4.2 Measures of Variability

Once we have located the central values of a set of data, it is important to measure the *variability*, whether the data values are tightly clustered or spread out. At the heart of Statistics lies variability: measuring it, reducing it, distinguishing random from real variability, identifying the various sources of real variability and making decisions in the presence of it. We need to know how "unstable" the data is and how much the values differ from its average or from other middle values. These numbers will have small values for closely grouped data (little variation) and larger values for more widely spread out data (large variation).

The measures of variation will also help us assess the reliability of our estimates and the accuracy of our forecasts.

Quantiles, percentiles and quartiles

Consider the primary data $X = \{x_1, \ldots, x_N\}$. The first two measures of variation give a very general idea of the spread in the data values.

Definition 4.1. *The **range*** (range) *of X is the difference*

$$x_{max} - x_{min}.$$

If the values of X are sorted in increasing order, then the range is $x_N - x_1$.

Definition 4.2. *The **mean absolute deviation*** (mad) *of X is the mean of the absolute value of the deviations from the mean, i.e. the value*

$$MAD_1 = \frac{1}{N}\sum_{i=1}^{N} |x_i - \overline{x}|.$$

*The **median absolute deviation*** (mad) *of X is the median of the absolute value of the deviations from the median, i.e. the value*

$$MAD_2 = \text{median}\{|x_i - \overline{M}|\}.$$

Like the median, the median absolute deviation is not influenced by extreme values, whereas the mean absolute deviation is.

Next, following the idea behind the definition of the median, we define values that divide the data into certain percentages. We simply replace 0.5 in its definition by some probability $0 < p < 1$.

Definition 4.3. *Let X be a set of data sorted increasingly, $p \in (0,1)$ and $k = 1, 2, \ldots, 99$.*

(1) *A **sample** p**-quantile** (quantile) is any number that exceeds at most $100p\%$ of the sample and is exceeded by at most $100(1-p)\%$ of the sample.*

(2) *A k**-percentile** (prctile) P_k is a $(k/100)$-quantile. So, P_k exceeds at most $k\%$ and is exceeded by at most $(100-k)\%$ of the data*

(3) *The **quartiles** of X are the values*

$$Q_1 = P_{25}, \;\; Q_2 = P_{50} = \overline{M} \;\text{ and }\; Q_3 = P_{75}.$$

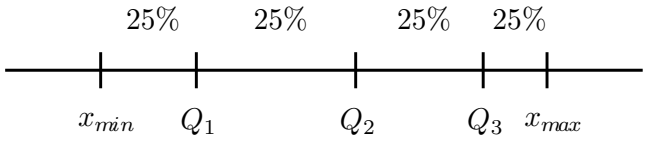


Fig. 1: Quartiles

Definition 4.4. *Let X be a set of sorted data with quartiles Q_1, Q_2 and Q_3.*

(1) *The **interquartile range** (iqr) is the difference between the third and the first quartile*

$$IQR = Q_3 - Q_1. \tag{4.1}$$

(2) *The **interquartile deviation** or the **semi interquartile range** is the value*

$$IQD = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2}. \tag{4.2}$$

(3) *The **interquartile deviation coefficient** or the **relative interquartile deviation** is the value*

$$IQDC = \frac{IQD}{\overline{M}} = \frac{Q_3 - Q_1}{2Q_2}. \tag{4.3}$$

Remark 4.5.
1. The interquartile deviation is an absolute measure of variation and it has an important property: the range $\overline{M} \pm IQD$ contains approximately 50% of the data.
2. The interquartile deviation coefficient $IQDC$ varies between -1 and 1, taking values close to 0 for symmetrical distributions, with little variation and values close to ± 1 for skewed data with large variation.

Example 4.6. Recall our example about the CPU times (in seconds) for $N = 30$ randomly chosen jobs (sorted ascendingly):

9	15	19	22	24	25	30	**34**	35	35
36	36	37	38	42	43	46	48	54	55
56	56	**59**	62	69	70	82	82	89	139

Let us compute various measures of variation.

Solution. For this example, the range is

$$139 - 9 = 130 \text{ seconds}$$

and the mean and median absolute deviations are

$$\begin{aligned} MAD_1 &= 19.6133, \\ MAD_2 &= 13.5. \end{aligned}$$

To determine the quartiles, notice that 25% of the sample equals $30/4 = 7.5$ and 75% of the sample is $90/4 = 22.5$ observations. From the ordered sample, we see that the 8th element, 34, has 7 observations to its left and 22 to its right, so it has *no more* than 7.5 observations to the left and *no more* than 22.5 observations to the right of it. Hence, $Q_1 = 34$.

Similarly, the third quartile is the 23rd smallest element, $Q_3 = 59$. Recall from last time that the second quartile (the median) si $Q_2 = \overline{M} = 42.5$. Then

$$\begin{aligned} IQR &= 59 - 34 = 25, \\ IQD &= IQR/2 = 12.5, \\ IQDC &= IQD/Q_2 = 0.2941. \end{aligned}$$

The interval

$$\overline{M} \pm IQD \ = \ [30, 55]$$

contains 14 observations.

The value of the $IQDC$ is close neither to 0, nor to the values ± 1. So the data doesn't show strong symmetry or strong asymmetry. This may be due to the extreme values 9 and/or 139.

■

Remark 4.7. For populations or very large data sets, calculating exact percentiles can be computationally very expensive since it requires sorting all the data values. Machine learning and statistical software use special algorithms (such as linear interpolation) to get an approximate percentile that can be calculated very quickly and is guaranteed to have a certain accuracy.

