1 Probability Theory

A probability is a measure of how frequent or likely an event will take place.

Probability interpretations

Frequentist: Fraction of positive samples, if we measured infinitely many samples.

Objectivist: Probabilities are due to inherent uncertainty properties.

Subjectivist: An agent's degree of belief (not external).

Bayesian: (Building on subjectivism) A reasonable expectation on the basis of a state of knowledge/evidence.

! → Also the frequentist view is subjective since you need to compare events on otherwise similar objects. Usually there are no completely similar objects, so you need to define them.

Q → The Bayesian view allows to give certainties to events, where we don't have samples on (e.g. disappearance of the south pole until 2030).

Probability Space The probability space is a triplet space containing a sample/outcome space Ω (containing all possible atomic events), a collection of events S (containing a subset of Ω to which we want to assign probabilities) and the mapping P between Ω and S.

Axioms of Probability The mapping P must fulfill the axioms of probability:

1. $P(a) \geq 0$
2. $P(\Omega) = 1$
3. $a, b \in S$ and $a \cap b = \{\} \Rightarrow P(a \cup b) = P(a) + P(b)$

Random Variable (RV) A RV is a function that maps points from the sample space Ω to some range (e.g. Real numbers or booleans). They are characterized by their distribution function. E.g. for a dice roll:

$$X(\omega) = \begin{cases} 0, \text{ if } \omega = heads \\ 1, \text{ if } \omega = tails. \end{cases}$$

Proposition A Proposition is a conclusion of a statistical inference that can be true or false (e.g. a classification of a datapoint). More formally: A disjunction of events where the logic model holds. An event can be written as a **propositional logic model**:
$A = true, B = false \Rightarrow a \wedge \neg b$. Propositions can be continuous, discrete or boolean.

1.1 Probability distributions(PDF)

Probability distributions assign probabilities to to all possible points in Ω (e.g. $P(Weather) = \langle 0.3, 0.4, 0.2, 0.1 \rangle$, representing Rain, sunshine, clouds and snow). Joint probability distributions give you a probability for each atomic event of the RVs (e.g. $P(weather, accident)$ gives you a 2×4 matrix.)

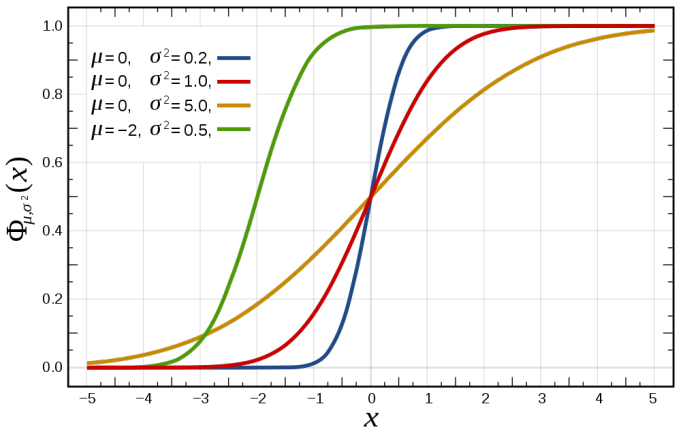

Fig. 1: Cumulative distribution function of a normal distribution for different mean (μ) and variance (σ). *Source: user Inductiveload on wikimedia.org.*

Cumulative Distribution Function (CDF) The CDF is defined as $F_X(x) = P(X \leq x)$ (See figure 1).

Probability Density Function (PDF) For continuous functions the PDF is defined by

$$p(x) = \frac{d}{dx} p(X \leq x). \tag{1}$$

The probability of x being in a finite interval is

$$P(a < X \leq b) = \int_a^b p(x)dx \tag{2}$$

A PDF is shown in figure

Properties of Distributions

The **expected value** (E) or **mean** (μ) is given by $E[X] = \sum_{x \in X} x * p(x)$ for discrete RVs and $E[X] = \int_X x * p(x)dx$ for continuous RVs.

The **variance** measures the spread of a distribution: $var[X] = \sigma^2 = E[(X - \mu)^2] = E[X]^2 - \mu^2$.

The **standard deviation** is given by: $\sqrt{var[X]} = \sigma$.

The **mode** is the value with the highest probability (or the point in the PDF with the highest value):

The **median** is the point at which all point less than the median and all points greater than the median have the same probability (0.5).

The **quantiles** (Q) divide the datapoints into sets of equal number. The Q_1 quartile has 25% of the values below it.

