Random Variables, Expectations, Variance, and Covariance

Alan R. Rogers

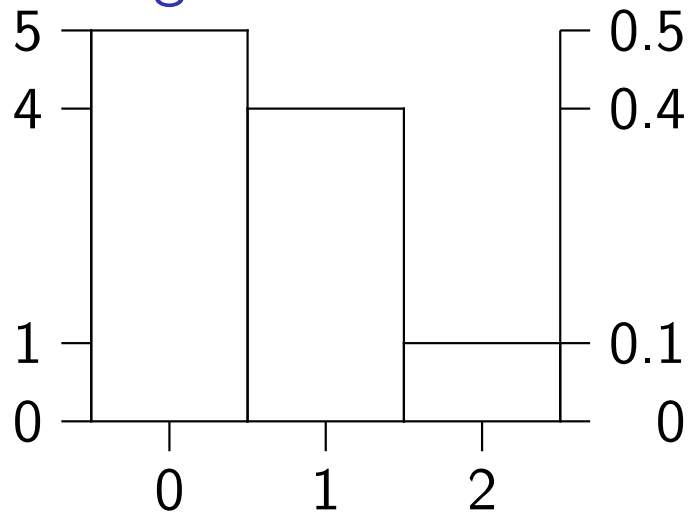
August 13, 2019

Distributions of counts and of relative frequencies

$$\text{data} = [0, 1, 0, 0, 1, 1, 1, 0, 0, 2]$$

Value	Count	Relative frequency
0	5	0.5
1	4	0.4
2	1	0.1
	10	1.0

Histogram



The mean (m)

$$m = \frac{1}{N} \sum_{i=1}^{N} x_i$$

where N is sample size and x_i is ith data value.
Example: If $x = [3, 2, 2]$, then

$$m = \frac{1}{3} \times (3 + 2 + 2) = 7/3$$

Using relative frequencies to calculate the mean

$$m = \sum_x x f_x$$

where x is a sample value and f_x is the relative frequency of that value.

Example: If $x = [3, 2, 2]$, then

$$\begin{aligned} f_3 &= 1/3 \\ f_2 &= 2/3 \\ m &= (3 \times 1/3) + (2 \times 2/3) = 7/3 \end{aligned}$$

Larger example

$$\text{data} = [0, 1, 0, 0, 1, 1, 1, 0, 0, 2]$$

Value	Relative frequency
0	0.5
1	0.4
2	0.1

$$\begin{aligned} m &= 0 \times 0.5 + 1 \times 0.4 + 2 \times 0.1 \\ &= 0.4 + 0.2 = 0.6 \end{aligned}$$

Measures of variation

- range of data
- interquartile range: range of middle half of data
- variance: average of $(x - m)^2$, where m is the mean
- square root of variance: the standard deviation

In population genetics, the variance is most useful.

Calculating the variance (v)

$$v = \sum_x (x - m)^2 f_x$$

where m is the mean, x is a sample value, and f_x is the relative frequency of that value.

What are the mean and variance of this data set: $[3, 2, 2]$?

Calculations

Frequency distribution:

Value	Relative frequency
2	2/3
3	1/3

Mean:

$$\begin{aligned} m &= 2 \times \frac{2}{3} + 3 \times \frac{1}{3} \\ &= 7/3 \approx 2.33 \end{aligned}$$

Variance:

$$\begin{aligned} V &= (2 - 2.33)^2 \times \frac{2}{3} \\ &\quad + (3 - 2.33)^2 \times \frac{1}{3} \\ &= 0.22 \end{aligned}$$

These ideas work not only for relative frequencies but also for probabilities.

- Frequency distributions become probability distributions.
- Means become expected values.
- Nothing else changes.

Probability distribution

- Assigns a probability to every event.
- When events have numeric values, the probability distribution translates one number (the event) into another (the probability).
- A set of events with associated probabilities is a *random variable* (r.v.).
- Distributions of numerical r.v.s are often described using mathematical functions.

A **random variable** is a variable whose values occur with particular probabilities.
(We would need to modify this slightly for variables that vary along a continuum, such as height or weight. But I'm going to ignore that distinction here.)

