Probability and expectations

Lecture 2 ▪ Quantitative Genetics

Grum Gebreyesus ▪ ASAP-Bio Knowledge Hub

asap-bio.github.io/lec-qg-probability.html

Contents

1 **Outcomes, events and probability** 1

2 **Combining events** 2

3 **Conditional probability and independence** 2

4 **Distributions** 3

5 **Expectations** 3

6 **Variance** 3

7 **Covariance, correlation and regression** 4
- 7.1 Covariance . . . 4
- 7.2 Correlation . . . 4
- 7.3 Regression . . . 4

8 **Conditional expectations** 4

9 **Reference: calculus of means and variances** 5
- 9.0.1 Means . . . 5
- 9.0.2 Variances and covariances . . . 5

10 **Exercises** 5

1 Outcomes, events and probability

The previous lecture ended with an uncomfortable fact: we cannot observe the effect of a single gene on a metric trait, so we must work with probabilities and expectations. This lecture assembles exactly the probability tools the rest of the course needs, and no more.

THE THREE BASIC OBJECTS

- An **outcome** is a possible result of an "experiment".
- An **event** is a set of outcomes.
- A **random variable** is a variable that can take different values (outcomes), each with some probability.

Write Ω for the set of all outcomes. Rolling a fair six-faced die, $\Omega = \{1, 2, 3, 4, 5, 6\}$. The event "the roll is even" is the subset $\{2, 4, 6\}$, and its probability is $3/6 = 1/2$.

2 Combining events

Two events can be combined in two ways, and the language matters because it maps directly onto genetic questions later ("this allele *and* that allele", "this genotype *or* that one").

- $A \cap B$, read "A **and** B", the outcomes in both.
- $A \cup B$, read "A **or** B", the outcomes in either.

PROBABILITY OF A COMPOSITE EVENT

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

Subtracting $\Pr(A \cap B)$ prevents double-counting the overlap.

Two events that can never occur together ($A \cap B = \varnothing$) are **mutually exclusive**. Then $\Pr(A \cap B) = 0$ and the rule simplifies to

$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$

Worked example: a fair die

Roll one fair six-faced die. Let A be "the roll is at least 4" and B be "the roll is even".
$A = \{4, 5, 6\}$, so $\Pr(A) = 3/6 = 1/2$. $B = \{2, 4, 6\}$, so $\Pr(B) = 1/2$. $A \cap B = \{4, 6\}$, so $\Pr(A \cap B) = 2/6 = 1/3$.
Hence $\Pr(A \cup B) = \frac{1}{2} + \frac{1}{2} - \frac{1}{3} = \frac{2}{3}$.
Contrast with mutually exclusive events: if $A = \{1\}$ and $B = \{2, 4, 6\}$, then $\Pr(A \cap B) = 0$ and $\Pr(A \cup B) = 1/6 + 3/6 = 2/3$ with nothing to subtract.

3 Conditional probability and independence

If the occurrence of one event affects the probability of another, the events are **dependent**. The probability of A *given that* B has occurred is

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

Rearranged, this gives the multiplication rule $\Pr(A \cap B) = \Pr(A \mid B)\,\Pr(B)$.

INDEPENDENCE

A and B are **independent**, written $A \perp B$, when knowing one tells you nothing about the other:

$$\Pr(A \mid B) = \Pr(A) \quad \iff \quad \Pr(A \cap B) = \Pr(A)\,\Pr(B)$$

Worked example: haplotype frequencies

A haplotype MQ has frequency f_{MQ}; allele M has frequency p_M and Q has frequency p_Q. If the two loci were independent (no linkage disequilibrium), we would expect $f_{MQ} = p_M\,p_Q$. The conditional probability of carrying Q given M is $\Pr(Q \mid M) = f_{MQ}/p_M$. Under independence this equals p_Q. Any departure is exactly what we later call **linkage disequilibrium**, so LD is nothing more mysterious than non-independence of two events.

4 Distributions

A **distribution** specifies the probabilities of the different values a random variable X can take.

- **Discrete case.** We specify $\Pr(X = x_i)$ for each possible value x_i. The binomial distribution is the standard example, and it is the natural model for counting alleles.
- **Continuous case.** We cannot give a positive probability to a single point, so we specify $\Pr(X \in [x_0, x_1])$, the probability of falling in an interval, via a density function.

