Probability Part II

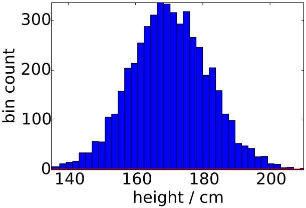
Lecture 8 of *Core Mathematics* in the MPH and MSc in Epidemiology Courses, Imperial College London

Chris Wymant

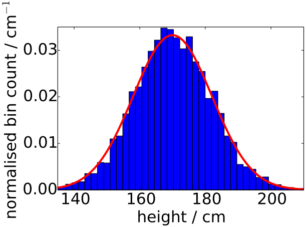
December 2015

Continuous Probability I

Continuous random variables take continuous values, e.g. they might be equal to *any* number in particular range, as opposed to any whole number. The probability of finding them equal to any specific value is vanishingly small; instead we ask what is the probability of finding them in a particular range.

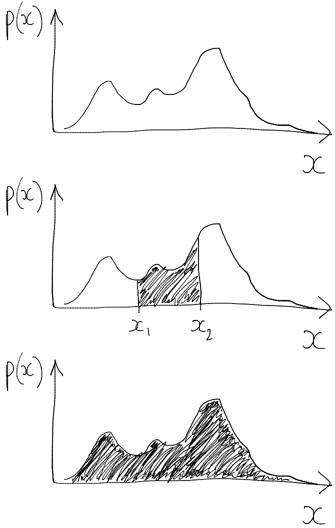


Imagine recording human heights in a large random sample, binning them, and plotting the bin counts in a histogram. What is the probability that a randomly selected individual has a height in some specified range? It's the fraction of counts falling in that range.



Unit-normalise the histogram by dividing bin counts by (the total count × bin width). The probability of having a height in a particular range is given by the area under this histogram in that range. Approximating the histogram shape with a continuous function gives you something you can integrate.

Continuous Probability II



Define a function $p(x)$ such that probability is given by integrating the function over the desired range. This is a *probability density function* (pdf).

$$P(x_1 \leq x < x_2) = \int_{x_1}^{x_2} p(x) dx$$

The probability that the variable has some value, any value, is 1. Therefore

$$\int_{-\infty}^{\infty} p(x) dx = 1$$

NB $p(x)$ itself can be greater than 1.

A Simple pdf Example

A continuous random variable x has the pdf

$$p(x) = \begin{cases} \frac{1}{B-A} & \text{if } A \leq x < B, \\ 0 & \text{otherwise} \end{cases}$$

where A and B are constants (do not depend on x).

1. Sketch $p(x)$. (Hint: consider the three regions $x < A$, $A \leq x < B$, and $x \geq B$ separately, but draw them all on the same graph.)
2. What is the probability that $x_1 \leq x < x_2$, if
   - 2.1 $x_1 < A$ and $x_2 < A$?
   - 2.2 $x_1 < A$ and $A < x_2 < B$?
   - 2.3 $A < x_1 < x_2 < B$?

   (Hint: the answer is $\int_{x_1}^{x_2} p(x) dx$ in each case, but $p(x)$ is equal to a different constant in the three different regions for x.)
3. How would you describe in words what this pdf represents?

The Mean and Variance I

To calculate the *mean* of a list of numbers, sum them then divide by how many there are.
e.g. the mean of 4, 4 and 7 is $(4+4+7)/3 = 5$.

Equivalently: assign each distinct value a probability equal to the number of times it occurs in the list divided by how many values there are in the list, then add all distinct values each multiplied by (or 'weighted by') its probability. In the above example, the probability of 4 is $\frac{2}{3}$ and the probability of 7 is $\frac{1}{3}$, so we add $4 \times \frac{2}{3}$ to $7 \times \frac{1}{3}$ to get 5.

Generalising this process, for a discrete random variable X the mean (or 'expectation') is given by multiplying each possible value X can take by its probability, then adding them all together:

$$E[X] = \langle X \rangle = \sum_{x} x \, P(X = x)$$

e.g. one can show that the mean of the binomial distribution is Np.

The Mean and Variance II

For a continuous random variable, calculating the mean is the same except we integrate instead of summing:

$$E[X] = \int_x x\, p(x) dx$$

The *variance* is the expected value of (the difference from the mean)^2. This is a measure of how disperse the distribution is – how usual it is to see values far away from the mean.
Letting $E[X] = \mu$ for clarity:

$$\begin{aligned} E[(X-\mu)^2] = & \sum_x (x-\mu)^2\, P(X=x), \quad \text{if } X \text{ is discrete} \\ & \int_x (x-\mu)^2\, p(x) dx, \quad \text{if } X \text{ is continuous} \end{aligned}$$

The Geometric Mean I

The mean just referred to is also called the *arithmetic mean*, to distinguish it from another quantity called the *geometric mean*. To calculate this, multiply all the values together then raise to the power 1/(how many there are). i.e.