Outliers

The interquartile range is also involved in another important aspect of statistical analysis, namely the detection of outliers. An *outlier*, as the name suggests, is basically an atypical value, "far away" from the rest of the data, that does not seem to belong to the distribution of the rest of the values in the data set.

We have seen how the mean is very sensitive to outliers. Other statistical procedures can be gravely affected by the presence of outliers in the data. Thus, the problem of detecting and locating an outlier is an important part of any statistical data analysis process.

How to classify a value as being "extreme"? First, we could use a simple property, known as the "3σ rule". This is an application of Chebyshev's inequality

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2}, \ \forall \varepsilon > 0.$$

If we use the classical notations $E(X) = \mu, V(X) = \sigma^2$, $\text{Std}(X) = \sigma$ for the mean, variance and standard deviation of X and take $\varepsilon = 3\sigma$, we get

$$\begin{aligned} P(|X - \mu| < 3\sigma) \ &\geq \ 1 - \frac{\sigma^2}{9\sigma^2} \\ &= \ \frac{8}{9} \ \approx \ .89. \end{aligned}$$

This is saying that it is *very* probable (at least 0.89 probable) that $|X - \mu| < 3\sigma$, or, equivalently, that $\mu - 3\sigma < X < \mu + 3\sigma$. In words, the 3σ rule states that *most of the values that any random variable takes, at least* 89%*, lie within* 3 *standard deviations away from the mean*. This property is

true in general, for any distribution, but especially for unimodal and symmetrical ones, where that percentage is even higher.

Based on that, one simple procedure would be to consider an outlier any value that is more than 2.5 standard deviations away from the mean, and an *extreme* outlier a value more than 3 standard deviations away from the mean.

A more general approach, that works well also for skewed data, is to consider an outlier any observation that is outside the range

$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right] = \left[Q_1 - 3IQD,\ Q_3 + 3IQD\right].$$

Also, the coefficient $3/2$ can be replaced by some other number to decrease or enlarge the interval of "normal" values (or, equivalently, the domain that covers the outliers):

$$\left[Q_1 - w \cdot IQR,\ Q_3 + w \cdot IQR\right],\ w = 0.5, 1, 1.5.$$

For our example on CPU times of processors, we have

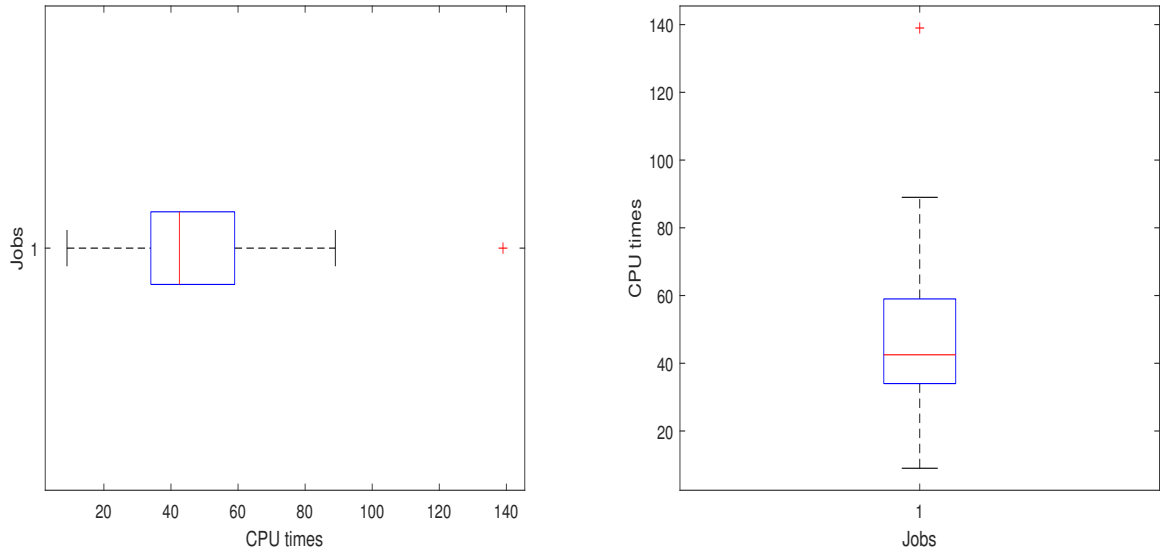
$$\begin{aligned} Q_1 - \frac{3}{2}IQR &= -3.5, \\ Q_3 + \frac{3}{2}IQR &= 96.5, \end{aligned}$$

so observations outside the interval $[-3.5, 96.5]$ are considered outliers. In this case, there is only one, the value 139.

