Probability theory and statistics overview

Erik Pärlstrand

November 2018

1 Probability theory

1.1 Random variable

A random variable is a variable whose possible values are outcomes of a random phenomenon.

A random variable has a probability distribution, which specifies the probability of its values. Random variables can be discrete with a probability mass function; or continuous, via a probability density function.

1.2 Distribution measures

Expected value

The expected value is an anticipated value for a random variable.

The expected value for a continuous random variable is:

$$E[X] = \int_{-\infty}^{\infty} xp(x)dx$$

The expected value for a discrete random variable is:

$$E[X] = \sum_{n=-\infty}^{\infty} np(n)$$

Variance

Variance is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of (random) numbers are spread out from their average value.

It is defined by:

$$V[X] = E[(X-\mu)^2] = E[X^2] - E[X]^2$$

The standard deviation is defined as:

$$\sigma[X] = \sqrt{V[X]}$$

Skewness

Skewness is a measure of the asymmetry of the probability distribution of a random variable about its mean. The skewness value can be both positive and negative.

It is defined by:

$$S[X] = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^3\Big]$$

Kurtosis

Kurtosis is a measure of the ”tailedness” of the probability distribution of a random variable. In a similar way to the concept of skewness, kurtosis is a descriptor of the shape of a probability distribution.

It is defined by:

$$K[X] = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^4\Big]$$

Covariance and correlation

The covariance between two random variables X and Y is defined by:

$$C[X,Y] = E[XY] - E[X]E[Y]$$

The correlation between two random variables X and Y is defined by:

$$\rho[X,Y] = \frac{C[X,Y]}{\sqrt{V[X]V[Y]}}$$

Note that if the correlation between X and Y is zero does not imply that they are independent. X and Y are independent i.f.f. $p(y|x) = p(y)$.

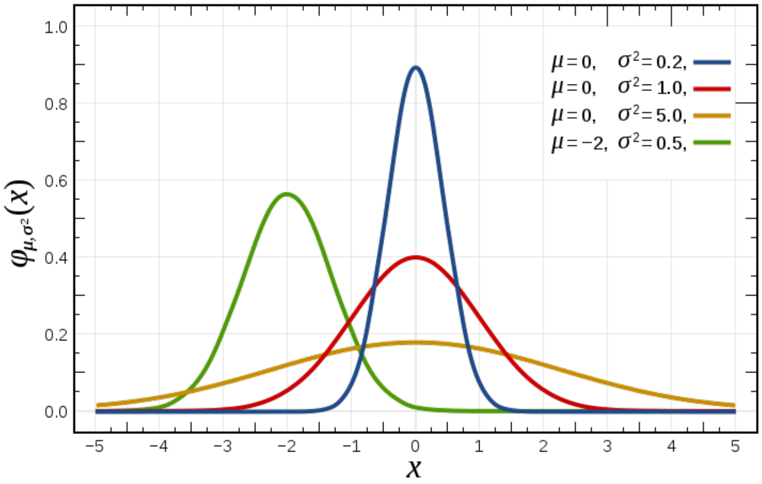
1.3 Common continuous probability distributions

Normal distribution

The PDF of a normal distribution is:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Below is an visualizing of the PDF for a normal distribution for different means and variances:



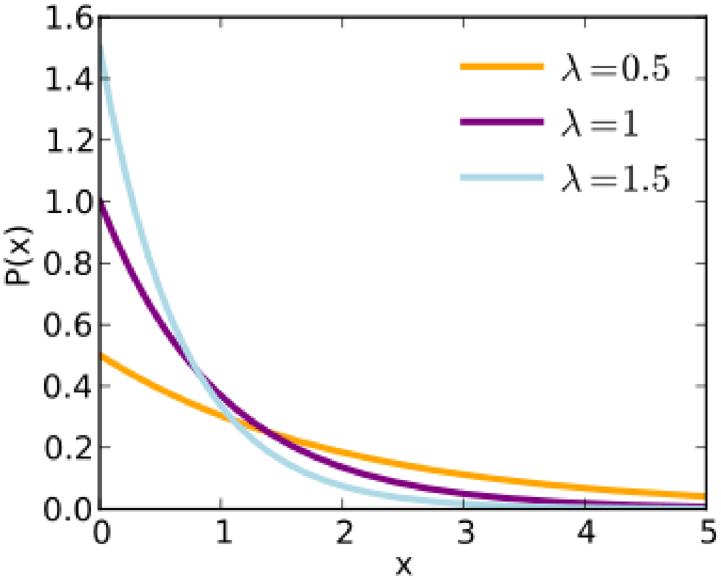
For a random variable following a normal distribution, we have that $E[X] = \mu$ and $V[X] = \sigma^2$.

