Geostatistics - Review of Probability and Statistics

Julián M. Ortiz - https://julianmortiz.com/

July 20, 2024

Summary

The estimation of a variable in space from nearby samples and the inference of its uncertainty call for a probabilistic framework. Geostatistics relies on the idea of random variables (and random functions) to characterize the unknown values at unsampled locations. In this section, we review the key concepts in probability and statistics that are required to lay out the probabilistic framework used in geostatistics. We will recall concepts such as population, sample, random variable, probability distribution, independence, and many others.

Contents

1 Statistics, population and sample

Statistics is concerned with principles and methods to collect, organize, summarize, display and analyze data, to obtain valid (and general) conclusions and take reasonable decisions based on that analysis. We see its use in everyday life, such as weather forecasts, insurance policy pricing, political polls, etc. Statistics is at the core of the spatial analyses required to quantify resources and reserves in mining applications. In this note, we provide the basic definitions used in statistics that are required to expand these concepts to the analysis of spatial data.

When dealing with data, one needs to keep in mind where the data come from. We define the population as the exhaustive set of measurements or logged features taken over a domain of interest. That is, the population is the collection of all the observations possible for the problem at hand. Obviously, we almost never have access to the population and sometimes, its definition is not completely clear. Therefore, we use a sample, which is a set of measurements or observations that we can actually collect. The hope is that the sample represents the population fairly.

We use statistics to analyze the information provided in the sample to make inference about a population, in the context of prediction and also uncertainty quantification. Statistics can also be used to design the sampling process. In an everyday example, we can think of a national election, where the population is everyone with the right to vote. We may be interested in their preference for President. However, it is impossible to survey everyone in the population to forecast the result of an election, as this would be too ex-

pensive and time consuming. Therefore, a samkple is taken, where a small group of people deemed representative of the population is surveyed and based on their preferences, a forecast is prepared, which will be subject to an error that depends on the sample size, as well as the sample selection process. To define the sample, several considerations must be accounted for. We would like the sample to represent the different groups in the population, therefore, we may try to distribute the observations within different age, socio-economic and geograhic strata to represent the diversity of the entire population. Finding how to weight those strata is challenging and is part of the sample design.

Populations can be characterized with parameters, which are summary numbers that describe the population. Oftentimes, we do not have access to the population to know the parameter (for example, the mean), therefore we need to estimate it. A statistic, on the other hand, is a number that describes the sample. Statistical techniques can be called descriptive, if their purpose is to describe the data. We talk about inferential statistics, when the goal is to draw general conclusions from the data about the population. Statistics are a mean to make inference about the population parameters.

Key Concepts:

Statistics: The body of science concerned with principles and methods to collect, organize, summarize, display and analyze data, to obtain valid (and general) conclusions and take reasonable decisions based on that analysis.

Population: The exhaustive set of measurements or logged characteristics taken over a domain of interest.

Sample: The set of measurements or observations that we collect from the population.

Parameter: A summary number that describes the population.

Statistic: A summary number that describes the sample. It is used as a means to infer the population parameter.

Descriptive or inferential statistics: Statistics can be used to describe the data (descriptive statistics), or to draw general conclusions from the data about the population (inferential statistics).

Inference: The process of reaching a conclusion about the population from the evidence provided by the sample.

2 The data

2.1 Types of data

Data may come in different formats and measure or characterize attributes of different types. We can distinguish the following data types:

- Qualitative or Categorical: this refers to a variable where the attribute is assigned to one of a number of categories, as counts or proportions. Discrete variables can refer to ordered or unordered categories. As an example, we can think of the number of truck loads during a shift in a mine (only integer numbers are possible), or the risk rating of a company (which can take discrete values coded with letters and a number, for example Aa1, Aa2, Aa3, A1, A2, A3, Baa1, Baa2, Baa3, etc). In particular, discrete data can be:
  - Nominal: these data are labelled with a character-string (although the labels can also be numerical codes) with a fixed number of categories. These are usually qualitative attributes that describe some feature of the observation that cannot be measured. A good example of this is the There is no order between the categories or distances to describe similarity. An example of nominal variable is a geological attribute such as the lithology of a rock volume.
  - Binary: binary variables are a special case of nominal data, where the data can only take two cat-

egories, typically complementary, such as yes or no, present or absent, etc.

