The risk of developing iron deficiency is especially high during pregnancy. The problem with detecting such deficiency is that some methods for determining iron status can be affected by the state of pregnancy itself. Consider the following data on transferrin receptor concentration for a sample of women with laboratory evidence of overt iron-deficiency anemia ("Serum Transferrin Receptor for the Detection of Iron Deficiency in Pregnancy," *Amer. J. Clin. Nutrit.*, 1991: 1077–1081):

$$
\begin{array}{llllll}
x_1 = 15.2 & x_2 = 9.3 & x_3 = 7.6 & x_4 = 11.9 & x_5 = 10.4 & x_6 = 9.7 \\
x_7 = 20.4 & x_8 = 9.4 & x_9 = 11.5 & x_{10} = 16.2 & x_{11} = 9.4 & x_{12} = 8.3
\end{array}
$$

The list of ordered values is

$$7.6 \quad 8.3 \quad 9.3 \quad 9.4 \quad 9.4 \quad 9.7 \quad 10.4 \quad 11.5 \quad 11.9 \quad 15.2 \quad 16.2 \quad 20.4$$

Since $n = 12$ is even, we average the $n/2 =$ sixth- and seventh-ordered values:

$$\text{sample median} = \frac{9.7 + 10.4}{2} = 10.05$$

Measures of variability: deviation from the mean

Definition 1.2 (Standard Deviation) Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. The standard deviation of this dataset is:

$$\begin{aligned}\mathsf{std}(\{x_i\}) &= \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}(x_i - \mathsf{mean}(\{x\}))^2} \\ &= \sqrt{\mathsf{mean}(\{(x_i - \mathsf{mean}(\{x\}))^2\})}.\end{aligned}$$

Properties of the standard deviation

Useful Facts 1.2 (Properties of Standard Deviation)

- Translating data does not change the standard deviation, i.e. $\mathsf{std}(\{x_i + c\}) = \mathsf{std}(\{x_i\})$.
- Scaling data scales the standard deviation, i.e. $\mathsf{std}(\{kx_i\}) = k\mathsf{std}(\{x_i\})$.
- For any dataset, there can be only a few items that are many standard deviations away from the mean. For N data items, x_i, whose standard deviation is σ, there are at most $\frac{1}{k^2}$ data points lying k or more standard deviations away from the mean.
- For any dataset, there must be at least one data item that is at least one standard deviation away from the mean, that is, $(\mathsf{std}(\{x\}))^2 \leq \max_i (x_i - \mathsf{mean}(\{x\}))^2$.

The standard deviation is often referred to as a scale parameter; it tells you how broadly the data spreads about the mean.

Variance

Definition 1.3 (Variance) Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. where $N > 1$. Their variance is:

$$\begin{aligned}\mathsf{var}\left(\{x\}\right) &= \frac{1}{N}\left(\sum_{i=1}^{i=N}(x_i - \mathsf{mean}\left(\{x\}\right))^2\right) \\ &= \mathsf{mean}\left(\left\{(x_i - \mathsf{mean}\left(\{x\}\right))^2\right\}\right).\end{aligned}$$

Interquartile range

Percentiles and Quartiles

Definition 1.5 (Percentile) The k'th percentile is the value such that $k\%$ of the data is less than or equal to that value. We write $\text{percentile}(\{x\}, k)$ for the k'th percentile of dataset $\{x\}$.

Definition 1.6 (Quartiles) The first quartile of the data is the value such that 25% of the data is less than or equal to that value (i.e. $\text{percentile}(\{x\}, 25)$). The second quartile of the data is the value such that 50% of the data is less than or equal to that value, which is usually the median (i.e. $\text{percentile}(\{x\}, 50)$). The third quartile of the data is the value such that 75% of the data is less than or equal to that value (i.e. $\text{percentile}(\{x\}, 75)$).

Percentiles and Quartiles

- If there are n data points, then the p quantile occurs at the position $1 + (n - 1)p$ with weighted averaging if this is between integers.
- For example the .25 quantile of the numbers

  $$10, 17, 18, 25, 28, 28$$

  occurs at the position 1+(6-1)(.25) $=$ 2.25. That is $1/4$ of the way between the second and third number which in this example is 17.25.