Exponential distribution

The PDF of an exponential distribution is:

$$f(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Below is an visualizing of the PDF for an exponential distribution for different λ:



For a random variable following an exponential distribution, we have that $E[X] = \frac{1}{\lambda}$ and $V[X] = \frac{1}{\lambda^2}$.

Uniform distribution

The PDF of an uniform distribution is:

$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{else} \end{cases}$$

For a random variable following an uniform distribution, we have that $E[X] = \frac{1}{2}(a+b)$ and $V[X] = \frac{1}{12}(b-a)^2$.

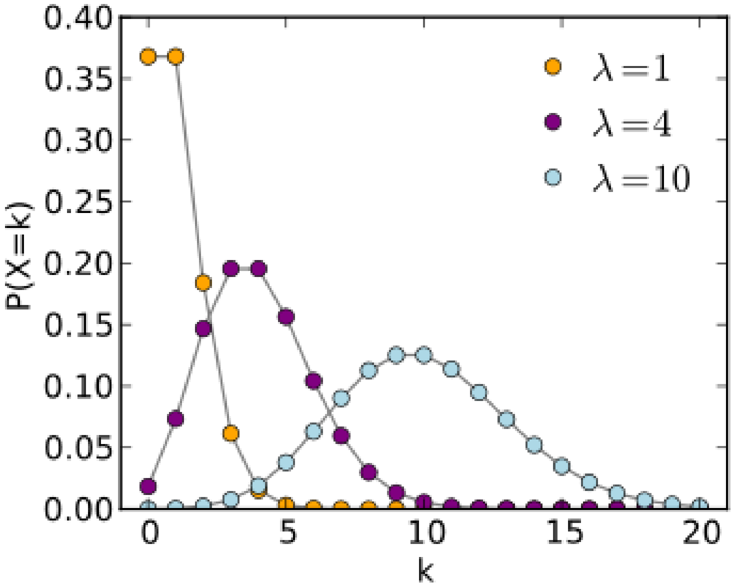
1.4 Common discrete probability distributions

Poisson distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time if these events occur with a known constant rate and independently of the time since the last event. The PMF is defined by:

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Below is an visualizing of the PMF for a Possion distribution for different λ:



For a random variable following a Possion distribution, we have that $E[X] = V[X] = \lambda$.

Bernoulli distribution

The PMF of an Bernoulli distribution is:

$$f(x|p) = p^x(1-p)^{(1-x)}$$

$x = \{0, 1\}$ and $p \in (0, 1)$.

For a random variable following a Bernoulli distribution, we have that $E[X] = p$ and $V[X] = p(1-p)$.

1.5 The law of large numbers

The law of large numbers states that the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

The law of large numbers is important because it guarantees stable long-term results for the averages of some random events. For example, while a casino may lose money in a single spin of the roulette wheel, its earnings will tend towards a predictable percentage over a large number of spins.

For example, let $\{X_1, ..., X_n\}$ be a random sample of size n, that is, a sequence of (i.i.d) random variables drawn from a distribution of expected value given by μ and finite variance. Suppose we are interested in the sample average:

$$S_n = \frac{X_1 + \cdots + X_n}{n}$$

By the law of large numbers, the sample averages converges to μ as $n \to \infty$.

1.6 Central limit theorem

The central limit theorem establishes that when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed.

For example, suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic mean of the observed values is computed. If this procedure is performed many times, the central limit theorem says that the distribution of the average will be closely approximated by a normal distribution.

Assume the average example in the law of large numbers. Then as $n \to \infty$, the random variables S_n convergence to a normal distribution $N(\mu, \frac{\sigma^2}{\sqrt{n}})$.

1.7 Bayes statistics

Two important rules when dealing with conditional distributions are:

1. Sum rule: $p(x) = \sum_y p(x, y)$ for discrete random variables and $p(x) = \int p(x, y)dy$ for continuous ones.
2. Product rule: $p(x, y) = p(x|y)p(y)$.

