Statistics (work in progress)

Joe Marsh Rossney

28th January 2019

Contents

Terminology, notation and abbreviations

- Capital letters, e.g. X, label random variables.
- Lower-case letters, e.g. x, label a possible value which X may take.
- When summing over many possible values, x_i is used, and for integration x' is used as a dummy variable when required.
- $P(X = x)$ is the probability of the random variable X taking the value x, often shortened to $P(x)$.
- $P(x|\alpha)$ has the variable to the left and a parameter to the right of the vertical line, and should be read as "the probability of $X = x$, given the parameter α."
- Probability mass functions (PMFs) and probability density functions (PDFs) are denoted by $f(x|\alpha)$.
- Cumulative distribution functions (CDFs) are denoted by $F(x|\alpha)$.
- The expectation value of a function $g(X)$ of a random variable X is written as $E[g(X)]$.
- Generally, the moments of X are denoted by $E[X^k]$ or μ_k. As an exception, the mean may be simply written as μ, or μ_X if there are multiple random variables.
- The central moments of X are denoted by $E[(X - \mu)^k]$ or ν_k. As an exception, the variance may be written as σ^2 or σ_X^2.
- The standard deviation is labelled by σ or σ_X.

Terminology: distribution, function, mapping.

A distribution is a more general concept than a function. As far as I'm aware, it is conventional to reserve the term 'function' for mappings $\mathbb{R}, \mathbb{C} \to \mathbb{R}, \mathbb{C}$, whereas a distribution can behave differently. For example, one cannot straightforwardly evaluate $\delta(x)$ (a distribution) at the point x, though $\int_\infty \delta(x)dx$ is well defined.

Mapping...

1 Probability distribution functions

1.1 Discrete random variables

Say we have a random variable X which can take various real values x. Formally, X is a mapping $X : \mathcal{S} \to \mathcal{A}$ from some abstract space $\mathcal{S}$ of all possible outcomes (the sample space) to a subset of the real numbers, $\mathcal{A} \subseteq \mathbb{R}$.

If X is a discrete random variable, we can seek a function $f_X : \mathcal{A} \to [0, 1]$ which directly assigns probabilities to individual elements $x \in \mathcal{A}$. $f_X(x) = P(X = x)$ is the *probability mass function* (PMF) of X. The total probability must always equal one, i.e

$$\sum_{x_i \in \mathcal{A}} f_X(x_i) = 1 \tag{1}$$

It turns out that a different function, $F_X(x) = P(X \leq x)$, is just as adequate as $f_X(x)$ for describing a distribution. $F_X(x)$ is the *cumulative distribution function* (CDF) of X, and is related to the PMF by

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i) \tag{2}$$

1.2 Continuous random variables

If X is a continuous random variable we need to be more careful. Since $\mathcal{A}$ is an infinite set, we cannot directly assign probabilities to individual elements while requiring that the total probability is finite. Instead, we can define a *probability density function* (PDF) over an infinitesimal interval dx as $f_X(x)dx = P(x < X \leq x + dx)$. Defining l and h as the lower and upper limits of $\mathcal{A}$, Eq (1) becomes

$$\int_l^h f_X(x)dx = 1 \tag{3}$$

The CDF for a continuous random variable is defined in an analogous way to Eq (2):

$$F_X(x) = \int_l^x f_X(x')dx' \tag{4}$$

The fundamental theorem of calculus then gives

$$f_X(x) = \frac{d}{dx} F_X(x) \tag{5}$$

and

$$\int_a^b f_X(x)dx = F_X(b) - F_X(a) = P(a < X \leq b) \tag{6}$$

2 Properties of distributions

Probability distribution functions are the most complete description possible for random variables. However, it is often quite inconvenient to work with distribution functions in their entirety; without god-like powers, For example, if we wanted to compare different distribution functions, simply knowing the value of $f_X(x)$ for all x may not give us much insight. We'd have far more success if we were able to somehow characterise the distributions via a small number of common features, and compare those.