  - Ordinal: these are also character-string labels, but that have a logical order. The risk rating described above are an example. Color intensities are another example. These ordinal variables are presented in a ranking scale. It should be noted that the categories are not necessarily equidistant, therefore, we should not assume that the difference between the first and second label is the same as that between the second and third category.
  - Continuous variables grouped into bins: in this case, there may be a small number of large bins defining few classes (high, medium and high), or may be smaller bins to "reduce" a continuous variable. Since the underlying variable is continuous, some assumption can be made about the distance between the data from different bins.
  - Integer variables where a distance can be used to measure the difference between observations, but due to the nature of the variable, it can only take integer values. Typical examples are counts, such as the number of truck loads mentioned earlier.

- Quantitative: this refers to the case where variables take numerical values that have a measure of distance associated, allowing comparisons and measuring similarity.

  - Discrete or Counted: in this case, the scale of measurement is not continuous and the variable can only take values in a discrete set.

- Continuous: this refers to a variable that is measured in a continuous scale, that is, where the observation can take any value within one or more ranges of numbers. A simple example is the length of a line or the concentration of zinc in an ore concentrate.

2.2 Univariate data description

The data available can be described in different ways, depending on their nature.

2.2.1 Qualitative data

Qualitative data can be summarized by computing the frequencies of each category, and calculating the proportion which is:

$$\text{proportion} = \frac{\text{count in category}}{\text{total count}} \tag{1}$$

These data can be displayed as a pie chart, where each sector of the chart is colored differently for easy identification and its area represents the proportion corresponding to each category. Pie charts only work when few categories are displayed, so you may need to combine small categories. Ordering the categories by proportion also helps visualizing and processing the information in the pie chart. An alternative to graphically display these data is to use a bar chart, where the height of each bar is equal to the proportion or the frequency of the category.

2.2.2 Quantitative data

Quantitative data can be summarized with different measures. First, there are measures of central tendency, which describe the typical values in the distribution:

- Mean: it is the average of the data. The sample mean is a statistic that helps infer the population mean which is a parameter. The sample mean is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

  where $x_1, x_2, \ldots, x_n$ is the sample of size n from the population. The mean is sensitive to outliers, that is values that are extreme in the distribution. It is a sensitive measure.

- Median: it represents the middle value of the ordered data. This means that 50% of the data fall below the median and 50% at or above this value. The median is not sensitive to outliers, therefore it is called a resistant measure.

- Mode: it is the value that occurs more frequently in the data. In some cases, there is more than one mode.

There are also measures of position, which help understanding the range where a given percentage of the data fall. These require the data to be solrted prior to computing the measure:

- Percentiles (or more generally quantiles): they represent the value at which a given percent (or proportion)

of the data fall under that value. The p^th percentile is the value such that p% of the data is below the percentile and $(100-p)$% is at or above the value. Percentiles are often associated to integer numbers, so the notion of a quantile is introduced to generalize this. For instance, the quantile $q_0.3251$ is the value such that 32.51% of the data is below that value and the remaining 67.49% is at or above it.

- Quartiles: quatriles are particular percentiles of interest. Usually, we identify the lower quartile as the 25^{th} percentile (or $q_0.25$), the median ($q_0.50$) and the upper quartile ($q_0.75$)

Additionally, the data can be described with measures of variability, that look a measuring the spread of the distribution:

- Range: it is the difference between the highest and lowest value in the data. The range is highly sensitive to outliers.

- Interquartile range (IQR): it is the difference between the upper and lower quartile. The IQR is thus a resistant measure.

- Variance and Standard deviation: the variance is the main measure of variability in statistics, although in practice, the standard deviation and coefficient of variation (see the next item) are preferred. The variance is the average squared difference with respect to the mean. It has the units of the original variable squared, which makes it awkward and somehow difficult to use.

The standard deviation is just the square root of the variance, thus it has the same units of the original data and can be used as an error.

- Population variance: this is the population parameter.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{3}$$

  where μ is the population mean and $x_1, x_2, ..., x_N$ are all the observations possible in the population, which has size N.

- Sample variance: this is the statistic that estimates the population variance, when a sample of size n is available.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{4}$$

  Here the sum is divided by $n-1$ to ensure the estimator of the population variance is unbiased. Notice that it also requires an estimate of the population mean.

- Population standard deviation: it is the square root of the population variance.

$$\sigma = \sqrt{\sigma^2} \tag{5}$$

- Sample standard deviation: it is the square root of the sample variance.

$$s = \sqrt{s^2} \tag{6}$$

- Coefficient of variation (CV): it is computed as the standard deviation divided by the mean, providing a unit-free statistic that allows us to compare the spread of different variables in relation to their "center".

$$CV = \frac{s}{\bar{x}} \tag{7}$$

There are many ways to summarize the data distribution and its features using graphs.

