# Introduction to probability distributions

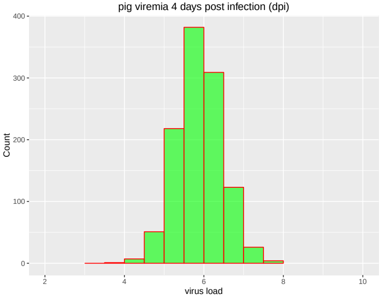
Osvaldo Anacleto
Genetics and Genomics, Roslin Institute
osvaldo.anacleto@roslin.ed.ac.uk

February 2018

# Overview

- Why do we need probability?
- Random variables
- Discrete probability distributions: binomial and Poisson
- Continuous probability distributions: normal
- Poisson processes

# **Example:** viraemia measurements from 1,120 pigs exposed to the PRRS virus in an infection experiment



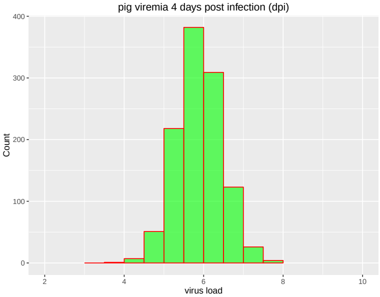

- these measurements **vary** from pig to pig in way that is unpredictable (random)
- to learn about the pig response to the virus, we need to represent and describe the variability in the data (i.e what sorts of measurements are “likely” and which are “unlikely”)
- this variability is represented using **probability models**

# Why do we need probability?

**randomness and probability are central to statistics**

- data are outcomes from experiments and/or complex systems
- experiments and complex systems are random (not reproducible, “unpredictable”)
- we need to represent the uncertainty underlying the nature of the data
- probability is a measure of uncertainty

# Random variables




- In the example, viraemia is a **random variable**
- this random variable varies from pig to pig
- random variables may take any values from a set of possible values, but some values may be more likely than others to occur (i.e. the variable is random, but there is some **structure** in it)
- probability models (distributions) are mathematical representations of this **structure**

# Probability distributions

- A **probability distribution** of a random variable X is the **set of probabilities** assigned to each possible value of X.
- This set of probabilities can be represented by a table, a graph of a **function**
  - Example: the pig histogram represents the distribution of viraemia (a random variable) among pigs based on the data

Probability distributions can vary according to the set of possible values of a random variable

- **discrete random variables** may take only a countable number of distinct values examples: faces of a dice, number of ticks in a sheep, number of infections over time
- **continuous random variables** may take values within a certain range (a continuum) examples: height, viremia, time to infection.

# Discrete distributions

- The distribution of a discrete random variable is represented by a **probability (mass) function (pmf)**
- This function assigns a probability to each possible value of a random variable
- All probabilities from the distribution must be non-negative and their sum must be 1

**Example: the binomial distribution**

Suppose a sample of 50 fish is collected independently from a lake and that some of these fish are infected with a disease of interest. What's the probability of

- having 2 infected fish in this sample?
- Or 30?
- less than 10 infected fish?

# Discrete distributions: the Binomial distribution

- We can define a **random variable** X as the number of infected fish from a sample of size N=50
- The binomial distribution can be used to represent the variability underlying $X$

The probability mass function of the binomial distribution is

$$P(X = x) = \binom{N}{x} p^k (1 - p)^{50-x}$$

where N=50 and **p** is the probability of having an infected fish at **each** collection of a fish from the lake (Bernoulli trial)

- $N$ and $p$ are the **parameters** of the binomial distribution
- The parameters of a distribution control its shape and can have useful interpretations
- A key task in statistics is how to estimate distribution parameters from data

**Notation:** if X follows a Binomial distribution, $X \sim Bin(N, p)$

# Discrete distributions: the Poisson distribution

- The Poisson distribution is suitable for **counts** of independent events that occur in some fixed region of **time** or **space**

Examples:

- number of mutations in the genome (e.g per region)
- number of traffic accidents along a stretch of a road
- number of infections per week

The probability mass function of the Poisson distribution is

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

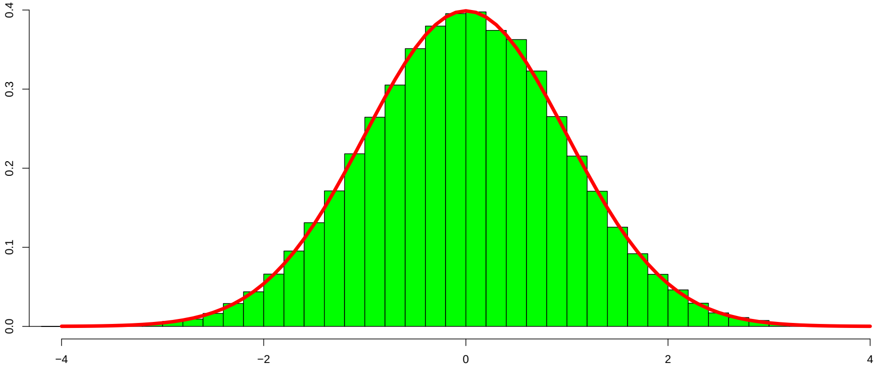
$\lambda$ is a parameter which represents a **rate** - the expected counts per unit of observed time (or region);

**Notation:** if X follows a Poisson distribution: $X \sim Poi(\lambda)$

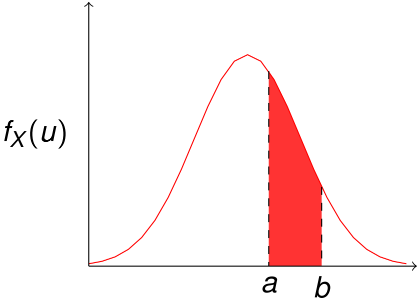
**note:** in probability and statistics, the upper-case $X$ is used for the random variable under consideration and lower-case $x$ is used to represent the possible values the random variable $X$ might take

# Continuous distributions

- A smooth curve can characterise the probabilities associated with continuous random variables
- This curve is described by the **probability density function (p.d.f)**



# Probability density functions




a probability density function:

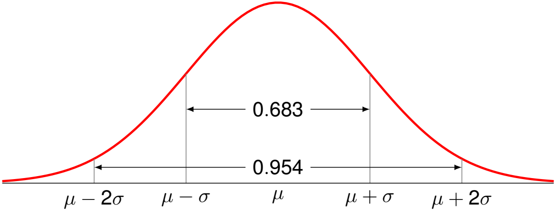
- must take only non-negative values
- their area under the curve must be 1

- then, the area under the curve between two values x and y gives the probability that the random variable will take a value somewhere between x and y (done by integration)

For example, for a random variable with a distribution given by a p.d.f $f_X(x)$:

$$P[a \leq X \leq b] = \int_a^b f_X(u) du$$

# Continuous distributions: the normal distribution




- The normal (or Gaussian) distribution is one of the most important distribution in statistics
- The normal distribution depends on two parameters
  - $\mu$: represents a central point where a distribution peaks (expectation of the random variable)
  - $\sigma^2$: which represents the dispersion or the degree of the variability in the outcome (variance of the random variable)

if $X \sim N(\mu, \sigma^2)$, its p.d.f is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{(2\sigma^2)}\right\}$$

# Why is the normal distribution important?

- it has separate parameters for the mean and variance of a random variable (quantities that are often of primary interest)
- the distribution is symmetric around $\mu$, and the mode, median and mean are all equal to $\mu$
- it is appropriate for data that result from the additive effects of a large number of factors (due to the central limit theorm - see inference lecture)

# Some remarks about probability distributions

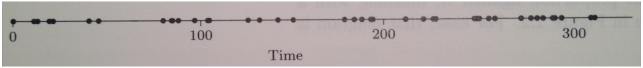
- There are also distributions for vectors of random variables (e.g. the multivariate normal distribution)
- A probabilility distribution is not defined by its density or mass function only
- When using probability distributions for statistical inference, a crucial step is checking if the chosen distribution fits the data well (e.g. by using q-q plots, goodness of fit tests, etc.)

