1 Distributions

Distributions

You have been introduced to frequency distributions describing a sample. The form of these distributions is based upon the collected data themselves.
In this section we are going to discuss frequency distributions that come from theoretical predictions. Generally, a theoretical frequency distribution is known as a *probability distribution* .

Discrete vs Continuous

Probability distributions can be used to describe counts or frequencies and these are known as *discrete* probability distributions.
They can also describe continuous processes and unsurprisingly are known as *continuous* probability distributions
Before we introduce the distributions there is some introductory information.

1.1 Random Sampling

Random Sampling

Random sampling is an attempt to

- include all types of events in a sample proportional to how often they occur.
- As Sokal and Rohlf (1995) state, "every individual of the population has an equal probability of being included in the sample".

If 50% of the indivudals in a population are male, then under true random sampling, the proportion of males in the sample should be close to 50%.

How to randomly sample

- use a random number source!
  - enumerate samples and use a random number table or list
  - make a choice based upon a random number whether to sample an individual as it is encountered
- Seemingly random, but not really
  - Drunken Walk
  - Haphazard sampling

1.2 Probability basics

Probability basics

The *sample space* refers to all of the possible outcomes that can occur.

Genetic example

- Genotype 'Aa' can produce 'A' and 'a' containing gametes
  This is the sample space for outcomes of meiosis for a single locus.
- A cross between 'Aa' and 'Aa' can produce 'AA', 'aA', 'Aa', 'aa' offspring
  these outcomes are the sample space for monohybrid cross.

1.2.1 Events

Events

An *event* is a single observation or set of observations that come from a sample space.
If a single member of the sample space is chosen it is also called a *simple event*
If a set are chosen it is called an event.

1.2.2 Sets

Sets

Make two sets $A = a, b, f, g$ and $B = b, c, d, e, f$

- Union. The union of the sets $(A \cup B) = a, b, c, d, e, f, g$
- Intersection. $(A \cap B) = b, f$

1.2.3 Probabilities

Probabilities

Probabilities describe how often events may be observed
In the classical sense a probability can be defined as the limit of how often an event occurs divided by the number of events sampled as the sample size increases
In the genetic example, a 'Aa' x 'Aa' cross produces a sample space whose simple events should occur at the same frequency based on Mendels laws.
Only after picking fairly large numbers of offspring, would it be possible actually estimate these probabilities in real data

1.3 Notation

Notation and Manipulation

Meiosis in 'Aa' individual

- Sample space 'a', 'A'
- Probability of event 'a' = $P(a)$
- Probability of event 'A' = $P(A)$

Because the probabilities of all events should total 1.0, $P(a) = 1 - P(A)$. (more general $P(a) = P(A^C)$)

1.4 Probabilities of combined events

Sets of events

Referring back to genetic cross. These probabilities refer to the genotype of a single offspring

- $P[Aa \cup aA] = P[Aa] + P[aA] - P[Aa \cap aA]$

$P[Aa \cap aA]$ is the empty set. Therefore Aa and aA are *mutually exclusive* .

1.4.1 Sampling events from distributions

Sampling issues

- Sampling with or without replacement
- Independent sampling

Combining probabilities

- Probability of 'a' and 'b' $=P(a) \times P(b)$ (under independence) (intersection)
- Probability of 'a' or 'A' $=P(a) + P(A)$ (union)

2 Discrete probability distributions

Discrete distributions

Discrete probability distributions describe events that can be placed into categories.

- binary (survive/not survive)
- quantitative discrete (number of offspring)
- qualitative (genotypes)

2.1 Binomial Distribution

Binomial Dist

Pulling alleles from a gene pool with two genetic types can be modeled using a binomial distribution.
Probabilities of picking alleles $P(A) = p$, $P(a) = q$.
Distribution of outcomes when one sample is taken:

Allele	A	a
Probability	p	q

Increasing sample size

If two alleles are picked:

Genotype	AA	Aa	aa
Probabilities	p^2	$2pq$	q^2

These are examples of *binomial expansion* where $(p + q)^k$. k is the sample size.