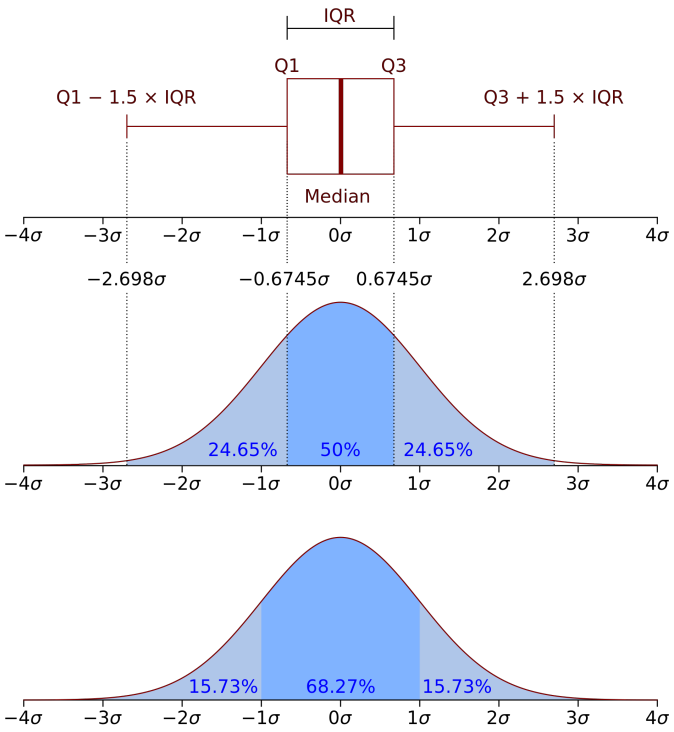

Fig. 2: Probability density function of a normal distribution with variance (σ). In red a range from a Box-plot is shown with quartiles (Q1, Q3) and interquartile range (IQR). For the cutoffs (borders to darker blue regions) the IQR (on top) and σ are chosen. Another common cutoff is the confidence interval with light blue regions having a probability mass of $2 * \alpha/2$. *Source: user Jhguch on wikimedia.org.*

Dirac delta function is a function that is infinite at one point and 0 everywhere else:

$$\delta(x) = \begin{cases} \infty, & \text{if } x = 0 \\ 0, & \text{if } x \neq 0 \end{cases} \qquad \text{and} \int_{-\infty}^{\infty} \delta(x)dx = 1$$

1.1.1 Uniform distribution

The uniform distribution has the same probability throughout a specific interval:

$$\text{Unif}(a, b) = \frac{1}{b-a}\mathbb{1}(a < x \leq b) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a, b] \\ 0, & \text{else} \end{cases}$$

$\mathbb{1}$ is a vector of ones.

1.1.2 Discrete distributions

Used for random variables that have discrete states.

Binomial distribution Used for experiments with two outcomes (e.g. coin flips).

$$X \sim \text{Bin}(n, \theta) \quad ,\text{Bin}(k|n, \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \quad ,\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

where n is the number of total experiments, k is the number of sucessful experiments and θ is the probability of success of an experiment.

Bernoulli distribution Is a special case of the binomial distribution with $n = 1$.

$$X \sim \text{Ber}(\theta) \quad ,\text{Ber}(x|\theta) = \theta^{\mathbb{1}(x=1)}(1-\theta)^{\mathbb{1}(x=0)} = \begin{cases} \theta, & \text{if } x = 1 \\ 1-\theta, & \text{if } x = 0 \end{cases}$$

Multinomial distribution Used for experiments with k different outcomes (e.g. dice rolls).

$$\text{Mu}(x|n, \theta) = \binom{n}{x_1, ..., x_K} \prod_{i=1}^{K} \theta_j^{x_j} = \frac{n!}{x_1!, ..., x_k!} \prod_{i=1}^{K} \theta_j^{x_j},$$

where k is the number of outcomes, x_j is the number times that outcome j happens. $X = (X_1, ..., X_K)$ is the *random vector*.

Multinoulli distribution Is a special case of the multinomial distribution with $n = 1$. The random vector is then represented in *dummy-* or *one-hot-encoding* (e.g. $(0, 0, 1, 0, 0, 0)$ if outcome 3 takes place).

$$\text{Mu}(x|1, \theta) = \prod_{j=0}^{K} \theta_j^{\mathbb{1}(x_j=1)}$$

Empirical distribution

$$\mathsf{p}_{\mathsf{emp}}(A) = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}(A), \quad \delta_{x_i} = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases},$$

w where $x_1, ..., x_N$ is a data set with N points. The points can also be weighted:

$$p(x) = \sum_{i=1}^{N} w_i \delta_{x_i}(x)$$

1.1.3 Continuous distributions

Used for random variables that have continuous states.