Example 1: a fair coin

Suppose that X (a random variable) is the number of heads in one toss of a fair coin. The *probability distribution* of X is

X	Probability (p_X)
0	1/2
1	1/2

Probabilities

- lie between 0 and 1,
- sum to 1.

Example 2: a loaded die

Let X be the number obtained on a roll of the die. This die is "loaded," so that 1s and 2s are twice as probable as other values.

X	(p_X)
1	0.250
2	0.250
3	0.125
4	0.125
5	0.125
6	0.125
	1.0000

The mean (or expectation) of a random variable

The mean of X is written $E(X)$ and equals

$$E(X) = \sum_i p_i x_i$$

where x_i is the ith value that X can take, and p_i is its probability. If X is the number obtained on a roll of our loaded die, then

$$\begin{aligned} E[X] &= 1 \times 0.25 + 2 \times 0.25 + 3 \times 0.125 \\ &\quad + 4 \times 0.125 + 5 \times 0.125 + 6 \times 0.125 \\ &= 3 \end{aligned}$$

The same as an average, except that p_i is a probability rather than a relative frequency.

Allele frequency as expectation

G'type	G'type freq	Cond. allele freq
A_1A_1	P_{11}	1
A_1A_2	P_{12}	0.5
A_2A_2	P_{22}	0

Allele frequency

$$\begin{aligned} p_1 &= 1 \times P_{11} \\ &\quad + 0.5 \times P_{12} \\ &\quad + 0 \times P_{22} \end{aligned}$$

The variance

If μ is the mean of X, then its variance is

$$V[X] = E[(X - \mu)^2] \tag{1}$$

For our loaded die, the mean was $\mu = 3$. The variance is

$$\begin{aligned} V[X] &= (1-3)^2 \times 0.25 \\ &\quad + (2-3)^2 \times 0.25 \\ &\quad + (3-3)^2 \times 0.125 \\ &\quad + (4-3)^2 \times 0.125 \\ &\quad + (5-3)^2 \times 0.125 \\ &\quad + (6-3)^2 \times 0.125 \\ &= 3 \end{aligned}$$

A single toss of an unfair coin

The probability of "heads" is an unknown value p.
Your winnings: $X = 1$ for heads and $X = 0$ for tails. What's the probability distribution of X? The mean? The variance?

Properties of expectations

If X and Y are random variables and a is a constant,

$$E[a] = a \tag{2}$$
$$E[aX] = aE[X] \tag{3}$$
$$E[X + Y] = E[X] + E[Y] \tag{4}$$

See JEPr for details.

Using rules of expectations to re-express the variance

Let $\mu = E[X]$. The variance of X is

$$
\begin{aligned}
V &= E[(X-\mu)^2] && \text{(by Eqn. 1)} \\
&= E[X^2 - 2\mu X + \mu^2] && \\
&= E[X^2] - E[2\mu X] + E[\mu^2] && \text{(by Eqn. 4)} \\
&= E[X^2] - E[2\mu X] + \mu^2 && \text{(by Eqn. 2)} \\
&= E[X^2] - 2\mu E[X] + \mu^2 && \text{(by Eqn. 3)} \\
&= E[X^2] - 2\mu^2 + \mu^2 && \text{(by definition of } \mu\text{)} \\
&= E[X^2] - \mu^2 && (5)
\end{aligned}
$$