Formal expression of Binomial

The binomial expansion can be generalized:

$$\binom{k}{Y} p^Y q^{k-Y} = \frac{k!}{Y!(k-Y)!} p^Y q^{k-Y}$$

Binomial Example

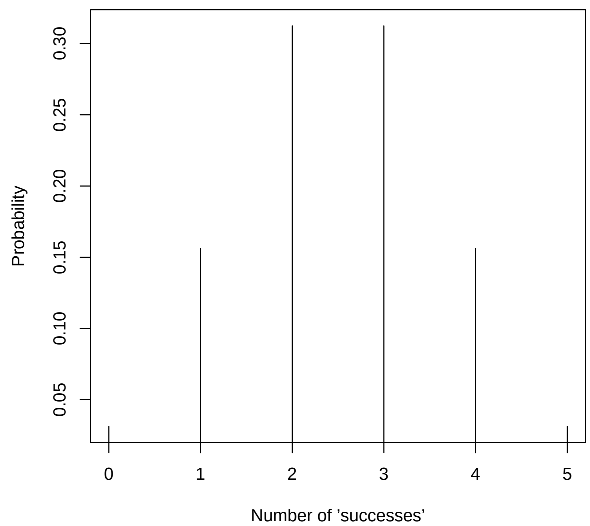
The gender of children makes a nice example.
$P[M] = P[F] = 0.5$ ($p = q = 0.5$)
So the probability of having two boys in a family of 3 children is:

$$\frac{k!}{Y!(k-Y)!} p^Y q^{k-Y} = \frac{3!}{2!1!} 0.5^2 0.5^1 = 0.375$$

In R, the `dbinom()` function calculates these probabilities

Binomial Example II

```
> plot(0:5, dbinom(c(0:5), 5, 0.5), type = "h", xlab = "Number of 'successes'",
+     ylab = "Probability")
```



2.2 Poisson Distribution

Poisson

The Poisson distribution models processes that produce relatively few rare events.
These events include

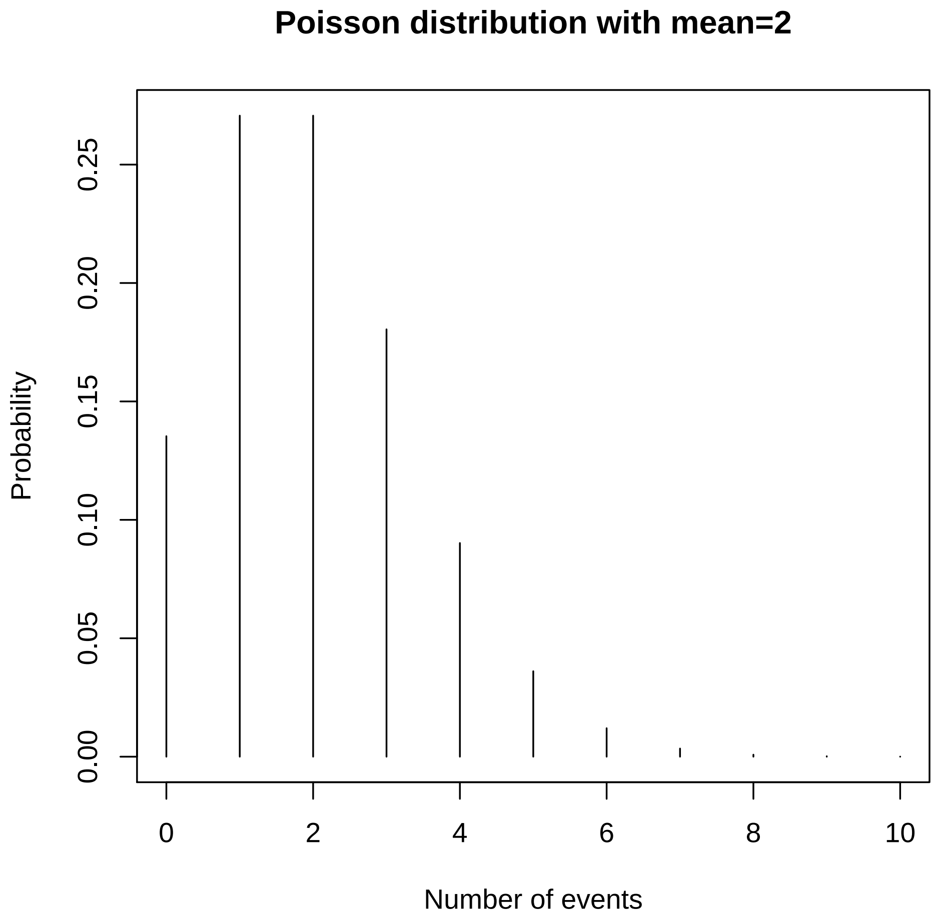
- numbers of offspring per mother
- numbers of plants observed in a plot
- incidence of disease over time