The most straightforward way to present the data distribution is showing a histogram, which looks similar to a bar chart, but is used for quantitative data. To create a histogram, data are grouped into bins or class intervals and the frequency (or the relative frequency) is plotted in the y-axis, with the bar width equal to the class interval in the x-axis.

Another way to summarize the distribution is to create a boxplot, which depicts the quartiles of the distribution (lower or first quartile Q1, second quatile or median Q2 and upper or third quartile Q3) and the minimum and maximum values. The three quartiles are signaled by vertical lines within a box, and the extremes are horizontal lines extending left and right. It is common to also display the outliers as dots, in which case, the minimum and maximum values are replaced by a lower and upper limits, calculated as:

$$\text{Lower limit} = Q1 - 1.5 \cdot IQR \tag{8}$$

$$\text{Upper limit} = Q3 + 1.5 \cdot IQR \tag{9}$$

Finally, we should mention that there are other statistics and parameters that characterize the shape of the data distributions:

- Skewness: this parameter (and its corresponding statistic) indicates whether the distribution is symmetric or skewed. It is calculated as the ratio between the average deviation of the observations with respect to the mean to the power three and the standard deviation to the power of three. It is a unit-free parameter:

$$g_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}\right)^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^3}{\sigma^3} \qquad (10)$$

  If the distribution is unimodal, a negative value means the distribution has a long tail to the left, that is few low values and the majority of the observations are concentrated in high values. A value of zero means the distribution is perfectly symmetric. A positive value indicates a long tail to the right, with most values concentrated in the lower values and few observations taking high values.

- Kurtosis: this parameter measures how heavy are the tails with respect to the center of the distribution. It is used to compare symmetric distributions with the normal distribution:

$$\alpha_4 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2\right)^2} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^4}{\sigma^4} \qquad (11)$$

  Kurtosis is often used to compare the shape with a normal distribution by calculating the excess kurtosis, simply subtracting 3 to the kurtosis, since a normal distribution has a kurtosis of 3.

2.2.3 Observations over time

When the observations are associated to time, we can build other displays to analyze the data, including time series plots and statistical control charts. A time series plot is simply a plot of the observation value ordered by time (x-axis). A control chart typically includes three reference values: the mean of the observations which is shown as a solid horizontal line, and two additional dotted lines representing the control limits at the average plus and minus two standard deviations ($\bar{x} - 2s$ and $\bar{x} + 2s$). These control limits are defined because under some circumstances the interval between the control lines represents approximately a 95% interval, thus, points falling outside these limits can be considered unusual.

2.2.4 Representative data

One key assumption about the data that belong to the sample is that they represent the population. However, in practice, there are many issues with data collection, which may lead to a bias.

To ensure that the sample is representative, a probabilistic method should be used for collecting the samples:

- Simple random sampling: the selection is made such that each possible member of the population has an equal chance of being selected. There is also an assumption that no element that does not belong to the population is ever selected (contamination).

- Stratified random sampling: here the population is broken into a number of groups or strata and then random

sampling is performed within each group.

- Regular sampling: this means the selection is done with a consistent rule about the "spacing" of the sample. For example, we select the 1^{st}, 1001^{st}, 2001^{st}, etc. observation for our sample.

2.3 Bivariate data description

It is common that observations include more than one variable. Therefore, it is interesting to see what the relationship between the different variables is. We start by analizing pairs of variables in what is called bivariate analysis.

2.3.1 Qualitative data

If we are dealing with two qualitative attrbutes, we can summarize their relationship in a contingency table. In this table we record the frequencies of particular categories for the first and second atrribute. If we use relative frequencies, we get a total sum of 1, and the sums per row or column are the marginal frequencies, that is the frequencies of a single attribute, without considering the other. These tables can be used to test hypothesis about the effect of one variable over another.

2.3.2 Quantitative data

In the case of quantitative data, a typical summary is provided thorugh a scatter plot, where each attribute is associ-

ated with one axis of the plot, thus each obervation (x_i, y_i) can be represented with a point in this plot.

The notion of variance can be extended by defining the covariance between the two attributes. This is defined as:

$$C(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)\cdot(y_i - \mu_Y) \tag{12}$$

where μ_X and μ_Y are the means of the populations of X and Y.

The linear correlation coefficient also summarizes this bivariate relationship, and provides a measure for the degree of linear relationship between the two attributes. It is defined as the ratio between the covariance and the product of the standard deviations of the two variables:

$$\rho(X,Y) = \frac{C(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)\cdot(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2 \cdot \sum_{i=1}^{N}(y_i - \mu_Y)^2}} \tag{13}$$

The sample correlation coefficient is noted with r.