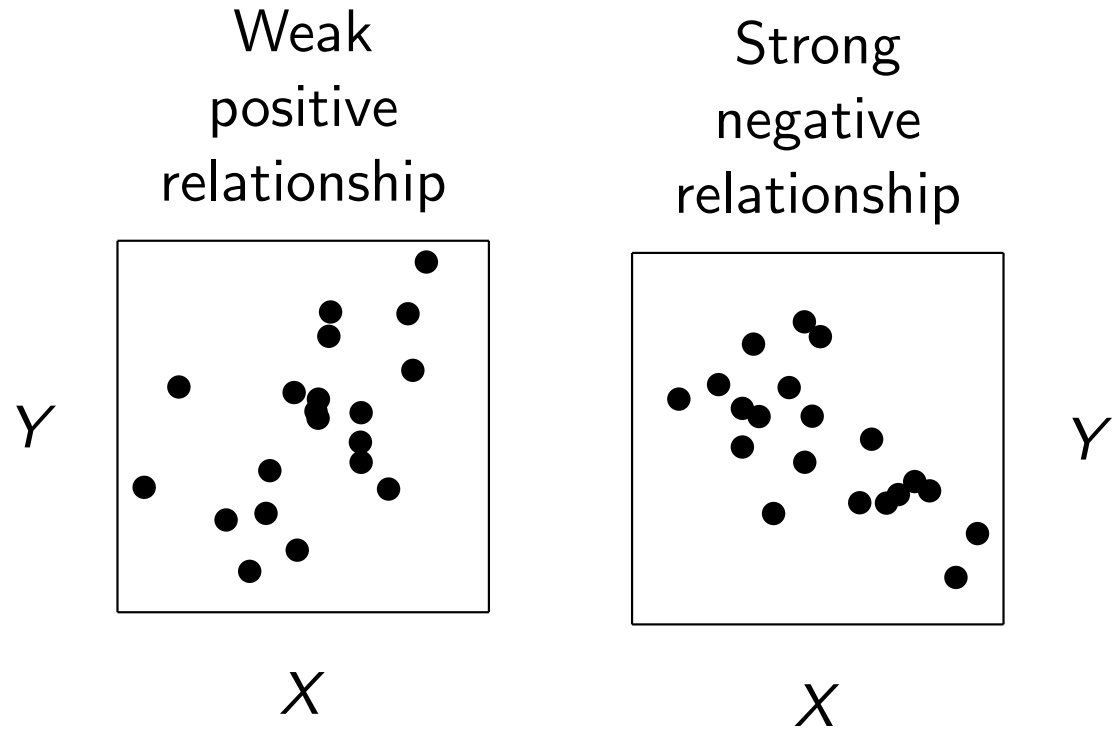
Variance of our loaded die

A moment ago, we found for our loaded die that $E[X] = V[X] = 3$. Let us recalculate this using Eqn. 5. We need

$$\begin{aligned} E[X^2] &= 1^2 \times 0.25 + 2^2 \times 0.25 \\ &\quad + 3^2 \times 0.125 + 4^2 \times 0.125 \\ &\quad + 5^2 \times 0.125 + 6^2 \times 0.125 \\ &= 12 \end{aligned}$$

$$V = E[X^2] - \mu^2 = 12 - 3^2 = 3$$

Association between variables



Positive and negative relationships between variables.

Covariance: a measure of association

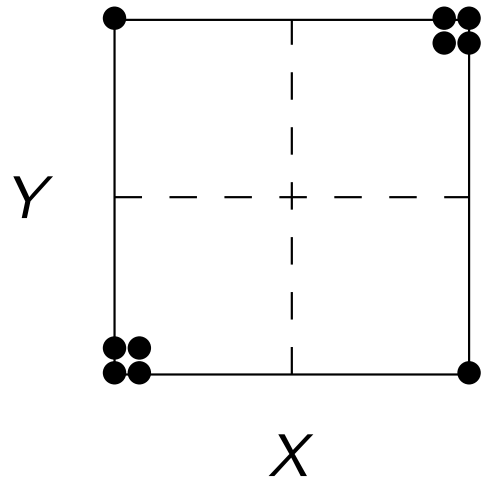
$$
\begin{aligned}
C(X, Y) &= \sum_{x,y} (x - E[X])(y - E[Y]) P_{x,y} \\
&= E\Big[(X - E[X])(Y - E[Y])\Big] \\
&= E[XY] - E[X]E[Y]
\end{aligned}
$$

When X and Y are independent, $C(X, Y) = 0$.

A bivariate probability distribution

X	Y	$P_{X,Y}$	$(X - E[X])(Y - E[Y])$
0	0	0.4	+0.25
0	1	0.1	–0.25
1	0	0.1	–0.25
1	1	0.4	+0.25

Note: the $P_{X,Y}$ column lists the probabilities of the (X, Y) pairs. In column 4, $E[X] = E[Y] = 0.5$.