Interquartile range

Definition 1.7 (Interquartile Range) The interquartile range of a dataset $\{x\}$ is $\mathsf{iqr}\{x\} = \mathsf{percentile}(\{x\}, 75) - \mathsf{percentile}(\{x\}, 25)$.

Example

Consider the previous example:

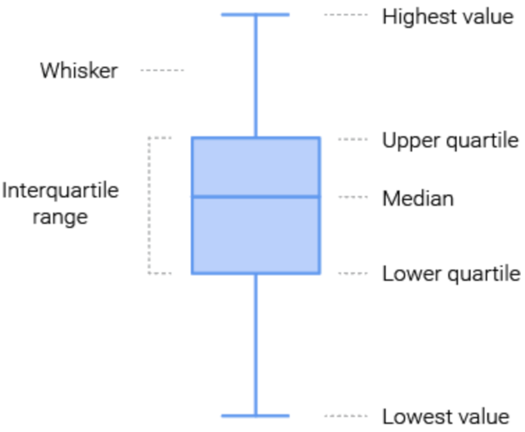
$$\begin{array}{llllll} x_1 = 15.2 & x_2 = 9.3 & x_3 = 7.6 & x_4 = 11.9 & x_5 = 10.4 & x_6 = 9.7 \\ x_7 = 20.4 & x_8 = 9.4 & x_9 = 11.5 & x_{10} = 16.2 & x_{11} = 9.4 & x_{12} = 8.3 \end{array}$$

The list of ordered values is

$$7.6 \quad 8.3 \quad 9.3 \quad 9.4 \quad 9.4 \quad 9.7 \quad 10.4 \quad 11.5 \quad 11.9 \quad 15.2 \quad 16.2 \quad 20.4$$

Compute the Interquartile range of this dataset.

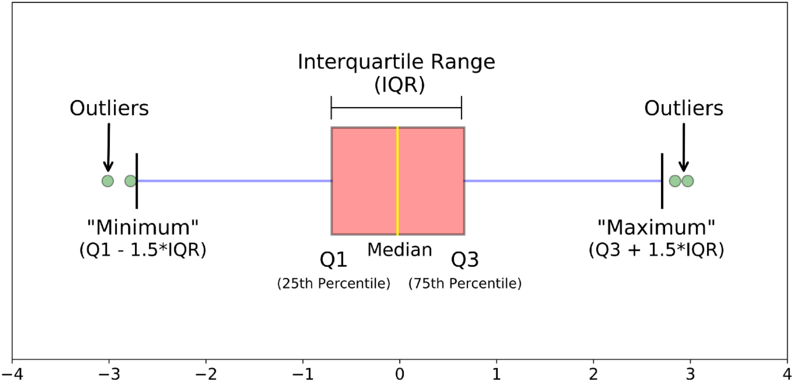
Boxplot


Highest value
Whisker
Upper quartile
Interquartile range
Median
Lower quartile
Lowest value

Boxplot with outliers

- Convention: any point further than 1.5*[Interquartile range] from the closest quartile is called *an outlier*
- Boxplot with outliers: The whisker is shorten to just include non-outliers. Outliers are plotted by points.

Boxplot with outliers



Final note: Standard coordinates

- It is often possible to get some useful insights about one univariate dataset from visualizations
- However, they are hard to compare because each is in a different set of units

Definition 1.8 (Standard Coordinates) Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. We represent these data items in standard coordinates by computing

$$\hat{x}_i = \frac{(x_i - \mathsf{mean}(\{x\}))}{\mathsf{std}(\{x\})}.$$

We write $\{\hat{x}\}$ for a dataset that happens to be in standard coordinates.

Problem

Definition 1.8 (Standard Coordinates) Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. We represent these data items in standard coordinates by computing

$$\hat{x}_i = \frac{(x_i - \text{mean}(\{x\}))}{\text{std}(\{x\})}.$$

We write $\{\hat{x}\}$ for a dataset that happens to be in standard coordinates.

Prove that:

- $mean(\{\hat{x}\}) = 0$
- $std(\{\hat{x}\}) = 1$

Standard coordinates

- We could then normalize the data by subtracting the location (mean) and dividing by the standard deviation (scale)
- The resulting values are unitless, and have zero mean