The covariance in the numerator can take negative values, while the denominator cannot. The correlation coefficient is bound to the interval $[-1, 1]$. $\rho(X,Y) = 1$ if the relationship between X and Y is perfectly linear and has a positive slope, that is, larger X values imply larger Y values (and in this case the increase is perfectly proportional). Similarly, a $\rho(X,Y) = -1$ implies a perfectly linear relationship but with negative slope (Y decreases when X increases). A correlation coefficient of 0 means there is no linear relationship between X and Y, which does not should be interpreted

as lack of relationship. We are just saying that if it exists, this relationship is not linear.

The correlation coefficient may be misleading when data appear in two clusters, or when there are some outliers. It is also important to mention that correlation does not imply causation and that it is quite frequent to find spurious correlations.

2.3.3 Linear regression

We can use the idea of correlation to try to create a predictive model based on a linear relationship bewteen two variables. Linear regression fits a linear model to the relationship between two variables: X (independent variable) acts as input variable to predict Y (dependent variable).

The model is fit by finding the best line to the cloud of points $(x_i, y_i), for i = 1, ..., n$. The best line is defined as the line that minimizes the sum of squared errors between the true and the predicted values for the x_i in the sample data.

This leads to a linear model with two parameters: the intercept $\hat{\beta}_0$ and the slope $\hat{\beta}_1$:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{14}$$

where:

$$
\begin{aligned}
\text{Slope} \quad \hat{\beta}_1 &= \frac{C(X,Y)}{s_x^2} = r(X,Y) \cdot \frac{s_Y}{s_X} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\text{Intercept} \quad \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \cdot \bar{x}
\end{aligned}
$$

Key Concepts:

Qualitative or categorical data: Refers to counts or proportions of a variable assigned to one of a number of categories. These can be nominal (labelled without a particular order), binary (taking one of two possible complementary values), or ordinal (labelled, where the categories have an order). These data can be displayed in a pie chart or a bar chart.

Quantitative data: Referes to variables taking numerical values that represent a measured attribute. These can be discrete or counted, or continuous. These data can be described with a histogram or a boxplot.

Description of quantitative data: These data can be described through measures of central tendency, such as the mean, the median and the mode, measures of position, such as quantiles and quartiles, and measures of variability or spread, such as the range, interquartile range, variance and standard deviation, and coefficient of variation. There are also measures of shape, such as the skewness and kurtosis coefficients.

Observations over time: When observations are linked to time, they can be analyzed with time series plots and control charts.

Representative data: Data collection may be subject to bias. This can be avoided by using a probabilitic method such as simple random sampling, stratified random sampling or regular sampling.

Bivariate data: The relationship between two qualitative variables can be summarized in contingency tables. For quantitative variables, the scatter plot serves to depict the relationship.

Correlation coefficient: This statistic summarizes the linear relationship between two variables, by measuring the defree of linear relationship between the two attributes. it is cimputed as the covariance between the variables divided by the standard deviation of each variable.

Linear regression: Provides a linear model to predict the dependent variable conditioned to the independent variable, through a linear fit that minimizes the sum of squared errors between the true and the predicted values of the dependent variable.

3 Probability

3.1 Random variable and sample space

A random variable is a variable that takes values according to a probability distribution. It provides a description of the values that the variable can take. The probability distribution indicates how likely each value is.

Random variables can be continuous or categorical. The description provided by the probability distribution depends on the type of variable we are dealing with. In the case of a categorical variable, the probability distribution can be a list of the probability of the variable taking each one of its discrete values, or can be a function that describes these probabilities (some categorical random variables can have infinite outcomes, so listing all of the probabilities may not be possible). In the case of a continuous variable, the probability distribution is a function that describes the "probability densities" for each value of the variable.

Formally, a random variable X is a real-valued function defined on a sample space S. In other words, X associates a numerical value with each event (a collection of outcomes)

of an experiment. We can use set operations on these events. Some important concepts are:

- Union: The union of two events A and B, noted $A \cup B$, contains all the outcomes of A and B (this includes those that occur in both).
- Intersection: The intersection of two events A and B, noted $A \cap B$, contains only the outcomes that occur in both A and B.
- Complement: The complement of an event A, noted as $\bar{A}$, A^C or A', contains all outcomes in the sample space that are not in A. Therefore $P(A) = 1 - P(\bar{A})$.
- Mutually exclusive: Two events A and B are mutually exclusive or disjoint, if there are no outcomes in their intersection.
- Empty set: The empty set is an event that contains no outcome and is denoted as $\emptyset$.

