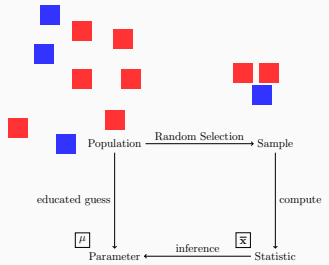

Figure 1: Statistical inference means inferring something about a population using information from samples [Has22].

Population includes all of the elements from a set of data. Sample consists of one or more observations from the population.

Parameter Characteristic of a distribution describing a population, such as the mean or standard deviation of a normal distribution. Often notated using Greek letters.

Statistics: A numerical value that represents a property of a random sample. Examples of statistics are

- the mean value of the sample data.
- the range of the sample data.
- deviation of the data from the sample mean.

Machine learning and statistics are both concerned with the construction of models to explain data. In statistics data is observed and we try to figure out a process that explains the data. In machine learning we look for a best-fit [DFO20].

Sample space Ω
The samples space contains all possible outcomes of an experiment. A coin toss, for example, can have two outcomes heads (h) or tails (t). Which leads to the set $\{h, t\}$. Two sucessive tosses generate the larger space $\{hh, tt, ht, th\}$.

Event space $\mathcal{A}$
A set of events, an event is a set of outcomes from the sample space.

Probability P
With each event $\mathcal{A}$ we associate a number P($\mathcal{A}$). This number measures the probability that the event will occur.

Lets consider a hypothetical bag. The bag contains dollar \$ and euro € coins. Of ten coins in the bag four are euro coins. If we draw twice the probability space will be the set $\{€€, \$€, €\$, \$\$\}$. We can now construct a discrete function for the random variable X. $X(€, €) = 2$, $X(€, \$) = 1$, $X(\$, €) = 1$, $X(\$, \$) = 0$

$$P(X = 2) = P((€, €)) = P(€) \cdot P(€) = 0.4 \cdot 0.4 \qquad = 0.16 \quad (1)$$

$$P(X = 1) = P((€, \$) \cup P((\$, €)) = P((€, \$) + P((\$, €))$$

$$= 0.4 \cdot (1 - 0.4) + (1 - 0.4) \cdot 0.4 \qquad = 0.48 \quad (2)$$

$$P(X = 0) = P(\$) \cdot P(\$) = 0.6 \cdot 0.6 \qquad = 0.36 \quad (3)$$

Marginal and conditional probability

We can compute marginal probabilities by summing rows or columns.

$$P(X = x_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{3} n_{ij}}{N} \tag{5}$$

$$P(X = y_1) = \frac{r_j}{N} = \frac{\sum_{i=1}^{3} n_{ij}}{N} \tag{6}$$

The marginal probabilities allow us to define conditional probability:

$$P(Y = y_i | X = x_i) = \frac{n_{ij}}{c_i} \tag{7}$$

$$P(X = x_i | Y = y_i) = \frac{n_{ij}}{r_j} \tag{8}$$

Made up data:

		≤ 1 h	> 1 h
y_1	Bonn	0.24	0.08
y_2	Cologne	0.04	0.32
y_3	Siegburg	0.16	0.16
		x_1	x_2

The table entries contain normalized frequencies. Row and column frequency sums are the so-called marginal probabilities.

$$\Rightarrow P(X = x_1) = 0.44. \tag{9}$$

$$\Rightarrow P(Y = y_2) = 0.36. \tag{10}$$

Probability for more than one hour from cologne:

$$P(X = x_1 | Y = y_2) = 0.32/(0.32 + 0.04) = 0.89 \tag{11}$$

Bayes Law [DFO20]

Sometimes, we have no direct way of observing a property. We are forced to infer knowledge indirectly. In such cases, Bayes law helps. Bayes states

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}. \tag{26}$$

The law is a consequence of our ability to factorize distributions as $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$. If we cant observe $\mathbf{x}$ directly, we may have expectations of its distribution $p(\mathbf{x})$, and the likelihood $p(\mathbf{y}|\mathbf{x})$. Bayes allows us to find a posterior $p(\mathbf{x}|\mathbf{y})$ given evidence $p(\mathbf{y})$.

Say 50 in 100k people of a population have a given illness.
We have $P(S) = 0.0005$, and $P(H) = 1 - (50/100000) = 0.9995$. We have a test that detects the disease with an accuracy of 98%. In other words, $P(T|S) = 0.98$. Unfortunately, it also yields a positive result for 1% of healthy people $P(T|H) = 0.01$. What happens if we use the test to look for the disease in the general population?

$$P(S|T) = \frac{P(T|S)P(S)}{P(T|S)P(S) + P(T|H)P(H)} \tag{27}$$

$$= \frac{0.98 \cdot 0.0005}{0.98 \cdot 0.0005 + 0.01 \cdot 0.9995} = 0.05 \tag{28}$$

In this case, testing is probability not a great idea. Note, total probability for exclusive events: $P(T) = P(T|S)P(S) + P(T|H)P(H)$ if H is not S.

Gaussian mixture models

A Gaussian mixture model has the density [McN16]

$$f(\mathbf{x}|\theta) = \sum_{g=1}^{G} \rho_g \phi(\mathbf{x}|\mu_g, \Sigma_g). \tag{35}$$

With the normal distribution ϕ defined as before. ρ_g denotes the global probability with which a data value could originate from gaussian g. The gs number the gaussians, and G is the total number of Gaussians in the mix. We will use two. ϕ denotes the Gaussian function. Parameters μ_g and Σ_g are mean vector and covariance matrix.

Typically we want as many g as we have classes in the data. I.e. one for healthy and one for diabetic. The data vectors are p dimensional $\mathbf{x} \in \mathbb{R}^p$. Sampling $\phi(\mathbf{x})$ tells us how likely it was to see the point we have. Big values mean it was likely small mean it was not.

Likelihood

Likelihood models the probability of data originating from a distribution as a function of the parameters. The gaussian case is modelled by [McN16]

$$\mathcal{L}_c(\theta) = \prod_{i=1}^{n} \prod_{g=1}^{G} [\rho_g \phi(\mathbf{x}_i | \mu_g, \Sigma_g)]^{z_{ig}}. \tag{36}$$

We want to maximize the likelihood.
In other words, we want to transform the bells in such a way, that they explain the points as plausible as possible.

To maximize *phi* it needs to sit on top of the points it labels. When a gaussian sits on top of many points it's ρ_g should be large. Finally, when this works well we want a big weight from z_{ig}.

After guessing an initial choice for all $\hat{\mu}_g$ and $\hat{\Sigma}_g$ [McN16],

$$\hat{z}_{ig} = \frac{\rho_g \phi(\mathbf{x_i}|\hat{\mu}_g, \hat{\Sigma}_g)}{\sum_{h=1}^{G} \rho_h \phi(\mathbf{x_i}|\hat{\mu}_h, \hat{\Sigma}_h)} \tag{38}$$

tells us the probability with which point x_i came from gaussian g. It creates an association between the data points and the Gaussians. Numerically evaluation results in a matrix $\mathbf{Z} \in \mathbb{R}^{G \times n}$. Use the maxima in it's output to select the points which belong to each class.

The z_{ig} are the true labels, $\hat{z}_{ig}$ is our estimation. The $\hat{z}_{ig}$ are the expected value of the complete data log-likelihood. Why? ϕ is a pdf. A pdf can be interpreted as providing a relative likelihood that the value of the random variable would be equal to that sample [1]. We ask for all gaussians and every point and normalize.