Luckily, it is often possible to adequately (and sometimes even completely) describe a distribution function in this way. Sometimes the important features can be easily identified and described mathematically, such as the average value of the random variable. In more complicated cases, important features which distinguish between different distributions are obscured by a combination of noise and unimportant features with little discriminatory power. This is a realm where deep learning is already making important contributions.

2.1 Expectation values

The most common way to define characteristic features of a probability distribution is with expectation values. The expectation value of a function $g(X)$ of a random variable X is given by [1]

$$E[g(X)] = \sum_{x_i \in \mathcal{A}} g(x_i) f_X(x_i) \tag{7}$$

for discrete random variables and

$$E[g(X)] = \int_l^h g(x) f_X(x) dx = \int_l^h g(x)\, dF_X(x) \tag{8}$$

for continuous random variables, where the last equality used Eq (5). The square brackets are conventional, and appropriate since $E[g(X)]$ is a *functional* (i.e. it takes a function as its argument and returns a number).

Expectation values also have the following properties:

1. $E[ag(X)] = aE[g(X)]$ for a constant a.
2. If $g(X) = s(X) + t(X)$ then $E[g(X)] = E[s(X)] + E[t(X)]$.

[1] Eqs (7) and (8) are not definitions, nor trivial results from other definitions – they must be rigorously derived. Apparently, a lack of appreciation of this fact (something I myself have never been guilty of) led to them being dubbed the *"law of the of the unconscious statistician"*.

2.2 Moments

For a random variable X, the k^{th} moment of its distribution is given by $\mu_k \equiv E[X^k]$, which yields the following formulae for discrete and continuous X, respectively:

$$\mu_k = \sum_{x_i \in \mathcal{A}} x_i^k f_X(x_i) \tag{9}$$

$$\mu_k = \int_l^h x^k f_X(x) dx = \int_l^h x^k \, dF_X(x) \tag{10}$$

For $k = 0$, Eqs (9) and (10) reduce to Eqs (1) and (3) – i.e. the 'zeroth' moment is the total probability and the fact that $\mu_0 = 1$ is a statement of global conservation of probability. This isn't particularly interesting, but it's good to know that it checks out.

The first moment $\mu_1 \equiv E[X]$ is also significant; it's the *mean value* of X, usually simply written as μ.

2.3 Central moments

Higher order moments, as defined above, turn out to not be particularly effective at describing the shape of typical distributions, due to the fact they are defined around $X = 0$. We're free to make any translation $x \mapsto x + y$ and take moments around the point y. However, it makes the most sense to define moments around μ, since most distributions we're interested in will be at least vaguely symmetric around their mean value.

This leads to the following formulae for the k^{th} central moment $\nu_k \equiv E[(X - \mu)^k]$:

$$\nu_k = \sum_{x_i \in \mathcal{A}} (x_i - \mu)^k f_X(x_i) \tag{11}$$

$$\nu_k = \int_l^h (x - \mu)^k f_X(x) dx = \int_l^h (x - \mu)^k \, dF_X(x) \tag{12}$$

Firstly, it is clear that $\nu_0 = \mu_0$ and that $\nu_1 = 0$. What about the second central moment, $\nu_2 \equiv E[(X - \mu)^2]$? This is just another name for the *variance* of the distribution, σ^2.

There is a way of expressing ν_2 in terms of μ and μ_2, which turns out to be quite useful.

We will derive it here:

$$\begin{aligned}
\nu_2 &= E[(X-\mu)^2] \\
&= E[X^2 - 2\mu X + \mu^2] \\
&= E[X^2] - 2\mu E[X] + \mu^2 \\
&= \mu_2 - 2\mu\mu + \mu^2 \\
&= \mu_2 - \mu^2
\end{aligned} \tag{13}$$

where in line 3 we used properties of the expectation value (subsection 2.1). In most cases, people care more about expectation values and variances than moments and central moments, so it's more common to see this written as something like $\sigma^2 = E[X^2] - E[X]^2$.

2.4 Standardised moments

2.5 Looking ahead: generating functions

2.6 Properties defined without expectation values

Mode, median etc.

3 Generating functions

3.1 Probability generating functions

3.2 Moment generating functions

3.3 Cumulant generating functions

3.4 Characteristic functions

4 Joint distributions

4.1 Covariance and correlation