Normal/Gaussian distribution Often chosen for random noise because it is simple and needs few assumptions (see sect. 1.1.4). The PDF is given by:

$$p(x|\mu\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$$

where μ is the mean and σ^2 is the variance. The CDF is given by:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{x} e^{\frac{-t^2}{2}} dt$$

Multivariate normal/Gaussian distribution For T datapoints with k dimensions (features). The pdf is:

$$p(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left[-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right],$$

where x now has multiple dimension $(x_1, x_2, ..., x_k)$ and Σ is the $k \times k$ covariance matrix: $\Sigma = \mathsf{E}[(X-\mu)(X-\mu)]$. The covariance between features is: $\mathsf{Cov}[X_i, X_j] = \mathsf{E}[(X_i - \mu_i)(X_j - \mu_j)]$

Beta distribution defined for $0 \leq x \leq 1$ (see figure 3). The pdf is:

$$f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

The beta function is there to normalize and ensure that the total probability is 1.

Dirichlet distribution The multivariate version of the Beta distribution (see fig. 4). The PDF is:

$$\mathsf{Dir}(x|\alpha) \triangleq \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \quad \sum_{i=1}^{K} x_i = 1, \quad x_i \geq 0 \ \forall i$$

Marginal distributions Are the probability distributions of subsets of the original distribution. Marginal distributions of normal distributions are also normal distributions (see fig. 5).

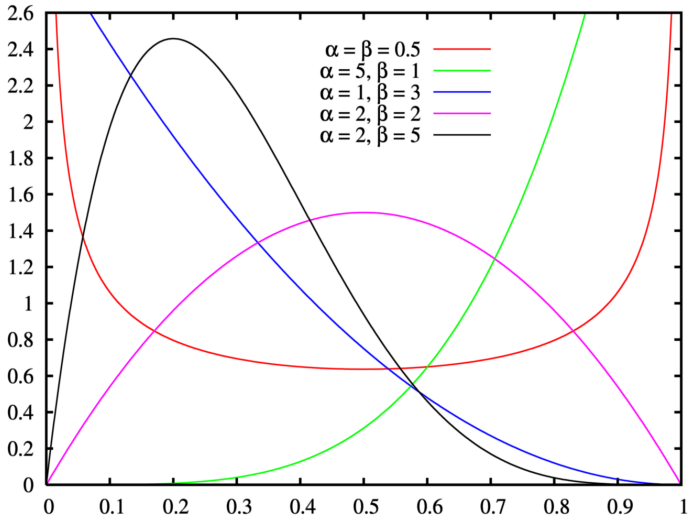

Fig. 3: Probability density function of a beta-distribution with different parameter values. *Source: user MarkSweep on wikimedia.org.*

1.1.4 Central limit theorem

In many cases the sum of random variables will follow a normal distribution as n goes to infinity.

1.2 Conditional/Posterior Probability

Expresses the probability of one event (Y) under the condition that another event (E) has occurred. (e.g. $C =$ "gets cancer", $S =$ "is a smoker" $\rightarrow p(C|S) = 0.2$, meaning: "given the *sole information* that someone is a smoker, their probability of getting cancer is 20%.")

The conditional probability can be calculated like this:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)} = \alpha P(A, B),$$

where α is used as a normalization constant. If you have hidden variables (confounding factors) you need to sum them out like so:

$$P(Y|E = e) = \alpha P(Y, E = e) = \alpha \sum_h P(Y, E = e, H = h)$$

where X contains all variables, Y is called *query variable*, E is called *evidence variable*, $H = X - Y - E$ is called *hidden variable*. You get the joint probabilities by summing out the

! → Usually $p(A|B) \neq p(B|A)$

! → Priors are often forgotten: E.g. P("COVID-19") is confused with P("COVID-19"|"Person is getting tested") (because only people with symptoms go to the testing station).

! → Base rate neglect: Under-representing prior probability. E.g. You have a test with a 5% false positive rate and a incidence of disease of 2% in the population. If you are tested positive in a population screening your probability of having the disease is only 29%.

🔍 → Conditional distributions of Gaussian distributions are Gaussian distributions themselves.

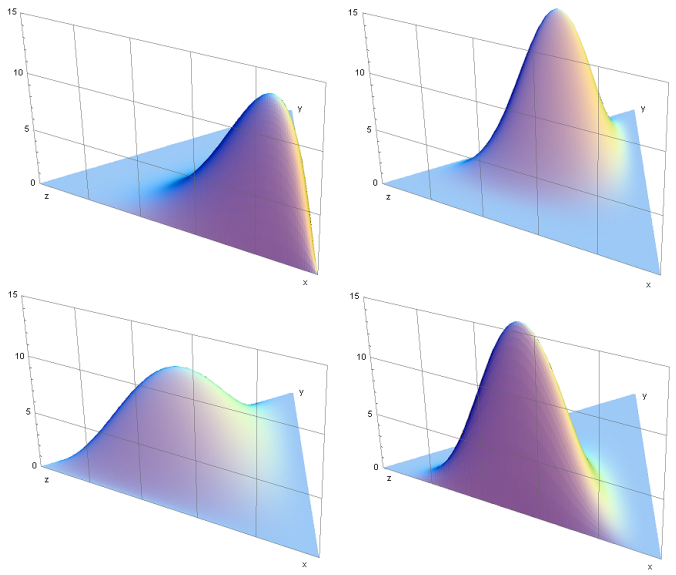

Fig. 4: Probability density function of a Dirichlet-distribution on a 2-simplex (triangle) with different parameter values. Clockwise from top left: α = (6,2,2), (3,7,5), (6,2,6), (2,3,4). *Source: user ThG on wikimedia.org.*