Based on these definition, we can deduce that:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

or if A and B are mutually exclusive:

$$P(A \cup B) = P(A) + P(B)$$

There are typical examples to illustrate the concept of a probabilistic experiment:

1. Experiment: Throwing a coin once. Random variable: number of heads ontained. Possible outcomes: 0 or 1.

2. Experiment: Throwing a coin n times. Random variable: number of heads obtained. Possible outcomes: 0, 1, 2,..., or n.

3. Experiment: Throwing a dice. Random variable: top face value. Possible outcomes: 1, 2, 3, 4, 5, or 6.

4. Experiment: Snapping turtles crossing a street in an hour. Random variable: number of turtles. Possible outcomes: 0, 1, 2,..., ∞.

5. Experiment: Taking MINE 467. Random variable: Final grade. Possible outcomes: Any value from 0 to 100%.

Now, in each case, the random variable has a probability distribution associated to the outcomes. Probabilities must comply with some rules (axioms). For example, they cannot be negative and cannot be larger than 1. Also, for a probability distribution to be valid, the sum of the probabilities for all possible outcomes must be 1.

We review these properties in the next section.

3.2 Distribution functions

3.2.1 Definitions and properties

The probability distribution of a random variable can be characterized by its cumulative distribution function (cdf) (notice that we simplify the notation by dropping the subscript X):

$$\forall x \in \mathbb{R}, \quad F_X(x) = F(x) = Prob(X < x) \tag{15}$$

From the cumulative distribution function, we can infer the probability density function (pdf) for continuous variables:

$$\forall x \in \mathbb{R}, \quad f_X(x) = f(x) = \frac{dF(x)}{dx} \tag{16}$$

$$= lim_{dx \to 0} \frac{Prob(x < X < x + dx)}{dx} \tag{17}$$

For categorical variables we use the probability mass function (pmf):

$$\forall i \in \mathbb{N}, \quad P_X(i) = P(i) = Prob(X = i) \tag{18}$$

Probabilities must satisfy some properties:

1. $0 \leq P(A) \leq 1$, for all events A
2. $P(A) = \sum_{\text{all events } a \text{ in } A} P(a)$
3. $P(S) = \sum_{\text{all events } a \text{ in } S} P(a) = 1$

The cdf has the following properties:

- F is a non-decreasing function: if $a < b$, then $F(a) \leq F(b)$.
- $lim_{b \to \infty} F(b) = 1$
- $lim_{b \to -\infty} F(b) = 0$
- $Prob(a < X \leq b) = F(b) - F(a), \quad \forall a < b$

The pdf, on the other hand, has the following properties:

- $Prob(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$
- $Prob(a < X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$
- $Prob(X = a) = \int_{a}^{a} f(x)dx = 0$
- $Prob(X \leq b) = \int_{-\infty}^{b} f(x)dx = F(b)$
- $\frac{dF(a)}{da} = f(a)$

It is clear from the previous definitions that the cdf and the pdf are equivalent. If you know one, you can deduce the other.

3.2.2 Interpretations

Probabilities can be assigned following different interpretations:

1. Classical interpretation (equally likely elementary outcomes):

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of possible outcomes in } S}$$

2. Belief or subjective interpretation: here the interpretation is that the probability reflects some subjective belief which may change among individuals.

3. Empirical interpretation: probabilities are interpreted as the long-run relative frequency of an outcome, that

is, over a large number of repetition of an experiment

$$P(A) \approx \frac{\text{number of outcomes in } A}{\text{number of experiments attempted}}$$

3.2.3 Conditional Probability and Independence

Sometimes one event B may have an effect on the probability of another event A. In that case, the probability of A must be modified to account for the new information (that event B has occured). The concept of conditional probability refers to the probability of one event occurring, given that another event is known to have occurred. This probability is noted as $P(A|B)$. A conditional probability can be computed as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{19}$$

from which we can deduce the probability of the intersection of dependent events:

$$P(A \cap B) = P(B) \cdot P(A|B) \tag{20}$$

The probability of an event without reference to any conditioning event is called the marginal probability.

It shoud be noted that in most cases $P(A|B) \neq P(B|A)$.

We can then define independence between events A and B, as follows:

$$P(A|B) = P(A) \tag{21}$$

This is equivalent to $P(B|A) = P(B)$ and to $P(A \cap B) = P(A) \cdot P(B)$. The interpretation is that the probability of A occurring does not change based on the knowledge of the outcome of B.

3.2.4 Bayes' Theorem

Bayes' theorem allows us to calculate the conditional probability of an event considering prior knowledge of related conditional events.