Note that these events *must be independent*

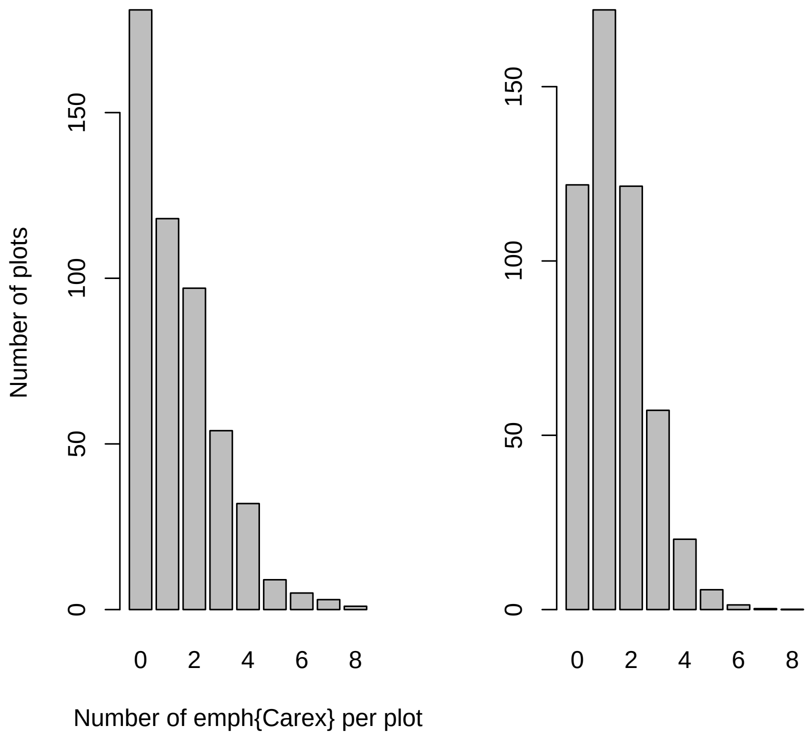
Poisson Distribution

The Poisson distribution is given by an infinite series:
$\frac{1}{e^{\mu}}, \frac{\mu}{1!e^{\mu}}, \frac{\mu^2}{2!e^{\mu}}, \ldots$
Corresponding to the probabilities of observing:
$0, 1, 2, \ldots$

rare events.



Comparing real data to a Poisson dist



3 Continuous probability distributions

Continuous Probability Distributions

As you no doubt guess, these distributions describe continuous phenomena, the familiar

- Length
- Width
- Mass

References

Sokal, R. R. and Rohlf, F. J. (1995). *Biometry. The principles and practice of statistics in biological research*, third edn, W. H. Freeman, San Francisco, California.