Boxplots

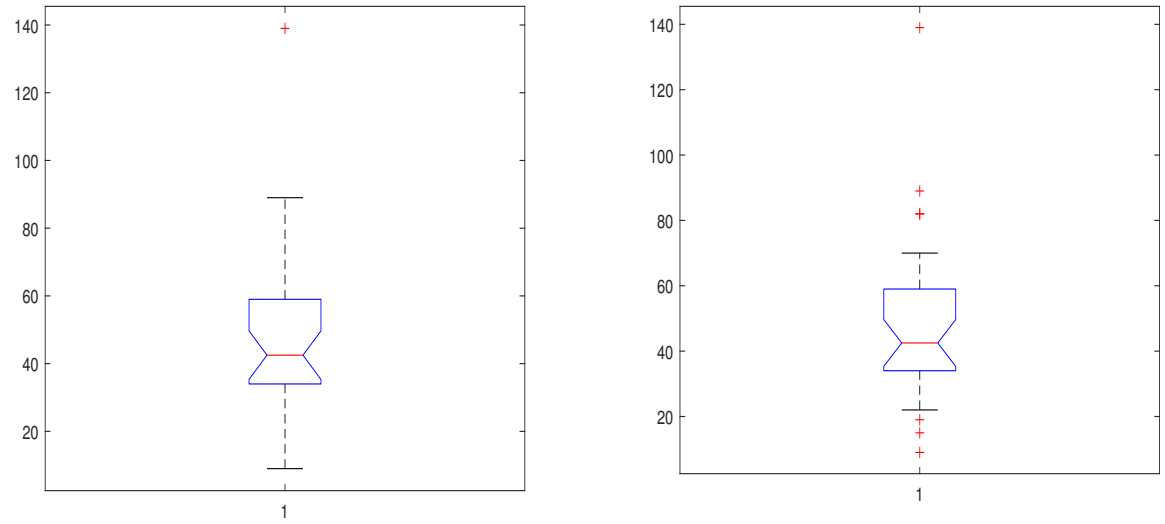
All the information we discussed above is summarized in a graphical display, called a **boxplot** (boxplot), a plot in which a rectangle is drawn to represent the second and third quartiles (so the interquartile range), with a line inside for the median value and which indicates which values are considered extreme. The "whiskers" of the boxplot are the endpoints of the interval on which normal values lie (so everything outside the whiskers is considered an outlier).

For the data in Example 4.6, the boxplot is displayed in Figure 2 and it can be drawn vertically (default) or horizontally.



(a) horizontally (b) vertically

Fig. 2: Quartiles, Interquartile Range, Outliers



(a) boxplot with a notch (b) whisker $w = 0.5$

Fig. 3: Boxplots

The box can have a "notch" (indentation) at the value of the median, as in Figure 3(a). The width of the interval of the whiskers can be changed. The interval that determines the outliers (i.e., outside of which values are considered too extreme, outliers) is

$$[Q_1 - w \cdot IQR, Q_3 + w \cdot IQR].$$

The default value is $w = 1.5$. With the smaller whiskers, boxplot displays more data points as outliers. In Figure 3(b), the whisker size is set to $w = 0.5$. Then, outliers are all the values outside the interval $[Q_1 - 0.5 \cdot IQR, Q_3 + 0.5 \cdot IQR] = [21.5, 71.5]$. These would be $9, 15, 19$ (too small) and $82, 89, 139$ (too large).

Boxplots are also very useful when we want to compare data from different samples (see Figure 4). We can compare the interquartile ranges, to examine how the data is dispersed between each sample. The longer the box, the more dispersed the data.

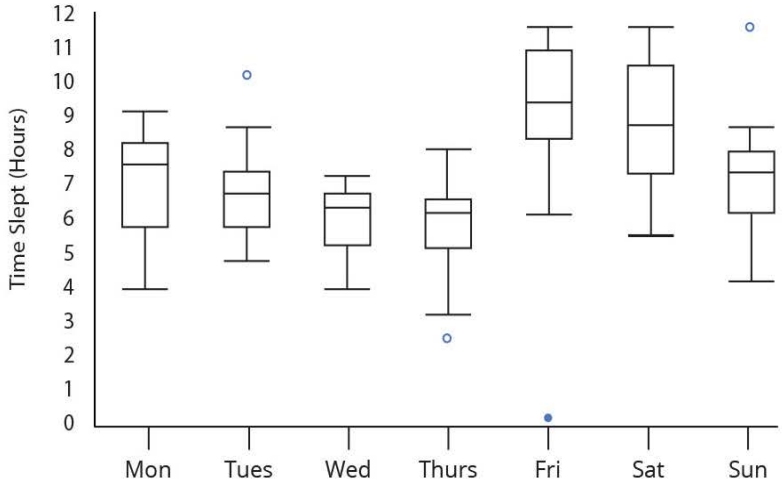


Fig. 4: Multiple boxplots

Moments, variance, standard deviation and coefficient of variation

The idea of the mean can be generalized, by taking various powers of the values in the data.

Definition 4.8.

(1) *The **moment of order k** is the value*

$$\overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{N} x_i^k, \;\; \overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^k, \tag{4.4}$$

for primary and for grouped data, respectively.

(2) *The **central moment of order k*** (moment) *is the value*

$$\overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^k, \;\; \overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^k \tag{4.5}$$

for primary and for grouped data, respectively.

(3) *The **variance*** (var) *is the value*

$$\overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2, \;\; \overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^2 \tag{4.6}$$

*for primary and for grouped data, respectively. The quantity $\overline{\sigma} = \sqrt{\overline{\sigma}^2}$ is the **standard deviation*** (std).

Remark 4.9.

1. A more efficient computational formula for the variance is

$$\overline{\sigma}^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\Big(\sum_{i=1}^{N} x_i\Big)^2\right) = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\right), \tag{4.7}$$

which follows straight from the definition.