One can see that the product rule is symmetric w.r.t. x and y which gives:

$$p(x|y)p(y) = p(y|x)p(x) \implies p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Which is the Bayes rule. Here:

- $p(x)$ is the evidence distributions.
- $p(y)$ is the prior distributions, i.e. one's beliefs about the distributions before data is taken into account.
- $p(x|y)$ is the likelihood distributions, i.e. how likely is the data given the parameters.
- $p(y|x)$ is the posterior distributions, i.e. one's beliefs about the distributions after data is taken into account.

We have that $p(x) = \int p(x|y)p(y)dy$. This integral typically does not have an analytic solution and is intractable, making Bayesian inference hard.

One solution to this problem is to choose conjugate priors, i.e. choose a likelihood and prior such that the posterior has an analytic solution. One example of this is normal likelihood and normal prior, yielding a normal posterior.

2 Statistics

2.1 Parameter estimation

Parameter estimation is the task of estimating the values of parameters based on measured empirical data that has a random component.

There are numerous approaches, but some common ones are:

1. Maximum likelihood (ML) estimations.
2. Maximum a posteriori (MAP) estimations.

For example, assume that $X \sim N(\mu, \sigma^2)$ and we wish to estimate $\theta = \{\mu, \sigma\}$ with ML. Now assume that we draw some samples $D = \{x_i\}$, $i = 1, ..., N$ which are i.i.d. Then the likelihood is given by:

$$p(D|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

Now, we simply have to choose θ s.t. the likelihood is maximized.

2.2 Confidence intervals

Confidence interval is a interval estimate, computed from the statistics of the observed data, that might contain the true value of an unknown population parameter. The interval has an associated confidence level that, quantifies the level of confidence that the parameter lies in the interval.

The confidence level represents the frequency of possible confidence intervals that contain the true value of the unknown population parameter. In other words, if confidence intervals are constructed using a given confidence level from an infinite number of independent sample statistics, the proportion of those intervals that contain the true value of the parameter will be equal to the confidence level.

Suppose $\{X_1, ..., X_n\}$ is an i.i.d. samples from a normal distribution population with unknown μ and σ^2. Let:

$$\bar{X} = \frac{X_1, ..., X_n}{n}$$

$$S^2 \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Where $\bar{X}$ is the sample mean and S^2 is the sample variance. Then:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a Student's t-distribution with $n-1$ degrees of freedom. Note that the distribution of T does not depend on the values of the unobservable parameters μ and σ. Suppose we wanted to calculate a 95% confidence interval μ. Then, denoting c as the 97.5th percentile of this distribution,

$$p(-c \leq T \leq c) = 0.95$$

Consequently:

$$p(\bar{X} - \frac{cS}{\sqrt{n}} < \mu < \bar{X} + \frac{cS}{\sqrt{n}}) = 0.95$$

2.3 Hypothesis testing

A hypothesis is a premise or a claim that we want to test (statistically). We need to formulate two hypothesis:

H_0: Commonly, the currently accepted value for a parameter or that the observations are the result of pure chance.

H_a: Commonly, involves the claim to be tested or that the observations show a real effect.

For example, with linear regression, we might have: $H_0 : \beta_1 = 0$; $H_a : \beta_1 \neq 0$. Note that H_0 and H_a are "opposite". We can have two outcomes from this hypothesis testing:

- Reject H_0.
- Fail to reject H_0.

Next, we need to identify a test statistic in order to assess the truth of the null hypothesis. For example, with linear regression we have that $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. The test statistic (assuming σ^2 is unknown) is:

$$T = \frac{\hat{\beta} - 0}{\sqrt{MS_{Res}(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}} \sim t_{n-k}$$

Next we compute the p-value. The p-value is the probability of obtaining a sample "more extreme" than the ones observed in your data, assuming that H_0 is true.

Next, we compare the p-value to an acceptable significance value α. If $p \leq \alpha$, that the observed effect is statistically significant, the null hypothesis is rejected.

When doing hypothesis testing, we can do two types of errors:

1. Type-1 error: rejecting H_0 when H_0 is true.
2. Type-2 error: not rejecting H_0 when H_a is true.