MATHEMATICAL METHODS IN BIOLOGY

PART 2

EXERCISES

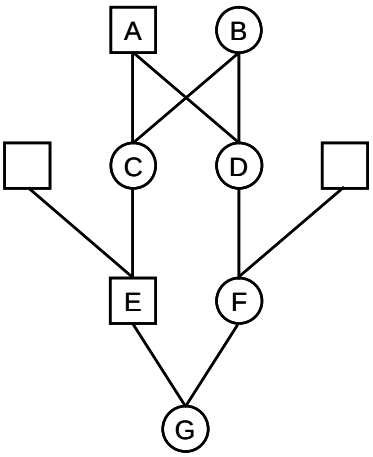
Eva Kisdi
Department of Mathematics and Statistics
University of Helsinki

EXERCISES 1-10: CALCULUS OF PROBABILITIES

1. *Genetic risks.* A couple is facing the risk that their children may suffer from a genetic disorder, because both the husband and the wife are known to be heterozygote (*Aa*) carriers of the harmful recessive allele *a* (the homozygote *aa* individuals are affected by the disorder, all others are healthy). The couple plans to have two children, and wants to know the prospects for their health: What is the probability that both children will be healthy?

2. *Marriage of relatives.* Close relatives are not allowed to marry but if they do, their children are very often affected by serious genetic problems. This is because most of us carry 2-3 recessive lethal genes and some more that are not lethal but harmful. In unrelated people, these harmful recessive alleles are most likely in different loci so that their children are healthy heterozygotes. Descendants of a single person may however carry harmful recessives in the *same* locus; and if they marry, their children can be recessive homozygotes.

The figure shows a pedigree with marriage between cousins. A and B are the shared grandparents; C and D are sisters, who marry unrelated persons; E and F are cousins; and G is their child. *a* denotes a harmful recessive allele present in A. Calculate the probability that G inherits *a* from both parents and is therefore a recessive homozygote *aa* exhibiting the symptoms of the disorder caused by *a*. Next, calculate the probability that G is *not* a recessive homozygote for any of the 3 unlinked harmful alleles that A carried and also not for another 3 unlinked alleles that were present in B. (In reality we cannot know that A and B carry exactly 3 harmful alleles each, but this illustrates the probability of having a healthy child in a cousin marriage with roughly realistic numbers.)



3. *Did Mendel cheat?* Mendel found that the seven traits of garden pea he studied were inherited independently, i.e., in modern terms as if all seven traits were on separate chromosomes. The garden pea happens to have seven chromosomes. Assuming that the chromosomes are equally long, calculate the probability that seven randomly picked loci are all on different chromosomes.

> Actually, this is not the case. The seven traits are seed shape (smooth/wrinkled); seed colour (yellow/green); seed coat colour (white/coloured); pod shape (smooth/constricted); pod colour (yellow/green); flower position (terminal/along stem); plant height (long/short). Of these, seed colour and seed coat colour are on chromosome 1, pod shape, flower position and plant height are on chromosome 4, and only the other traits are on separate chromosomes. The loci on the same chromosome are however far apart, so that they are inherited independently, except pod shape and plant height. Mendel did not perform all possible dihybrid crosses and so he happened not to investigate the pod shape - plant height dihybrid cross. (How many pairs of traits are there to test independent inheritance for?)

4. *Population genetics of sex-linked traits. Drosophila* males have a single *X* chromosome whereas females have two. White eye is a recessive X-linked trait of *Drosophila*. Suppose that we cross white-eyed females with wild (red-eyed) males.

Calculate the frequency of genotypes and the frequency of the white allele (*p*) in the females and in the males of the next generation. Next, suppose that we let the females and males born from the previous cross mate randomly among themselves. Plot the frequency of the white allele in females and in males for several generations, and calculate the equilibrium frequency of phenotypes.