Numerical value of covariance in previous slide

$$
\begin{aligned}
C(X, Y) &= 0.4 \times 0.25 \\
&\quad - 0.1 \times 0.25 \\
&\quad - 0.1 \times 0.25 \\
&\quad + 0.4 \times 0.25 \\
&= 0.15
\end{aligned}
$$

Probability Distributions

Alan R. Rogers

August 13, 2019

Probability distributions

A probability distribution is a function.

Input event

Output probability of event

- So far we have described probability distributions using tables.
- When events are numbers, distributions can be expressed as mathematical functions.

The Urn Metaphor

Imagine two urns: metaphors for a population in two successive generations. Urn 1 has 50 balls, some red, some white, representing parental gene copies. Urn 2 is empty until urn 1 has "reproduced" as follows:

1. Examine a random ball from urn 1.
2. Put a ball of the same color into urn 2.
3. Replace the ball from urn 1.
4. Repeat until there are 50 balls in urn 2.

The number of red balls in urn 2 is likely to differ from that in urn 1, because of random sampling. This metaphor is used as a model of genetic drift.

Binomial random variable

In probability theory, the number of red balls in urn 2 is a *binomial random variable*.

1. Balls drawn from the urn are statistically independent.
2. Each ball is red with probability p, the fraction of red balls in urn 1.

Probability of HT

Consider tosses of an unfair coin, for which the probability of "heads" is p and that of "tails" is $q = 1 - p$. Assume that the tosses are statistically independent.

Experiment	Toss a coin 2 times.
Result	HT
Probability	pq

This is an event of form $\Pr[A\&B]$, where A is the event that the first toss is H and B is the event that the 2nd is T. By assumption, $\Pr[A] = p$ and $\Pr[B] = q$. The tosses are statistically independent, so

$$\Pr[A\&B] = pq$$

by the multiplication law of probability.

Probability of HHT

Experiment	Toss a coin 3 times.
Result	HHT
Probability	p^2q

Probability of 2 heads in 3 tosses

There are 3 ways to get 2 heads in 3 tosses:

Event	Probability
THH	p^2q
HTH	p^2q
HHT	p^2q

The probability of 2 heads in 3 tosses is

$$\begin{aligned} P_2 &= 3p^2q \\ &= \binom{3}{2}p^2q \end{aligned}$$

where $\binom{3}{2}$ is pronounced "3 choose 2" and means the number of ways to choose 2 items out of a collection of 3.

Binomial distribution

The probability of x heads in K tosses is

$$P_X = \binom{K}{x} p^x q^{K-x}$$

$$E[X] = pK \qquad \text{mean}$$

$$V[X] = Kpq \qquad \text{variance}$$

Poisson distribution

Consider the lineage that connects me to an ancestor who lived t generations ago. The expected number of mutations along that lineage is $\lambda = ut$, where u is the mutation rate per generation. The number of mutations is a random variable (r.v.). If the mutation rate is constant, then the distribution of this r.v. is *Poisson*.

Poisson distribution function

If X is a Poisson-distributed r.v. with mean λ, then X takes value x with probability

$$P_x = \frac{\lambda^x e^{-\lambda}}{x!}$$

where e is the base of natural logarithms and $x!$ is "x factorial," or $x \cdot (x-1) \cdot (x-2) \cdots 1$.