2. We will see later that when the data represents a sample (not the entire population), a better formula is

$$\begin{aligned} s^2 &= \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \overline{x})^2 &&= \frac{1}{N-1}\Big(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\Big), \\ s^2 &= \frac{1}{N-1}\sum_{i=1}^{n} f_i (x_i - \overline{x})^2 &&= \frac{1}{N-1}\Big(\sum_{i=1}^{N} f_i x_i^2 - N\overline{x}^2\Big), \end{aligned} \tag{4.8}$$

for the *sample* variance for primary or grouped data. The reason the sum is divided by $N-1$ instead of N will have to do with the "bias" of an estimator and will be explained later on in the next chapter. To fully explain why using N leads to a biased estimate involves the notion of *degrees of freedom*, which takes into account the number of constraints in computing an estimate. The sample observations $x_1, \ldots, x_N$ are independent (by the definition of a random sample), but when computing the variance, we use the variables $x_1 - \overline{x}, \ldots, x_N - \overline{x}$. Notice that by subtracting the sample mean $\overline{x}$ from each observation, there exists a linear relation among the elements, namely

$$\sum_{k=1}^{N}(x_k - \overline{x}) = 0$$

and, thus, we lose 1 degree of freedom due to this constraint. Hence, there are only $N-1$ degrees of freedom. So, we will use (4.7) to compute the variance of a set of data that represents a population and (4.8) for the variance of a sample.

Example 4.10. Consider again our previous example on CPU times (in seconds) for $N = 30$ randomly chosen jobs:

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

Recall that for this data the sample mean was $\overline{x} = 48.2333$ seconds. The sample variance is

$$s^2 \;=\; \frac{(70-48.2333)^2 + \ldots + (19-48.2333)^2}{30-1} \;=\; \frac{20391}{29} \;\approx\; 703.1506 \text{ sec}^2.$$

Alternatively, using (4.7),

$$s^2 \;=\; \frac{70^2 + \ldots + 19^2 - 30 \cdot 48.2333^2}{30-1} \;=\; \frac{90185 - 69794}{29} \;\approx\; 703.1506 \text{ sec}^2.$$

The sample standard deviation is

$$s \;=\; \sqrt{703.1506} \;\approx\; 26.1506 \text{ sec}.$$

By the 3σ rule, using $\overline{x}$ and s as estimates for the population mean μ and population standard deviation σ, we may infer that at least 89% of the tasks performed by this processor require between $\overline{x} - 3s = -30.2185$ and $\overline{x} + 3s = 126.6851$ (so less than 126.6851) seconds of CPU time.

Definition 4.11. *The **coefficient of variation** is the value*

$$CV = \frac{s}{\overline{x}}.$$

Remark 4.12.
1. The coefficient of variation is also known as the **relative standard deviation (RSD)**.
2. It can be expressed as a ratio or as a percentage. It is useful in comparing the degrees of variation of two sets of data, even when their means are different.
2. The coefficient of variation is used in fields such as Analytical Chemistry, Engineering or Physics when doing quality assurance studies. It is also widely used in Business Statistics. For example, in the investing world, the coefficient of variation helps brokers determine how much volatility (risk) they are assuming in comparison to the amount of return they can expect from a certain investment. The lower the value of the CV, the better the risk-return trade off.

5 Correlation and Regression

So far we have been discussing a number of descriptive techniques for describing one variable only. However, a very important part of Statistics is describing the association between two (or more) variables, whether or not they are independent, and if they are not, what is the nature of their dependence. One of the most fundamental concepts in statistical research is the concept of correlation.

Correlation is a measure of the relationship between one dependent variable, called *response* and one or more independent variables, called *predictor(s)*. If two variables are correlated, this means that one can use information about one variable to predict the values of the other variable.

Regression is then the method or statistical procedure that is used to establish that relationship. Establishing and testing such a relation enables us:
- to understand interactions, causes, and effects among variables;
- to predict unobserved variables based on the observed ones;
- to determine which variables significantly affect the variable of interest.

Example 5.1 (World Population)**.** According to the International Data Base of the U.S. Census Bureau, population of the world grows according to Table 1. How can we use these data to predict the world population in years 2025 and 2030?

Figure 5 shows that the population (response) is tightly related to the year (predictor). It increases every year, and its growth is almost linear. If we estimate the regression function relating

Year	Pop. (mln. people)	Year	Pop.(mln.people)	Year	Pop.(mln.people)
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6970
1965	3350	1990	5287	2015	7405
1970	3712	1995	5700	2020	7821

Table 1: World Population 1950-2020

our response and our predictor (see the dotted line on Figure 5) and extend its graph to the year 2030, the forecast is ready.

A straight line that fits the observed data for years $1950 - 2020$ predicts the population of 8.06 billion in 2025 and 8.444 billion in 2030. It also shows that between 2020 and 2025, the world population reaches the historical mark of 8 billion (which actually happened last summer ...). How accurate is the forecast obtained in this example? The observed population during $1950 - 2020$ appears rather close to the estimated regression line in Figure 5. It is reasonable to hope that it will continue to do so through 2030.