# Counting processes

Suppose the data below show the **times between infections** during a disease outbreak:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 12 | 2 | 6 | 2 | 19 | 5 | 34 | 4 |
| 1 | 4 | 8 | 7 | 1 | 21 | 6 | 11 |
| 8 | 28 | 6 | 4 | 5 | 1 | 18 | 9 |
| 5 | 1 | 21 | 1 | 1 | 5 | 3 | 14 |
| 5 | 3 | 4 | 5 | 1 | 3 | 16 | 2 |

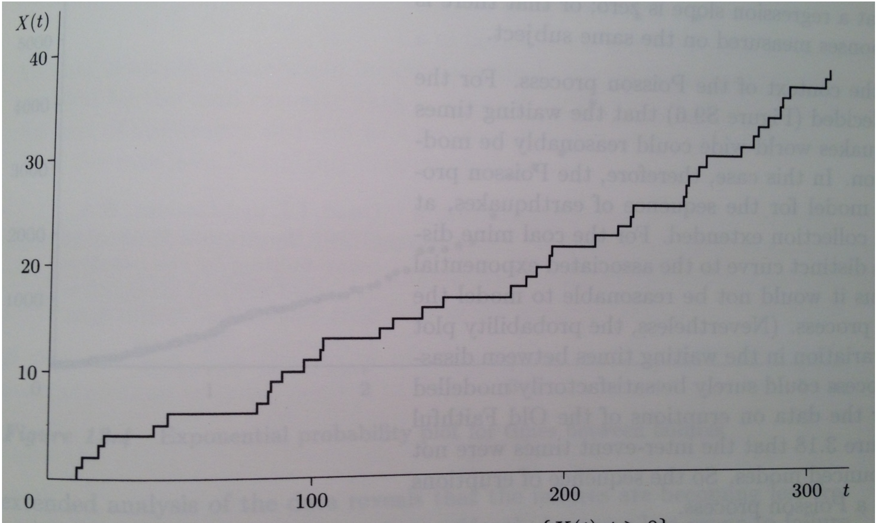
One way of viewing these data is plotting the time-to-infection on a time axis, such that the first infection is represented at time 0:




- The **order** of infections is crucial when modelling infectious disease data. So, we can define a random variable $X(t)$ which represent the number of infectious occurred by time $t$.
- The collection of all possible random variables $\{X(t), t \geq 0\}$ is a type of random (stochastic) process called **counting process**

# Counting processes

The data can then be viewed as a **realization** of the counting process $\{X(t), t \geq 0\}$:




A realization of the random process $\{X(t); t \geq 0\}$

# Poisson Processes

The counting process $\{X(t), t \geq 0\}$ is a (homogeneous) Poisson Process if:

- the number of events $X(t)$ occuring in time intervals of duration $t$ follows a **Poisson distribution** with mean $\lambda t$
- the times between consecutive events are independent observations of a continuous random variable following an **exponential distribution** with parameter $\theta = 1/\lambda$

**Poisson processes are extensively used when modelling and simulating stochastic epidemic models**
(see lectures on Tuesday and Thursday)

# References


- Diggle, P. J., & Chetwynd, A. G. (2011). Statistics and scientific method: an introduction for students and researchers. Oxford University Press.
- Daly, F., Hand, D. J., Jones, M. C., Lunn, A. D., & McConway, K. J. (1995). Elements of statistics. Addison-Wesley Publishing Company.
- Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2). Pacific Grove, CA: Duxbury.
- Grimmett, G., & Stirzaker, D. (2001). Probability and random processes. Oxford university press.