For two events A and B, we can write:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})} \tag{22}$$

This relationship can be easily computed by realizing that the denominator $P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})$ is equivalent to $P(B)$, which can be shown by realizing that $P(B|A) \cdot P(A) = P(B \cap A)$ and $P(B|\bar{A}) \cdot P(\bar{A}) = P(B \cap \bar{A})$, thus $P(B \cap A) + P(B \cap \bar{A}) = P(B)$.

This fact can be used to do Bayesian inversion, that is inverting the conditional probability $P(A|B)$ into $P(B|A)$ by knowing the prior probabilities $P(A)$ and $P(B)$.

This can be generalized to the case of multiple mutually exclusive events $A_1, A_2, \ldots, A_k$:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_i P(B|A_i) \cdot P(A_i)} \tag{23}$$

3.2.5 Moments

The probability distribution of a random variable can be summarized by looking at different moments (or statistics) of the distribution. For example:

- Expected value
  - Continuous case: $\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$

  - Categorical case: $\mu = E(X) = \sum_{i \in \mathbb{N}} i\, P(i)$

- Variance
  - Continuous case: $\sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2\, f(x)\, dx$
  - Categorical case: $\sigma^2 = Var(X) = \sum_{i \in \mathbb{N}} (i-\mu)^2\, P(i)$

It is interesting to notice that probability and frequency are related. When an experiment is repeated many times, the frequencies will approach the probability distribution. Therefore, a histogram that uses relative frequencies is an experimental representation of the density of the underlying variable and the moments of the random variable can be inferred experimentally from the statistics of that experimental distribution. Similarly, the cumulative histogram represents the cumulative distribution function.

Finally, it is important to point out some properties of the expected value and variance. These properties facilitate calculations for some probability distributions.

If X is a random variable with expected value m and variance σ^2, then:

- $E(aX+b) = aE(X)+b = am+b$, that is, the expectation is a linear operator.
- $Var(aX+b) = a^2 Var(X) = a^2\sigma^2$, this can be deduced from the properties of the expected value.

3.3 Probability distributions

We now look at some simple distributions and apply some of these concepts.

3.3.1 Discrete uniform

Consider a fair dice. The outcome is the number obtained in the face that looks up. Since the dice is fair, each one of the outcomes has equal probability. The possible outcomes are 1, 2, 3, 4, 5, and 6.

The probability distribution of throwing a dice is therefore a discrete uniform distribution between 1 and 6.

The following table provides the outcomes and probabilities:

Table 1: Fair dice outcomes and probabilities.

Outcome	1	2	3	4	5	6
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

The expected value is:

$$\begin{aligned} m = E(X) &= \sum_{i \in \mathbb{N}} iP(i) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= 3.5 \end{aligned}$$

The variance is:

$$\begin{aligned}\sigma^2 = Var(X) &= \sum_{i \in \mathbb{N}} (i-m)^2 P(i) \\ &= (1-3.5)^2 \cdot \frac{1}{6} + (2-3.5)^2 \cdot \frac{1}{6} + (3-3.5)^2 \cdot \frac{1}{6} \\ &+ (4-3.5)^2 \cdot \frac{1}{6} + (5-3.5)^2 \cdot \frac{1}{6} + (6-3.5)^2 \cdot \frac{1}{6} \\ &= 2.9167\end{aligned}$$

3.3.2 Bernouilli distribution

Let us consider now a binary random variable X_i, where one of the outcomes has a probability p of occurring. We can call this event "success". The complement of this event (we can call it "failure") has a probability $1-p$.

This random variable has a Bernouilli distribution, where the mean and variance are given by:

$$\mu = E(X_i) = p \tag{24}$$

$$\sigma^2 = Var(X_i) = p(1-p) \tag{25}$$

We can also call this variable an indicator variable.

3.3.3 Binomial distribution

When an event is repeated multiple times, and each time the result is independent of the previous experiments, this is noted as "independent and identically distributed" random variables (or i.i.d.).

We can now consider n Bernouilli trials X_i drawn independently. The probability of X defined as the number of successes in n identical trials follows a binomial distribution with parameters:

$$\mu = E(X) = np \tag{26}$$

$$\sigma^2 = Var(X) = np(1-p) \tag{27}$$

$$\sigma = SD(X) = \sqrt{np(1-p)} \tag{28}$$

The probability mass distribution for X is:

$$f(x) = P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad \text{for } x = 1, 2, \ldots, n \tag{29}$$

Recall that the factorial operation is $x! = x \cdot (x-1) \cdot \ldots \cdot 2 \cdot 1$ and $0! = 1$.

3.3.4 Continuous uniform

Consider now the case of a continuous random variable that can take values according to a uniform distribution within the range [0,1].