To describe where a distribution sits, we use the **mode** (most frequent value), the **median** (middle value) and the **mean** or expected value. For the symmetric, unimodal distributions of quantitative genetics these coincide, but it is the mean that has the algebra we need.

5 Expectations

In most situations the average of a sample approaches the expected value as the sample grows. The expectation is the theoretical counterpart of the sample mean.

$$\text{Discrete:} \quad E(X) = \sum_{i=1}^{n} x_i \Pr(X = x_i) \qquad\qquad \text{Continuous:} \quad E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

The two properties used constantly in this course are that expectation is **linear**, and that it passes through sums regardless of dependence:

$$E(aX + b) = aE(X) + b \qquad \text{and} \qquad E(X + Y) = E(X) + E(Y)$$

APPLICATION

The second identity is the reason we could write $E(g) = E\left(\sum_j g_j\right) = \sum_j E(g_j)$ in the previous lecture *without* assuming the loci were independent. Expectations always add. Variances do not, as shown in the next section.

6 Variance

Variance measures dispersion, how spread out a distribution is around its mean.

$$V(X) = \sigma_X^2 = E\left[(X - \mu_X)^2\right] = E(X^2) - [E(X)]^2$$

The second form is usually the easier one to compute. Two rules follow, and the contrast between them is the single most important thing in this section:

$$\begin{aligned} V(aX + b) &= a^2 V(X) \\ V(X + Y) &= V(X) + V(Y) + 2\,\mathrm{Cov}(X, Y) \\ V(X + Y) &= V(X) + V(Y) \qquad \text{only if } X \perp Y \end{aligned}$$

ADDITIVITY OF VARIANCES

Adding a constant b does not change the variance, but scaling by a multiplies it by a^2. And variances of sums only add when the variables are **uncorrelated**. When we later decompose $\sigma_y^2 = \sigma_A^2 + \sigma_D^2 + \sigma_E^2$, we are quietly relying on exactly this condition, and when it fails, for instance under genotype-environment correlation, the decomposition acquires an extra covariance term.

7 Covariance, correlation and regression

Some variables tend to vary together. Three related measures describe this.

7.1 Covariance

$$\sigma_{XY} = \mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y)$$

A covariance is positive when the two tend to be large together, negative when one is large as the other is small. Its size depends on the units of both variables, which makes it awkward to interpret on its own.

7.2 Correlation

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{V(X)\,V(Y)}}, \qquad -1 \leq \rho_{XY} \leq 1$$

Correlation is the covariance standardised by both standard deviations, hence unit-free and bounded. This is the form used for genetic correlations between traits.

7.3 Regression

$$b_{Y|X} = \mathrm{Reg}(Y \mid X) = \frac{\mathrm{Cov}(X, Y)}{V(X)}$$

The regression coefficient is the covariance standardised by the variance of the *predictor only*. It answers: if X increases by one unit, by how much does Y change on average? If the relationship really is linear, $y = b_{Y|X}x + a$, then knowing $X = x$ gives

$$E(Y \mid X = x) = b_{Y|X}\,x + a$$

WHERE EACH ONE REAPPEARS

Covariance between relatives is how we measure resemblance and estimate heritability. **Correlation** is how we express genetic relationships between traits. **Regression** is the parent-offspring regression, and it is also the form the breeder's equation ultimately takes. Three faces of the same quantity.

8 Conditional expectations

Just as probabilities can be conditioned on an event, so can expectations. If we know an event has occurred, the expectation is modified accordingly:

$$E(Y \mid A) = \sum_i y_i \,\Pr(Y = y_i \mid A)$$

This is the exact form in which the one-locus model is written. When we say the expected phenotype of an AA animal is $E(y \mid AA) = x_{AA}$, that is a conditional expectation: the mean of the trait *given* the genotype. The whole one-locus model is built from three such conditional expectations, one per genotype.

9 Reference: calculus of means and variances

These identities are worth keeping to hand; almost every derivation later in the course uses one of them. Throughout, X, Y, Z are random variables, a, b are constants, and $X \perp Y$ means X and Y are independent.