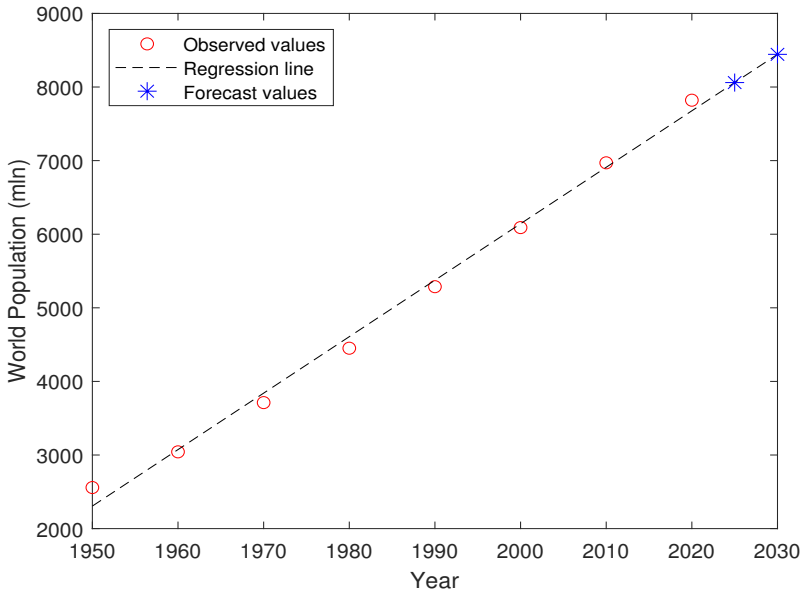


Fig. 5: World population and regression forecast

Example 5.2 (House Prices)**.** Seventy house sale prices in a certain county (in the U.S.) are depicted in Figure 6 along with the house area.

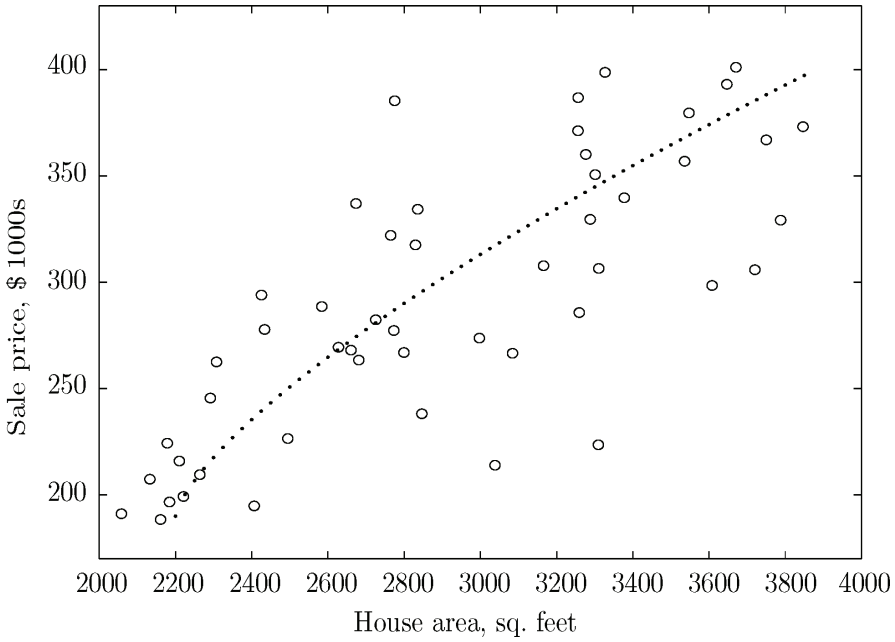


Fig. 6: House prices and square footage

First, we see a clear relation between these two variables, and in general, bigger houses are more expensive. However, the trend no longer seems linear.

Second, there is a large amount of variability around this trend. Indeed, area is not the only factor determining the house price. Houses with the same area may still be priced differently. Then, how can we estimate the price of a 3200-square-foot house? We can estimate the general trend (the dotted line in Figure 6) and plug 3200 into the resulting formula, but due to obviously high variability, our estimation will not be as accurate as in Example 5.1.

To improve our estimation in the last example, we may take other factors into account: location, the number of bedrooms and bathrooms, the backyard area, the average income of the neighborhood, etc. If all the added variables are relevant for pricing a house, our model will have a closer fit and will provide more accurate predictions.

5.1 Univariate Regression, Curves of Regression

We will restrict our discussion to the case of **univariate regression**, predicting response Y based on *one* predictor X.

So, we have two vectors X and Y of the same length. We can get a first idea of the relationship between the two, by plotting them in a **scattergram**, or **scatterplot**, which is a plot of the points with coordinates $(x_i, y_i)_{i=\overline{1,k}}$, $x_i \in X$, $y_i \in Y$, $i = \overline{1,k}$. We group the N primary data into mn classes and denote by (x_i, y_j) the class mark and by f_{ij} the absolute frequency of the class (i, j), $i = \overline{1,m}$, $j = \overline{1,n}$. Then we represent the two-dimensional characteristic (X, Y) in a *correlation table*, or *contingency table*, as shown in Table 2.

$X \setminus Y$	y_1	$\dots$	y_j	$\dots$	y_n	
x_1	f_{11}	$\dots$	f_{1j}	$\dots$	f_{1n}	$f_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	f_{i1}	$\dots$	f_{ij}	$\dots$	f_{in}	$f_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_m	f_{m1}	$\dots$	f_{mj}	$\dots$	f_{mn}	$f_{m.}$
	$f_{.1}$	$\dots$	$f_{.j}$	$\dots$	$f_{.n}$	$f_{..} = N$

Table 2: Correlation Table

Notice that

$$\sum_{j=1}^{n} f_{ij} = f_{i.}, \ \sum_{i=1}^{m} f_{ij} = f_{.j}, \ \sum_{i=1}^{m} f_{i.} = \sum_{j=1}^{n} f_{.j} = f_{..} = N.$$

Now we can define numerical characteristics associated with (X, Y).