The expected value and variance are:

$$m_Z = E(Z) = \int_{-\infty}^{\infty} z\, f(z) dz \tag{30}$$

$$= \int_0^1 z\, \frac{1}{1}\, dz \tag{31}$$

$$= \frac{z^2}{2}\Big|_0^1 = \frac{1}{2} \tag{32}$$

$$\sigma_Z^2 = Var(Z) = \int_{-\infty}^{\infty} (z-m)^2\, f(z) dz \tag{33}$$

$$= \int_0^1 (z - \frac{1}{2})^2\, \frac{1}{1}\, dz \tag{34}$$

$$= \frac{1}{12} \tag{35}$$

We should be able to compute the same parameters in the case that the uniform distribution can take values within an arbitrary range $[a, b]$.

3.3.5 Normally distributed variable

Consider a variable that is normally distributed, that is, one that follows the Gaussian distribution:

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} exp^{-\frac{1}{2}(\frac{z-\mu}{\sigma})^2} \tag{36}$$

The deduction of the mean and variance from the probability distribution function is not simple, so we will just state

that the mean and variance are the two parameters in the pdf expression:

$$\mu_Z = E(Z) = \mu \tag{37}$$
$$\sigma_Z^2 = Var(Z) = \sigma^2 \tag{38}$$

This means that the probability distribution of a normal random variable requires only two parameters to be fully defined. The mean defines its center, and the variance determines its spread. It is noted $N(\mu, \sigma^2)$The distribution is symmetric and ranges from $-\infty$ to $+\infty$, although values in the tails beyond $\mu \pm 3\sigma$ have a very low probability.

A particular and important case of the normal distribution is the so called standard normal distribution (or $N(0, 1)$), which has a mean of 0 and variance of 1.

$$f(y) = \frac{1}{\sqrt{2\pi}} exp^{-\frac{1}{2}y^2} \tag{39}$$

We mentioned that dealing with the pdf of the normal distribution is tricky when solving integrals to calculate its parameters or probabilities of not exceeding a particular value. Recall that we can compute this probability is equivalent to calculating the cdf for that particular value:

$$\begin{aligned} P(Z \leq z) = P(Z < z) &= \int_{-\infty}^{z} f(z)\, dz \\ &= \int_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\pi}} exp^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}\, dz \end{aligned}$$

In order to compute this value, we can standardize the variable and reset its mean to 0 and its variance o 1, by defining:

$$Y = \frac{Z - \mu}{\sigma} \tag{40}$$

Notice that this operation can be done with any random variable, even if it is not normally distributed. The resulting variable will have a mean of 0 and variance of 1, but will only follow a standard normal distribution, if the original variable was normally distributed.

We can compute the probability of not exceeding a value for the standard normal distribution:

$$\begin{aligned} P(Y \leq y) = P(Y < y) &= \int_{-\infty}^{y} f(y)\, dy \\ &= \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} exp^{-\frac{1}{2}y^2}\, dy \end{aligned}$$

Again, this integral cannot be easily obtained, therefore, the probabilities are tabulated for the standard normal distribution, which allows us to infer $P(Z \leq z)$ for any normal distribution with parameters μ and σ^2, by using the standardization presented earlier.

A general rule for distributions that are close to normal, is to use the compute probability intervals associated to the parameters of the distribution. Typically, the following approximations are used:

- 68% of the observations are within the range defined by the mean plus or minus one standard deviation.

- 95% of the observations are within the range defined by the mean plus or minus two standard deviations.
- 99.7% of the observations are within the range defined by the mean plus or minus three standard deviations.

3.3.6 Other important probability distributions

Student's t-distribution

This distribution is similar to the normal distribution, but with heavier tails. This distribution depends on a single parameter called degrees of freedom and stems from the estimation of the mean of a normallu distributed variable, from a small sample of size n. The random variable $T_{n-1} = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a t-distribution with $n-1$ degrees of freedom. The expression for the pdf and cdf are quite complex, but the relevant tables to determine the probability values are easily obtained in books or software.

Chi-square distribution

This distribution (also noted χ^2 distribution) is a right-skewed distribution that also depends on a single parameter called degree of freedom. It arises as the distribution of a random variable defined as the sum of squares of n independent standard normal random variables. The random variable $\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$ has a χ^2 distribution with $n-1$ degrees of freedom. We also omit the expressions for the pdf and cdf.

F distribution

This distribution is also right-skewed, but depends on

two parameters (both referred as degrees of freedom). It is used in ANOVA (analysis of variance).