Means

$$
\begin{aligned}
E(X) &= \textstyle\sum_i x_i \Pr(X = x_i) & E(aX) &= aE(X) \\
E(a) &= a & E(aX+b) &= aE(X) + b \\
E(X+Y) &= E(X) + E(Y) & E(XY) &= E(X)E(Y) \ \text{ if } X \perp Y
\end{aligned}
$$

Variances and covariances

$$
\begin{aligned}
V(X) &= E(X^2) - [E(X)]^2 & V(a) &= 0 \\
V(aX+b) &= a^2 V(X) & V(X+Y) &= V(X) + V(Y) + 2\,\mathrm{Cov}(X,Y) \\
\mathrm{Cov}(X,Y) &= E(XY) - E(X)E(Y) & \mathrm{Cov}(X,X) &= V(X) \\
\mathrm{Cov}(aX,bY) &= ab\,\mathrm{Cov}(X,Y) & \mathrm{Cov}(X,Y) &= 0 \ \text{ if } X \perp Y \\
\mathrm{Cov}(X+Y,Z) &= \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)
\end{aligned}
$$

10 Exercises

Use the language of events and the equations above, not intuition alone. Showing the events explicitly is the point of the exercise.

Exercise 2.1

Die roll. Roll two regular six-faced dice, one blue and one red.

1. What is the probability that the blue die shows a 4?
2. What is the probability that the blue die shows at least 4?
3. What is the probability that the red die shows 4, given that the blue die shows at least 4?
4. What is the probability that both dice show at least 4?
5. What is the probability that at least one of the dice shows at least 4?
6. What is the probability that one die shows at least 4 and the other shows less than 4?

1. $B = \{4\}$, $\Pr(B) = 1/6$.
2. $B = \{4, 5, 6\}$, $\Pr(B) = 3/6 = 1/2$.
3. $R = \{4\}$, $B = \{4, 5, 6\}$. $\Pr(R \mid B) = \Pr(R \cap B)/\Pr(B) = \Pr(R) = 1/6$, because the two dice are independent ($R \perp B$).
4. $R = \{4, 5, 6\}$, $B = \{4, 5, 6\}$: $\Pr(R \cap B) = \Pr(R)\Pr(B) = \frac{1}{2} \times \frac{1}{2} = 1/4$.
5. $\Pr(R \cup B) = 1 - \Pr(\bar{R} \cap \bar{B}) = 1 - \Pr(\bar{R})\Pr(\bar{B}) = 1 - \frac{1}{2} \times \frac{1}{2} = 3/4$.
6. Two disjoint possibilities, so $\Pr\Big[(R \cap \bar{B}) \cup (\bar{R} \cap B)\Big] = \Pr(R)\Pr(\bar{B}) + \Pr(\bar{R})\Pr(B) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = 1/2$.

Note how 5 is easiest via the complement, and 6 via mutually exclusive events. Choosing the right decomposition is most of the work.

Exercise 2.2

Birthday. Assume the same number of people are born on every day of the year, and consider a non-leap year. A given person was born on the 31st of some month. What is the conditional probability that this person was born in July? Demonstrate the result using the language of conditional probability.

Let D be "born on the 31st" and J be "born in July". Seven months have a 31st, so $\Pr(D) = 7/365$. The intersection $J \cap D$ is simply "born on 31 July", so $\Pr(J \cap D) = 1/365$.

$$\Pr(J \mid D) = \frac{\Pr(J \cap D)}{\Pr(D)} = \frac{1/365}{7/365} = \frac{1}{7}$$

The intuitive answer, 1/7, is correct here, but only because the seven months with a 31st are equally likely. The conditional-probability machinery is what tells you *when* the intuitive shortcut is safe.

Exercise 2.3

Show, starting from the definition $\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$, that $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$. Then use it to confirm that $\mathrm{Cov}(X,X) = V(X)$.

Expand the product inside the expectation and use linearity:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E[XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y] \\
&= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X\mu_Y \\
&= E(XY) - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y \\
&= E(XY) - E(X)E(Y)
\end{aligned}$$

Setting $Y = X$ gives $\mathrm{Cov}(X,X) = E(X^2) - [E(X)]^2 = V(X)$, which is the definition of variance. Variance is just covariance of a variable with itself.