Independence For independent variables it holds: $P(A|B) = P(A)$ or $P(B|A) = P(B)$

Conditional independence Two events A and B are independent, given C: $P(A|B,C) = P(A|C)$. A and B must not have any information on each other, given the information on C. E.g. for children: $P(\text{"vocabulary"}|\text{"height"},\text{"age"}) = P(\text{"vocabulary"}|\text{"age"})$.

1.2.1 Bayes Rule

Bayes rule:

$$P(\text{hypothesis}|\text{evidence}) = \frac{P(\text{evidence}|\text{hypothesis})P(\text{hypothesis})}{P(\text{evidence})}$$

often used as:

$$P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model})P(\text{model})}{P(\text{data})}$$

Terminology:

- $P(\text{hypothesis}|\text{evidence})$ = Posterior
- $P(\text{evidence}|\text{hypothesis})$ = Likelihood
- $P(\text{hypothesis})$ = Prior (How probable hypothesis was before seeing evidence)
- $P(\text{evidence})$ = Marginal (How probable evidence is under all possible hypotheses)
- $\frac{P(\text{evidence}|\text{hypothesis})}{P(\text{evidence})}$ = Support B provides for A

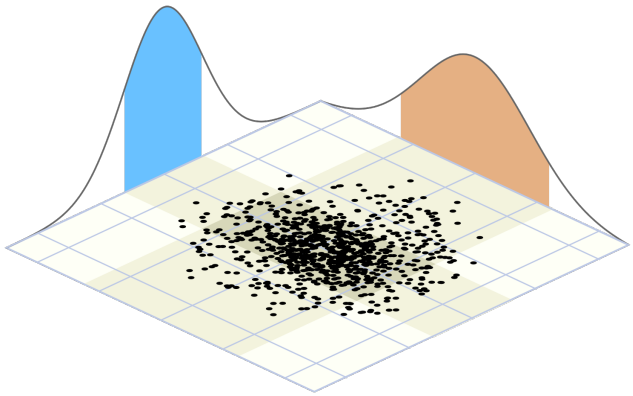

Fig. 5: Data following a 2D-Gaussian distribution. Marginal distributions are shown on the sides in blue and orange. *Source: user Auguel on wikimedia.org.*

- $P(\text{data}|\text{model})P(\text{model}) = \text{joint probability } (P(A,B))$

The proof (see above): $P(A|B)P(B) = P(A,B) = P(B|A)P(A)$

Example for Bayes Rule using COVID-19 Diagnostics

$$P(\text{COVID-19}|\text{cough}) = \frac{P(\text{cough}|\text{COVID-19})P(\text{COVID-19})}{P(\text{cough})} = \frac{0.7 * 0.01}{0.1} = 0.07$$

Estimating $P(\text{COVID-19}|\text{cough})$ is difficult, because there can be an outbreak and the number changes. However, $P(\text{cough}|\text{COVID-19})$ stays stable, $P(\text{COVID-19})$ and $P(\text{cough})$ can be easily determined.

1.3 Further Concepts

Convergence in Probability of Random Variables You expect your random variables (X_i) to converge to an expected random variable X. I.e. after looking at infinite samples, the probability that your random variable X_n differs more than a threshold ϵ from your target X should be zero.

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

Bernoulli's Theorem / Weak Law of Large Numbers

$$\lim_{n\to\infty} P(|\frac{\sum_{i=1}^{n} X_i}{n} - \mu| > \epsilon) = 0,$$

where $X_1, ..., X_n$ are independent & identically distributed (i.i.d.) RVs. $\Rightarrow$ With enough samples, the sample mean will approach the true mean. The **strong law of large numbers** states that $|\frac{\sum_{i=1}^{n} X_i}{n} - \mu| < \epsilon$ for any $\epsilon > 0$.

Bias of an estimator You have a model with a parameter $\hat{\theta}$ that is an estimator for the true θ. You want to know whether your model over- or underestimates the true θ systematically.

$$\text{Bias}[\hat{\theta}] = \mathsf{E}_{X|\mathcal{D}}[\hat{\theta}] - \theta$$