

Overview/reminder of basic concepts in statistics and genetics

SISG, Visscher-Goddard Module

Random variables, expected values and (co)variance

A discrete random variable can assume only a countable number of values

Probability mass function:

$$p(x) = P(X = x)$$

Expected value:

$$\mu = E(X) = \sum xp(x)$$

As a function of random variable:

$$E[h(X)] = \sum h(x)p(x)$$

Variance:

$$Var(X) = E[(X - \mu)^2]$$

Random variables, expected values and (co)variance

A discrete random variable can assume only a countable number of values

allele x :

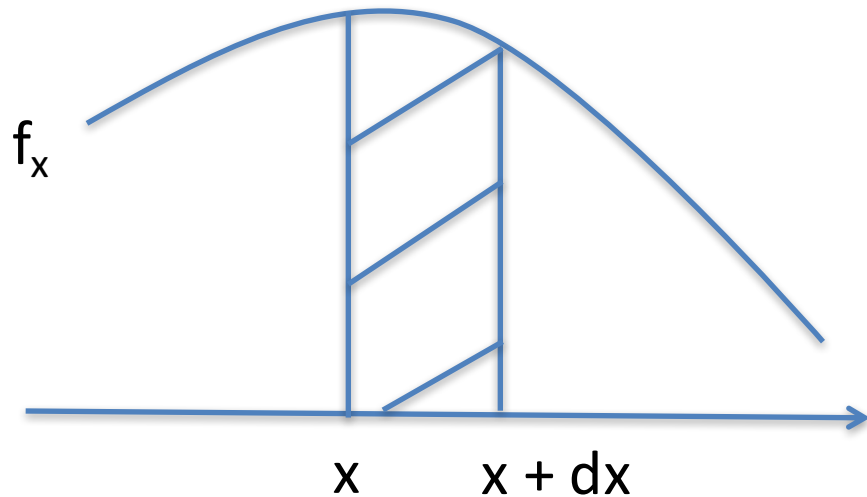
$$p(x) = \left\{ \begin{array}{ll} p & x = 1 \\ 1 - p & x = 0 \end{array} \right\}$$

$$E(X) = 0(1 - p) + 1(p) = p$$

$$Var(X) = p - p^2 = p(1 - p)$$

Random variables, expected values and (co)variance

A continuous random variable can be any value within a range



probability of being in shaded area

$$= f_X(x)dx$$

the interval should contribute

$$= xf_X(x)dx$$

the expected value and variance

$$E(X) = \mu_X = \int_{-\infty}^{\infty} xf_X(x)dx$$

$$Var(X) = E((X - \mu_x)^2)$$

Random variables, expected values and (co)variance

Covariance

Let X and Y be a pair of continuous random variables, with respective means μ_x and μ_y . The expected value of $(X - \mu_x)(Y - \mu_y)$ is called the covariance between X and Y .

$$\text{Cov}(X, Y) = E\left[(X - \mu_x)(Y - \mu_y)\right]$$

If the random variables X and Y are independent, then the covariance between them is 0. However, the converse is not true.

Summary (co)variance rules

$$\text{Var}(x) = E[x - E(x)]^2$$

$$\text{Var}(cx) = c^2 \text{Var}(x)$$

$$\text{Var}(x + y) = \text{Var}(x) + \text{Var}(y) + 2\text{Cov}(x, y)$$

$$\text{Var}(x + c) = \text{Var}(x)$$

$$\text{Cov}(x, y) = E[(x - E(x))(y - E(y))]$$

$$\text{Cov}(cx, y) = c\text{Cov}(x, y)$$

$$\text{Cov}(x, y + z) = \text{Cov}(x, y) + \text{Cov}(x, z)$$

Bayes' Theorem

Identify people who are liable to suffer from a genetic disease later in life.

1 in 1000 people are a carrier of the disease

No test is perfect - probability that a carrier tests negative is 1%

- probability that a non-carrier tests positive is 5%

A = the event “the patient is a carrier”

B = the event “the test result is positive”

Hence: $P(A) = 0.001$; $P(A') = 0.999$; $P(B|A) = 0.99$; $P(B|A') = 0.05$

A patient has a positive result. **Q:** What is the probability that the patient is a carrier?

Answer

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A')P(A')} \\ &= \frac{0.99 * 0.001}{(0.99 * 0.001) + (0.05 * 0.999)} = 0.0194 \end{aligned}$$

Hardy-Weinberg equilibrium

Mathematical relation between **allele** frequencies and the **genotype** frequencies is:

$$AA: p^2 \quad Aa: 2pq \quad aa: q^2$$

Allele
Frequency

A
p

a
q

Allele Frequency

A p

a q

AA	Aa
p^2	pq
aA	aa
qp	q^2

HWE and SNPs

If SNP genotypes are coded $X = 0, 1$ and 2 (alleles) and the allele frequency is p , then:

$$E(X) = (1-p)^2 * 0 + 2p(1-p) * 1 + p^2 * 2 = 2p$$

$$\text{var}(X) = (1-p)^2 * (0-2p)^2 + 2p(1-p) * (1-2p)^2 + p^2 * (2-2p)^2 = 2p(1-p)$$