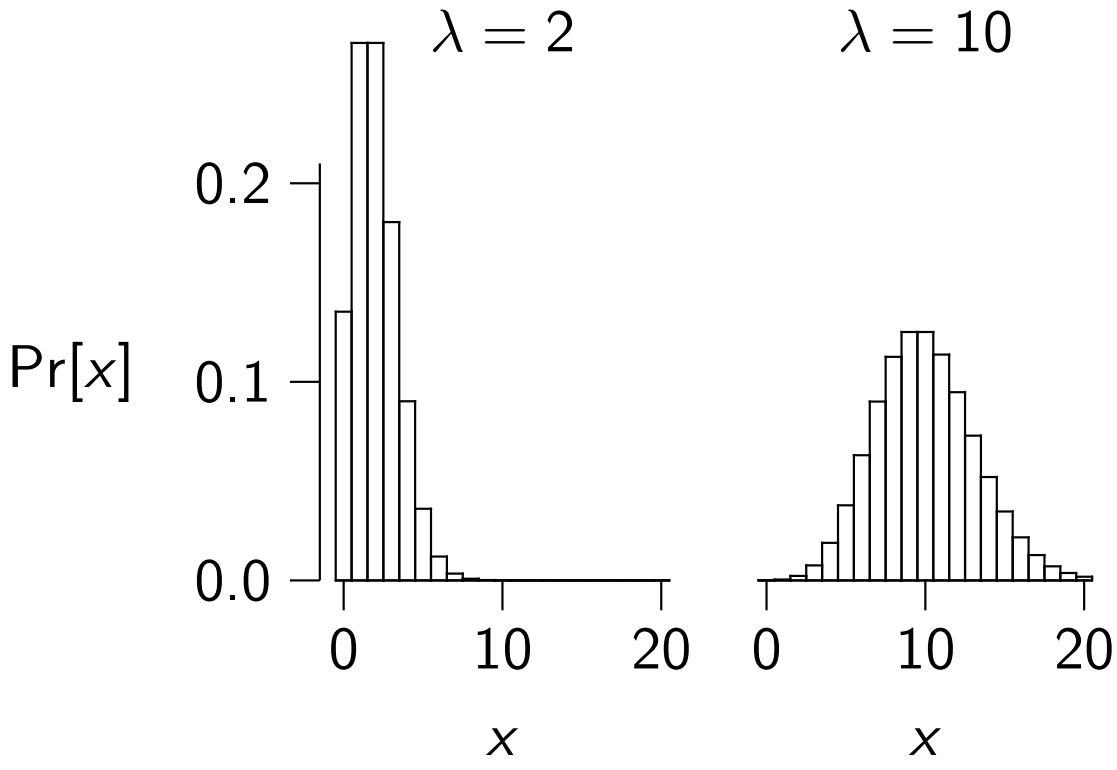
Mean equals variance.

$$E[X] = V[X] = \lambda$$

What is P_0? (Hint: $0! = 1$ and $\lambda^0 = 1$.)

$$P_0 = e^{-\lambda}$$

Poisson distribution



Poisson distribution functions

Mutation rates at autosomal nucleotide sites are roughly 10^{-9} per year. Consider a nucleotide in you. If you could trace its ancestry back across the last 10^9 years, what is the probability that you would find no mutations?

The expected number of mutations is $\lambda = ut$, where $u = 10^{-9}$ and $t = 10^9$. Thus, $\lambda = 1$. The probability of no mutations is

$$e^{-1} \approx 0.37$$

Raisin data

```
Date: Aug 28, 2009
N=41, Mean=21.756098, Var=26.239024, Max=33
 1- 3: *
 4- 6: *
 7- 9:  *
10-12: -*
13-15: --*
16-18: -----*-
19-21: ----------  *
22-24: -------*--
25-27: ------- *
28-30: --- *
31-33: *

Key: ---- Poisson distribution w/ mean 21.756098
         * Data
```

Raisin data

```
Date: Sept 6, 2013
N=32, Mean=20.375000, Var=15.080645, Max=34
 1- 3: *
 4- 6: *
 7- 9: *
10-12: *
13-15: --*
16-18: -------  *
19-21: --------*
22-24: -------   *
25-27: --*-
28-30: *
31-33: *

Key: ---- Poisson distribution w/ mean 20.375000
         * Data
```

Raisin data

```
Date: Sept 6, 2017
N=36, Mean=14.111111, Var=13.473016, Max=21
 1- 3: *
 4- 6: *
 7- 9: ---  *
10-12: --------*
13-15: --------*--
16-18: --------  *
19-21: --- *

Key: ---- Poisson distribution w/ mean 14.111111
         * Data
```