$$\text{arithmetic mean}(x_1, x_2, \ldots x_N) = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots x_N}{N}$$

$$\text{geometric mean}(x_1, x_2, \ldots x_N) = \left(\prod_{i=1}^{N} x_i\right)^{1/N} = (x_1 \times x_2 \times \ldots x_N)^{1/N}$$

Equivalent to this definition of the geometric mean is to take logarithms of each of the values, calculate the arithmetic mean of those, then exponentiate:

$$\text{geometric mean}(x_1, x_2, \ldots x_N) = \exp\left(\sum_{i=1}^{N} \ln(x_i)/N\right)$$

The Geometric Mean II

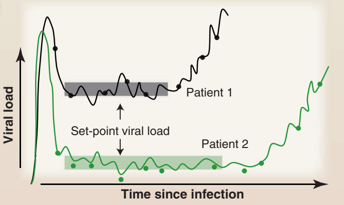
The arithmetic mean of two values is such that you can *add* a number to the smaller value and get the mean, then *add the same number again* to get the larger value.

The geometric mean of two values is such that you can *multiply* the smaller value by a number and get the mean, then *multiply by the same number again* to get the larger value.

e.g. the arithmetic mean of 3 and 27 is 15; $3 \xrightarrow{+12} 15 \xrightarrow{+12} 27$

the geometric mean is $(3 \times 27)^{1/2} = 9$; $3 \xrightarrow{\times 3} 9 \xrightarrow{\times 3} 27$

These properties means that the arithmetic mean is halfway between two numbers on a linear scale, and the geometric mean is halfway between them on a logarithmic scale.



A widely used characterisation of the severity of an HIV infection is the *set-point viral load* – the geometric mean of measurements of viral load taken after acute infection and before treatment or AIDS.

The Median

Wikipedia: "the number separating the higher half of a data sample, a population, or a probability distribution, from the lower half. The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one (e.g., the median of $\{3, 3, 5, 9, 11\}$ is 5). If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values (the median of $\{3, 5, 7, 9\}$ is $(5 + 7)/2 = 6)$".

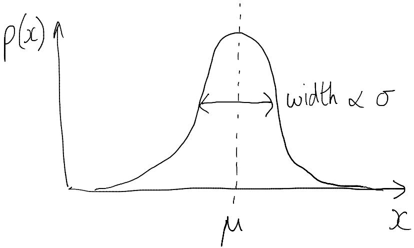
The median of a continuous random variable X is the value m such that there are equal chances of being larger than m and being smaller than m:

$$\int_{-\infty}^{m} p(x)dx = \int_{m}^{\infty} p(x)dx = 0.5$$

Don't need to know: the median of a discrete random variable X is tricky: it is any value m such that $P(X \leq m) \geq 0.5$ and $P(X \geq m) \geq 0.5$. There may be more than one such value.

The Normal Distribution I

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



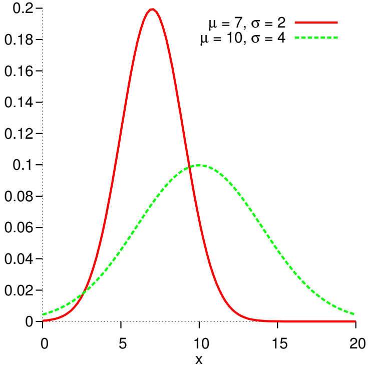
μ and σ are parameters of this pdf (its integral over all x equals 1 regardless of the value μ and σ).

The mean is μ and the variance is σ^2 (the standard deviation is σ).

The function is symmetric about the point $x = \mu$, i.e. $p(\mu - a) = p(\mu + a)$ for any a.

Many things which we think cluster symmetrically about some central value are modelled as following a normal distribution, for example random measurement error.

The Normal Distribution II



The integral of the function from $\mu - \sigma$ to $\mu + \sigma$ is 0.68 (2 d.p.): you have a roughly $\frac{2}{3}$ chance of being less than one standard deviation from the mean. The integral from $\mu - 2\sigma$ to $\mu + 2\sigma$ is 0.95.

It is important to note that

$$\int_{-\infty}^{0} p(x) dx > 0$$

What does this mean?

Hazards I

The *hazard* λ for something happening is the probability-per-unit-time of it happening. This means the probability of it happening in a small time window δt is approximately $\lambda\delta t$, provided $\lambda\delta t$ is much smaller than 1. A hazard constant in time implies that if the thing hasn't happened yet, the probability of it happening in the next small window δt is always $\lambda\delta t$ regardless of how much time has passed. Let $P(t)$ be the probability that whatever it is has *not* happened yet, at time t. The probability that it happens between t and $t + \delta t$ is

$$P(t) - P(t + \delta t) \approx P(t) \times \lambda\delta t$$

(The factor of $P(t)$ on the right-hand side accounts for the our requirement that it does not happen before t.) Rearranging,

$$\frac{P(t + \delta t) - P(t)}{\delta t} \approx -P(t) \times \lambda$$

Hazards II

Letting $\delta t \to 0$, the approximation becomes exact and the left-hand side becomes a derivative:

$$\frac{dP(t)}{dt} = -\lambda P(t) \quad \text{which is solved by} \quad P(t) = Ae^{-\lambda t} \quad \text{for any } A.$$

Defining $t = 0$ to be the last time we knew for sure the thing hadn't happened yet, $P(0) = 1$ (an *initial condition*) and so $A = 1$.