Definition 5.3. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 2 and let $k_1, k_2 \in \mathbb{N}$.*

(1) *The **(initial) moment of order (k_1, k_2)** of (X, Y) is the value*

$$\overline{\nu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} x_i^{k_1} y_j^{k_2}. \tag{5.1}$$

(2) *The **central moment of order (k_1, k_2)** of (X, Y) is the value*

$$\overline{\mu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} (x_i - \overline{x})^{k_1} (y_j - \overline{y})^{k_2}, \tag{5.2}$$

where $\overline{x} = \overline{\nu}_{10} = \dfrac{1}{N} \sum_{i=1}^{m} f_{i.} x_i$ and $\overline{y} = \overline{\nu}_{01} = \dfrac{1}{N} \sum_{j=1}^{n} f_{.j} y_j$ are the means of X and Y, respectively.

Remark 5.4. Just as the means of the two characteristics X and Y can be expressed as moments of (X, Y), so can their variances:

$$\begin{aligned} \overline{\sigma}_X^2 &= \overline{\mu}_{20} = \overline{\nu}_{20} - \overline{\nu}_{10}^2, \\ \overline{\sigma}_Y^2 &= \overline{\mu}_{02} = \overline{\nu}_{02} - \overline{\nu}_{01}^2. \end{aligned}$$

Definition 5.5. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 2.*

(1) *The **covariance*** (cov) *of (X, Y) is the value*

$$\operatorname{cov}(X, Y) = \overline{\mu}_{11} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}(x_i - \overline{x})(y_j - \overline{y}). \tag{5.3}$$

(2) *The **correlation coefficient*** (corrcoef) *of (X, Y) is the value*

$$\overline{\rho} = \overline{\rho}_{XY} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\overline{\mu}_{20}}\sqrt{\overline{\mu}_{02}}} = \frac{\overline{\mu}_{11}}{\overline{\sigma}_X \overline{\sigma}_Y}. \tag{5.4}$$

These two notions have been mentioned before, for two random variables. They are defined similarly for sets of data and they have the same properties. The covariance gives a rough idea of the relationship between X and Y. As before, if X and Y are independent (so there is no relationship, no correlation between them), then the covariance is 0. If large values of X are associated with large values of Y, then the covariance will have a positive value, if, on the contrary, large values of X are associated with small values of Y, then the covariance will have a negative value. Also, an easier computational formula for the covariance is $\operatorname{cov}(X, Y) = \overline{\nu}_{11} - \overline{x} \cdot \overline{y}$.

The correlation coefficient is then

$$\overline{\rho} = \frac{\overline{\nu}_{11} - \overline{x} \cdot \overline{y}}{\overline{\sigma}_X \overline{\sigma}_Y}$$

and, as before, it satisfies the inequality

$$-1 \leq \overline{\rho} \leq 1 \tag{5.5}$$

and, by its variation between -1 and 1, its value measures the linear relationship between X and Y. If $\overline{\rho}_{XY} = 1$, there is a *perfect positive correlation* between X and Y, if $\overline{\rho}_{XY} = -1$, there is a *perfect negative correlation* between X and Y. In both cases, the linearity is "perfect", i.e there exist $a, b \in \mathbb{R},\ a \neq 0$, such that $Y = aX + b$. If $\overline{\rho}_{XY} = 0$, then there is no linear correlation

between X and Y, they are said to be *(linearly) uncorrelated*. However, in this case, they may not be independent, some other type of relationship (not linear) may exist between them.

In our task of finding a relationship between X and Y, we may go the following path: knowing the value of one of the characteristics, try to find a probable, an "expected" value for the other. If the two characteristics are related in any way, then there should be a pattern developing, that is, the expected value of one of them, *conditioned* by the other one taking a certain value, should be a function of that value that the other variable assumes. In other words, we should consider *conditional means*, defined similarly to regular means, only taking into account the condition.

Definition 5.6. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 2.*

(1) *The **conditional mean** of Y, given $X = x_i$, is the value*

$$\overline{y}_i = \overline{y}(x_i) = \frac{1}{f_{i\cdot}} \sum_{j=1}^{n} f_{ij} y_j,\ i = \overline{1, m}. \tag{5.6}$$

(2) *The **conditional mean** of X, given $Y = y_j$, is the value*

$$\overline{x}_j = \overline{x}(y_j) = \frac{1}{f_{\cdot j}} \sum_{i=1}^{m} f_{ij} x_i,\ j = \overline{1, n}. \tag{5.7}$$

Definition 5.7. *Let (X, Y) be a two-dimensional characteristic.*

(1) *The curve $y = f(x)$ formed by the points with coordinates $(x_i, \overline{y}_i)$, $i = \overline{1, m}$, is called the **curve of regression** of Y on X.*

(2) *The curve $x = g(y)$ formed by the points with coordinates $(y_j, \overline{x}_j)$, $j = \overline{1, n}$, is called the **curve of regression** of X on Y.*

Remark 5.8. The curve of regression of a characteristic Y with respect to another characteristic X is then the mean value of Y, $\overline{y}(x)$, given $X = x$. The curve of regression is determined so that it approximates best the scatterplot of (X, Y).