Poisson distribution

This distribution serves to describe rare events, that is, events that have a small probability of occurrence. It defines the probability of a number of these events occur in a fixed interval of time, assuming they occur with a constant rate and are independent of the last occurrence. Its pdf is: $f(x) = \frac{e^{\lambda}\lambda^{x}}{x!}$, where x is the number of occurrences and the parameter λ is the mean and variance of the variable.

Key Concepts:

Event: A collection of outcomes.

Random variable: It is a variable that takes values according to a probability distribution. It associates a numerical value with each event of an experiment.

Probability mass function (pmf): It is a mathematical function, noted $f(x)$ that provides the probability for possible outcomes of a discrete random variable. In this case $f(x) = Prob(X = x)$.

Probability density function (pdf): It is a mathematical function, also noted $f(x)$, that provides the probability for possible outcomes of a continuous random variable. In this case $f(x) \neq Prob(X = x)$.

Cumulative distribution function (cdf): It is noted $F(x)$ and represents the probability that the random variable X is less than or equal to the value x: $F(x) = Prob(X \leq x)$

Sample space: This is the set of possible outcomes of a random experiment.

Set operations: Events of a random experiment can be combines through several set operations such as union, intersection, and complement.

Conditional probability: It refers to the probability of one event occurring, given that another event is known to have occurred.

Independence: The lack of dependence between events occurs when the conditioning event does not affect the probability of the event considered.

Bayes' theorem: It allows us to calculate the conditional probability of an event considering prior knowledge of related conditional events.

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_i P(B|A_i) \cdot P(A_i)}$$

Probability distributions: There are some important distributions, including uniform, Bernouilli, binomial, and the normal distribution.

4 Sampling distributions

As stated before, a parameter is a numerical feature of the population. However, we do not have access to the entire population and we only can obtain a limited sample. From this sample, we can calculate statistics that may help us make inference of the parameters of the population. A statistic has a probability distribution. This distribution is called a sampling distribution. Since the sample may change from one experiment to the next, by repeating the experiment, we can obtain multiple results for the statistic. In

general, we will consider a random sample of a fixed sixe from the population.

One important parameter we want to infer is the population mean. This parameter can be approximated using the sample mean:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \tag{41}$$

If the variable X_i is drawn from a distribution with population mean μ and population variance σ^2, that is, if these are independent and identically distributed events, then the mean $\bar{X}$ of a sample of size n has a distribution (the sampling distribution) with the following parameters:

$$E(\bar{X}) = \mu \tag{42}$$

$$Var(\bar{X}) = \frac{\sigma^2}{n} \tag{43}$$

In words, the sample mean is centered in the population mean (so it is useful to infer this parameter of the population) and its spread bcomes smaller as the number of observations increases, that is, the distribuion of the sample mean is more concentrated around the population mean than the population distribution.

Now, if the observations X_i are drawn from a normal distribution, then, the distribution of the sample average is also normal, with mean μ and variance $\frac{\sigma^2}{n}$, that is, $N(\mu, \sigma^2/n)$.

Now an interesting fact is that whatever the distribution of the population we are drawing the samples from, the distribution of the sample mean is approximately normal, when

the sample size is large. This is known as the Central Limit Theorem. This implies that:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \tag{44}$$

We will consider a sample large, when $n > 30$.

When a statistic is used to estimate a population parameter, it is called a point estimator, or simply an estimator. The statndard deviation of this estimator is called standard error, SE. The sample mean is a point estimator of the population mean. In order to know its accuracy in estimating the population mean, we need to determine the standard error of $\bar{X}$. Since we know that the standard deviation of $\bar{X}$ is $\sqrt{\sigma^2/n} = \sigma/\sqrt{n}$.

This standard error can be used to determine the probability that the error will be between some values. For instance, we can say that there is approximately a 95% probability that the error will be within $\pm 2SE$ of the population mean. An estimate will usually (and should) be reported with an associated error. It should be explicit what error is reported: the SE, $2SE$ or some other error associated to a probability.

5 Bibliographical notes

The concepts covered in this chapter are just a small portion of what can be found in any probability and statistics book. It should be noted that most statistics books deal with mathematical statistics and do not emphasize data analysis. An excellent textbook that covers probability concepts

is “A first course in probability” by S. Ross [2]. For mathematical statistics, many introductory books can be recommended, however a more comprehensive approach can be found in “Applied multivariate statistical analysis” by R.A. Johnson and D.W.Wichern [1].

References

[1] Johnson, R. A., and Wichern, D. W. *Applied Multivariate Statistical Analysis*, 6th ed. Pearson, 2008.

[2] Ross, S. M. *A First Course in probability*, 8th ed. Pearson Prentice Hall, 2010.

Index