5. *Test of independence.* An ecologist collects presence-absence data of two different species of plants in sample quadrats, and wants to know whether the plants occur independently. The data are

(a) Both plants: 25% of quadrats, plant A only: 25%, plant B only: 5%
(b) Both plants: 15% of quadrats, plant A only: 35%, plant B only: 15%

6. *Total probability.* About 33% of African American people develop high blood pressure during their lives, whereas in people of Caucasian origin, this occurs only with probability 25%. In a town where 40% of people are African Americans and the rest are Caucasians, what percentage of people will need care for high blood pressure?

7. *Blood transfusion.* Current medical protocols of blood transfusion require matching blood types of the AB0, Rh+/-, and also other minor blood groups, but before blood groups were discovered, transfusion was risky due to the blood group incompatibilities: If the donor has an antigen (A or B in the AB0 system) that the recipient does not have, then the recipient's body produces an immune reaction that is easily fatal. In the AB0 system, people with blood group 0 or A do not have antigen B and therefore may not receive B or AB blood; similarly people with blood group 0 or B may not receive A or AB blood. Blood of type 0 can be given to anyone and people of blood group AB may receive any blood.

In Finland, the frequencies of blood groups are

A	44.2%
B	16.6%
AB	8.1%
0	31.2%

Calculate what would be the probability that a blood transfusion is fatal if the AB0 blood groups were not known.

8. If a rare mutation is present with frequency *q* in a large population, how many individuals does one need to sample in order to be 99% sure that the sample contains at least one mutant for investigation?

9. *Eigen's paradox.* In prebiotic conditions, where polynucleotides replicated without enzymes, the probability of mutation per nucleotide could not be less than 10^{-2}. A sequence must produce at least one mutation-free copy out of ca 5 copies if it is to be maintained. What is the maximum length of a sequence (primitive "genome") that can be copied faithfully enough?

> The answer is a number much less than the length of the smallest genome. This is known as *Eigen's paradox*: The primitive genome is not long enough to code for an enzymatic replication system, but without enzymatic replication, the genome cannot be longer!

10. *The Luria-Delbrück fluctuation test.* This test is a simple method to estimate mutation rates in bacteria. Suppose we want to measure the rate of mutation that gives resistance against some toxic material. First, we inoculate 20 test tubes with a small number of non-resistant (wild type) bacteria, and grow the cultures in normal medium to 10^8 cells/ml. Then we take a 0.1 ml sample of each of the 20 cultures. The samples are spread on plates that contain the toxic substance, hence only resistant bacteria can grow. Out of the 20 samples, we find that 11 samples did not contain any mutant (no resistant bacteria found). Calculate the mutation rate.

Hints: figure out the number of bacteria in the sample. Because each tube started with a small number of bacteria but ended with a large number of them, almost every bacterium in the sample is the product of a cell division: hence the number of divisions is approximately the same as the number of cells. If μ is the probability of mutation in one division, you can calculate the probability that no mutation has occurred in any of the divisions leading to the cells in the sample; and this must match the fraction of samples found to be mutation-free.

EXERCISES 11-13: BAYES' THEOREM

11. *Rare disease screening.* A medical test picks out a disease in 100% of the cases when it really occurs, but also produces positive results in 5% of healthy people (false positives). The disease is known to affect 0.1% of the population. Your test comes back positive. What is the probability that you really have the disease?

12. *Prior vs posterior probabilities.* The frequency of identical (monozygotic) twin births among all human births is about 0.3% and is fairly constant over time and across populations. The frequency of fraternal (dizygotic) twins was about 1.7% a generation ago in the US.

(a) Calculate the prior probability that a pair of twins is monozygotic and the posterior probability that they are monozygotic if we know that they are of the same sex.

(b) The frequency of fraternal twins is increasing (the present value in the US is around 3%). How does this modify the prior probability of twins being identical? How does the posterior probability change if the prior decreases?

13. *Bayesian statistics.* In this problem, we consider a simple coin-tossing experiment for clarity. However, the same kind of statistics is now used widely from phylogenetics to artificial intelligence. In principle, it can be used any time when we want to estimate parameters from experimental data.