5.2 Least Squares Estimation, Linear Regression

One of the most popular ways of finding curves of regression is the *least squares method*.
Assume the curve of regression of Y on X is of the form

$$y = y(x) = f(x; a_1, \ldots, a_s).$$

We determine the unknown parameters $a_1, \ldots, a_s$ so that the *sum of squares error* (SSE) (the sum of the squares of the differences between the responses y_j and their fitted values $y(x_i)$, each counted with the corresponding frequency)

$$S = SSE = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - y(x_i)\Big)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - f(x_i; a_1, \ldots, a_s)\Big)^2$$

is minimum (hence, the name of the method).
We find the point of minimum $(\overline{a}_1, \ldots, \overline{a}_s)$ of S by solving the system

$$\frac{\partial S}{\partial a_k} = 0,\ k = \overline{1, s},$$

i.e.

$$-2\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - f(x_i; a_1, \ldots, a_s)\Big)\frac{\partial f(x_i; a_1, \ldots, a_s)}{\partial a_k} = 0, \tag{5.8}$$

for every $k = \overline{1, s}$.
Then the equation of the curve of regression of Y on X is

$$y = f\left(x; \overline{a}_1, \ldots, \overline{a}_s\right).$$

Let us consider the case of *linear regression* and find the equation of the *line of regression* of Y on X. We are finding a curve

$$y = ax + b,$$

for which

$$S(a, b) = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - ax_i - b\Big)^2$$

is minimum. The system (5.8) becomes

$$\begin{cases} \left(\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}x_i^2\right) a \;+\; \left(\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}x_i\right) b \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}x_i y_j \\ \left(\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}x_i\right) a \;+\; \left(\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\right) b \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}y_j \end{cases}$$

and after dividing both equations by N,

$$\begin{cases} \overline{\nu}_{20}a \;+\; \overline{\nu}_{10}b \;=\; \overline{\nu}_{11} \\ \overline{\nu}_{10}a \;+\; \overline{\nu}_{00}b \;=\; \overline{\nu}_{01}. \end{cases}$$

Its solution is

$$\overline{a} \;=\; \frac{\overline{\nu}_{11}-\overline{\nu}_{10}\overline{\nu}_{01}}{\overline{\nu}_{20}-\overline{\nu}_{10}^2} \;=\; \frac{\overline{\nu}_{11}-\overline{x}\cdot\overline{y}}{\overline{\sigma}_X^2} \;=\; \frac{\overline{\nu}_{11}-\overline{x}\cdot\overline{y}}{\overline{\sigma}_X\overline{\sigma}_Y}\cdot\frac{\overline{\sigma}_Y}{\overline{\sigma}_X} \;=\; \overline{\rho}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X},$$

$$\overline{b} \;=\; \overline{\nu}_{01}-\overline{\nu}_{10}\overline{a} \;=\; \overline{y}-\overline{a}\cdot\overline{x}.$$

So the equation of the line of regression of Y on X is

$$y-\overline{y}=\overline{\rho}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X}\,(x-\overline{x}) \tag{5.9}$$

and, by analogy, the equation of the line of regression of X on Y is

$$x-\overline{x}=\overline{\rho}\,\frac{\overline{\sigma}_X}{\overline{\sigma}_Y}\,(y-\overline{y})\,. \tag{5.10}$$

Example 5.9. Let us consider the world population data in Example 5.1 and find the equation of the line of regression.

Solution. For the world population $(1950-2020)$ data, we find

$$\begin{aligned} \overline{x} &= 1985,\ \overline{y} = 4991.5 \\ \overline{\sigma}_X &= 24.5,\ \overline{\sigma}_Y = 1884.6 \\ \overline{\rho} &= 0.9972 \end{aligned}$$

and the equation of the line of regression

$$y \;=\; 76.72x - 147300.5.$$

With this, we were able to forecast the values of 8.0604 billion for the year 2025 and 8.444 billion for 2030. Also, based on this model, the predicted population for 2023 is 7.9069 billion people.

■

Let us analyze linear regression further.

Remark 5.10.

1. The point of intersection of the two lines of regression (5.9) and (5.10) is $(\overline{x}, \overline{y})$. This is called the *centroid* of the distribution of the characteristic (X, Y).

2. The slope $\overline{a}_{Y|X} = \overline{\rho}\,\dfrac{\overline{\sigma}_Y}{\overline{\sigma}_X}$ of the line of regression of Y on X is called the *coefficient of regression* of Y on X. Similarly, $\overline{a}_{X|Y} = \overline{\rho}\,\dfrac{\overline{\sigma}_X}{\overline{\sigma}_Y}$ is the coefficient of regression of X on Y and we have the relation