In summary, when there is a constant hazard for something happening, the time taken for it to happen is a continuous random variable that's exponentially distributed.

e.g. people often assume the hazard for leaving a compartment in a compartmental model (by changing state through disease progression, recovery, death etc.) is constant…

Time Spent in Compartments

In a particular state in a compartmental model, with constant hazard for leaving λ,

- the probability for still being in the state a time t later – the 'survival time' – is $e^{-\lambda t}$,
- the pdf for *the total amount of time spent in the state* is $\lambda e^{-\lambda t}$,
- the mean time spent in the state is $\frac{1}{\lambda}$.

Exercises:

1. Verify that the pdf is correctly *normalised*, i.e. integrating over all possible values give 1.
2. Sketch the pdf for small, medium and large values of λ.
3. Calculate the probability that the time spent in the state is greater than its mean value.
4. What is the probability that the time spent in the state is greater than its median value?

Risk

The *risk* of something happening is the probability that it *does* happen, in some specified amount of time.

A constant hazard λ for something happening means the probability it *does not* happen in time t is $e^{-\lambda t}$.
The risk for a constant hazard[1] is therefore $1 - e^{-\lambda t}$.
This starts at 0, and for large times $t \gg \frac{1}{\lambda}$ approaches 1.

A result from maths: for $x \ll 1$, $e^x \approx 1 + x$.
This means that for $t\lambda \ll 1$, i.e. $t \ll \frac{1}{\lambda}$, the risk is $\approx \lambda t$: risk increases linearly in the short term.

Exercise: the annual risk for malaria in the Oshikuku district of Namibia is 42%[2]. Assuming constant hazard, what is the probability of an individual remaining malaria-free for three years?

[1] Keen beans see me after for the equation for time-varying hazards.

[2] http://www.malariajournal.com/content/13/1/52/table/T1

Conditional Probability

The vertical bar symbol | means 'given that'.
$P(A|B)$ is the probability that A is true given that B is true.
NB it's still fine to talk about this even if B isn't definitely true – this can represent the hypothetical '*if* B were to be true, *then* what would the probability of A be'.

$$P(A|B) = P(A \text{ and } B) \,/\, P(B)$$

$$\begin{aligned} P(A|B) &> P(A) \text{ if } B \text{ being true makes } A \text{ more likely to be true;} \\ &< P(A) \text{ if } B \text{ being true makes } A \text{ less likely to be true;} \\ &= P(A) \text{ if } A \text{ and } B \text{ are independent} \end{aligned}$$

Note that $P(A \text{ and } B) = P(A|B)P(B)$, but also
$P(A \text{ and } B) = P(B|A)P(A)$, therefore . . .

Bayes' Theorem I

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

$$P(\text{explanation}|\text{observation}) = \frac{P(\text{observation}|\text{explanation})P(\text{explanation})}{P(\text{observation})}$$

Something might explain what you have observed very well, but if it's sufficiently implausible *a priori*, you'll discount it as a possibility anyway. "Extraordinary claims require extraordinary evidence."

Important: by convention in this field, *statistically significant* means $P(\text{observation}|\text{null hypothesis})$, the "$p$-value", is less than 0.05. On its own, that tells you nothing about $P(\text{null hypothesis}|\text{observation})$!

NB the denominator $P(\text{observation})$, "the probability of making the observation", often doesn't really make sense. But if we want to compare two different explanations, that same denominator occurs for both: taking a ratio, it disappears.

Bayes' Theorem II

More advanced. Only for those intending to model.

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

- $P(\text{model}|\text{data}) =$ 'the posterior probability'
- $P(\text{data}|\text{model}) =$ 'the likelihood'
- $P(\text{model}) =$ 'the prior'

If the model depends on some parameters (e.g. transmission rates, recovery rates), then for fixed data, the likelihood and prior are both functions of the parameters. If you have uninformative/flat priors (i.e. you have no reason to believe any values of the parameters are more likely than any others), then the parameters that give the maximum likelihood will also give the maximum posterior probability.

Some Further Reading

An investigation of the false discovery rate and the misinterpretation of p-values, D. Colquhoun, Royal Society Open Science (2014): http://dx.doi.org/10.1098/rsos.140216
A non-mathematical explanation of what it says on the tin. "If you use $p = 0.05$ to suggest that you have made a discovery, you will be wrong at least 30% of the time." Caveat: claims to be independent of Bayesian ideas, but really it isn't.

Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Y. Benjamini and Y. Hochberg, Journal of the Royal Statistical Society Series B (1995): http://www.jstor.org/stable/2346101
A method for controlling for multiple testing ($\sim 30,000$ citations).

http://tinyurl.com/o4osslo by Glen Cowan. An outstanding lecture on probability and statistics for science, for those with strong mathematical ability.