Let q be the probability that when tossing a coin, we get a tail, and let 1-q be the remaining probability of getting a head. The coin may be fair (which gives tails and heads equally often, i.e., q=0.5) but may also be loaded either in favour of tails (q>0.5) or in favour of heads (q<0.5). We want to estimate parameter q of a given coin. To this end, we toss our coin 6 times. Suppose that we get 6 tails in a row. The traditional estimate would then be q=6/6=1, predicting that this coin will always give a tail. But we are not happy with this point estimate, because we had a pretty strong belief that our coin is fair (q=0.5). The Bayesian way of formalizing our prior "belief" is to specify a prior probability of the coin being fair and also the probability of deviating from fairness to different degrees.

For simplicity, here we assume that there are only three types of possible coins:

fair coins (q=0.5)
tail coins (q=1), and
head coins (q=0)

[This simplifying assumption is of course unrealistic for the coin experiment; we can relax this assumption later.]

(a) Our confidence in having a fair coin is formalized in the prior probabilities

P(fair coin) = 0.98
P(tail coin) = 0.01
P(head coin) = 0.01

Calculate the posterior probability of having a fair coin given that we have tossed 6 tails out of 6 trials. Calculate the posterior probabilities of head and tail coins, too.

(b) Repeat (a) with the uniform distribution as prior,

P(fair coin) = 1/3
P(tail coin) = 1/3
P(head coin) = 1/3

Observe how the posterior probabilities change due to changing the prior when you go from the uniform prior to the prior in (a).

> The uniform prior is often called the "uninformative prior", which does not "bias" the resulting posterior probabilities, and is often used for this reason. This view is however questionable; knowing that all coin types are equally likely is information just as knowing it otherwise. Typical coins in our purses are much closer to fair than to "tail" or to "head", so in this particular example, the "uninformative" prior is not a very sensible choice. Sometimes the prior is evident from the context (for example, if the prior has to specify the probability that a randomly chosen person is male or female, then 50-50% will be taken as prior) whereas at other times the prior is fully *ad hoc* - but it does matter.

(c) What happens to the posterior probabilities if we exclude a possibility in the prior, e.g. we assume P(tail coin)=0? What happens if we have unshakeable faith in the coin being fair, i.e., we use the prior P(fair coin) = 1?

EXERCISES 14-21: BINOMIAL AND POISSON DISTRIBUTIONS

14. *Cohort survival.* $n = 5$ birds are born in the same year. Each bird survives one year with probability 0.7, and they are independent of one another.

(a) Plot the distribution ($P(k)$ against k) of the number of birds alive after 1 year; 3 years; and 10 years.

(b) How long do we have to wait to be 95% sure that none of the birds is alive?

> *Hint:* use Excel or similar software for the calculations in (a). In Excel, the factorial n! is computed by FACT(n). If working with a calculator, compute P(k) after 1 year and outline how the rest could be done.

15. *Offspring number of highly fecund organisms.* A tree may produce tens to hundreds of thousands of seeds during its life, but in an expanding (!) population, on average only 1.1 of its seeds survives to become an established mature tree. The probability of survival is therefore very small, and the number of surviving seeds per tree follows a Poisson distribution with expectation $\lambda = 1.1$.

Calculate the probabilities that a tree has 0, 1, 2, 3, ..., k surviving seeds. Plot the data as a histogram. Take k high enough such that the probability of having more than k seeds is less than 1%.

16. *Compare the binomial and Poisson distributions.* Use Excel or similar software to make a histogram of the binomial probabilities (i.e., plot $P(k)$ against k) using the parameters $n = 10$ and $p = 0.25$. Then increase n and decrease p such that you keep the expectation $\lambda = np = 2.5$ constant. Prepare a series of histograms of the binomial distribution with increasing n, and compare them to the histogram of the Poisson distribution with parameter $\lambda = 2.5$.

17. *Stochastic behaviour of membrane channels.* A membrane channel, when open, goes closed at a rate α, and when closed, it goes open at a rate β. Consider first a very large number N of channels. Let $x(t)$ be the probability that a channel is open at time t, such that the number of open channels is $Nx(t)$.