$$\overline{\rho}^2 = \overline{a}_{Y|X}\;\overline{a}_{X|Y}.$$

3. For the angle α between the two lines of regression, we have

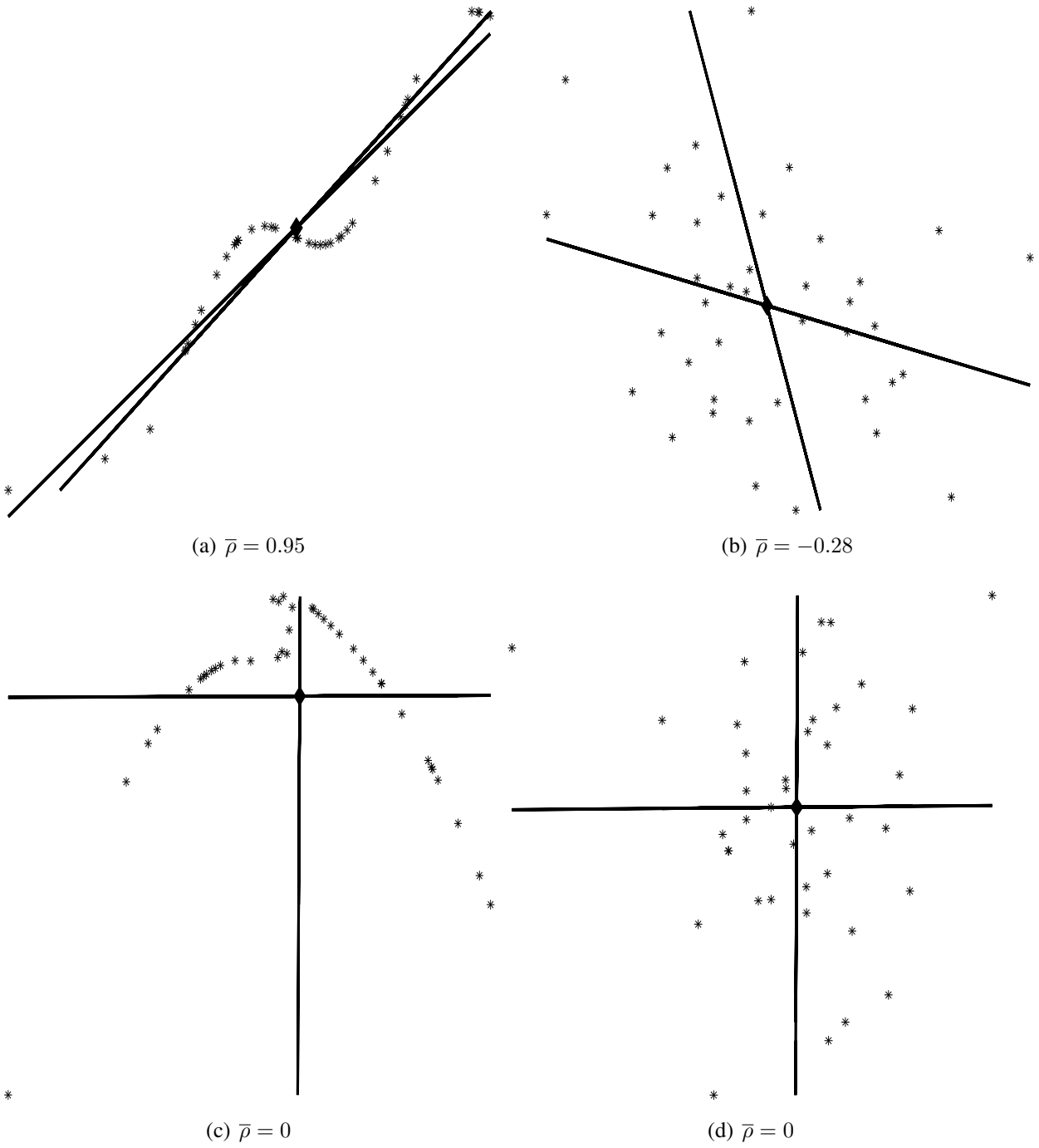
$$\tan\alpha \;=\; \frac{1-\overline{\rho}^2}{\overline{\rho}^2}\cdot\frac{\overline{\sigma}_X\overline{\sigma}_Y}{\overline{\sigma}_X^2+\overline{\sigma}_Y^2}.$$

So, if $\mid \overline{\rho} \mid = 1$, then $\alpha = 0$, i.e. the two lines coincide. If $\mid \overline{\rho} \mid = 0$ (for instance, if X and Y are independent), then $\alpha = \dfrac{\pi}{2}$, i.e. the two lines are perpendicular.

Example 5.11. Let us examine the situations graphed in Figure 7.

- In Figure 7(a) $\overline{\rho} \;=\; 0.95$, positive and very close to 1, suggesting a strong positive linear trend. Indeed, most of the points are on or very close to the line of regression of Y on X. The positivity indicates that large values of X are associated with large values of Y. Also, since the correlation coefficient is so close to 1, the two lines of regression almost coincide.

- In Figure 7(b) $\overline{\rho} = -0.28$, negative and fairly small, close to 0. If a relationship exists between X and Y, it does not seem to be linear. In fact, they are very close to being independent, since the points are scattered around the plane, no pattern being visible. The two lines of regression are very distinct and both have negative slopes, suggesting that large values of X are associated with small values of Y.



(a) $\overline{\rho} = 0.95$

(b) $\overline{\rho} = -0.28$

(c) $\overline{\rho} = 0$

(d) $\overline{\rho} = 0$

Fig. 7: Scattergram, Lines of Regression and Centroid

- In Figure 7(c) $\overline{\rho} = 0$, so the two characteristics are uncorrelated, no linear relationship exists between them. However they are not independent, they were chosen so that $Y = -X^2 + \sin\left(\frac{1}{X}\right)$. Notice also, that the two lines of regression are perpendicular.
- Finally, in Figure 7(d) $\overline{\rho} = 0$, again, so no linear relationship exists. In fact the two characteristics are independent, which is suggested by their random scatter inside the plane.

Remark 5.12. Other types of curves of regression that are fairly frequently used are
– *exponential* regression $y = ab^x$,
– *logarithmic* regression $y = a \log x + b$,
– *logistic* regression $y = \dfrac{1}{ae^{-x} + b}$,
– *hyperbolic* regression $y = \dfrac{a}{x} + b$.