(a) Verify that the number of open channels changes according to the differential equation

$$\frac{dNx}{dt} = -\alpha xN + \beta(1-x)N$$

and since N is a constant, the probability of being open changes according to

$$\frac{dx}{dt} = -\alpha x + \beta(1-x)$$

Find the equilibrium probability of the channel being open (this is the equilibrium fraction of open channels), and evaluate it assuming $\beta = 2\alpha$ (opening is twice as fast as closing).

(b) Suppose a cell has only $n = 18$ channels. Calculate the probability that at equilibrium, $k = 12$ of these are open.

18. *Genetic mapping.* To establish the distance between two genes on a chromosome, one performs a testcross of double heterozygote and recessive homozygote parents:

$$\frac{AB}{ab} \times \frac{ab}{ab}$$

The offspring obtained from this cross are partly non-recombinant (having either both dominant alleles *A* and *B* or neither dominant allele) or recombinant (having only *A* or only *B*). Recombinant offspring are produced when there is a crossover between the loci during the meiosis of the double heterozygote parent.

Crossover can occur at any base pair, and therefore the potential number of crossovers is very large. However, the probability of a crossover at any given base pair is small, so that the actual number of crossovers is only a few, and the number of crossovers follows a Poisson distribution. The expectation of the Poisson distribution, λ, is proportional to the physical distance between the two loci.

If the double heterozygote parent had no crossover at all between loci *A* and *B*, all offspring will be non-recombinant. Somewhat surprisingly, *any* nonzero number of crossovers results in 50% recombinant offspring (see
http://www.ncbi.nlm.nih.gov/books/NBK21819/figure/A1116/
for a figure demonstrating this fact).

(a) Derive the fraction of recombinant offspring, RF = # of recombinant offspring / # of all offspring x 100%, and plot it as a function of the physical distance measured by λ.

(b) Traditionally, a map unit is defined such that 1 map unit corresponds to 1% recombinant offspring. Argue that this definition cannot be extended to large map distances associated with high percentages of recombinant offspring; in other words, finding 30% recombinant offspring does not imply that the two loci are 30 times farther apart than two loci that exhibit 1% recombinant offspring.

(c) Based on (a), calculate λ from the measured fraction of recombinant offspring, RF. Show that a distance of $(\lambda/2)\cdot 100$ map units is consistent with the traditional definition for small distances.

> $(\lambda/2)\cdot 100$ is called the distance in corrected map units. Calculating λ from RF and using the formula $(\lambda/2)\cdot 100$ yields the true map distance also for large distances, and eliminates the problem that several crossovers can happen between genes far apart yet these do not increase the fraction of recombinant offspring compared to a single crossover (cf the result in (a)).
>
> For part (c), use the approximation $e^{-x} \approx 1-x$ or $\ln(1-x) \approx -x$, which holds as long as *x* is small; you can prove this approximation by simple differentiation. Alternatively, you can demonstrate (c) just with numerical examples: show that $(\lambda/2)\cdot 100$ gives roughly the % value of RF if RF was low, but not if RF was high.

19. *Generalizing the Skellam model to perennial plants.* In the lecture, we constructed the model

$$x_{t+1} = 1 - \exp(-\alpha x_t) \qquad (1)$$

for the fraction of living sites *x* occupied by an annual plant species with per capita seed number α.

(a) Construct an analogous model for a perennial species, which matures at age 1 and survives each subsequent year with probability *p* (the model above is the special case *p* = 0). Adult plants are competitively superior to seedlings, i.e., a seed cannot survive in a site occupied by a surviving plant.

(b) Find the equilibrium fraction of occupied sites and using Excel, plot it as a function of fecundity (α) for several different values of *p*.

> *Hint:* it is not possible to solve the equilibrium equation explicitly. Instead, solve the equation for α and plot α as a function of the equilibrium fraction of occupied sites; and then swap the axes to plot the equilibrium fraction of occupied sites as a fraction of α.

20. *Coexistence by the competition-colonisation trade-off.* Generalise the Skellam model in equation (1) above to the case of two species. Both species are annual. Species 1 with fecundity α is competitively superior to species 2 with fecundity β, i.e., if a site contains seed(s) of both species, it will be occupied by species 1.

(a) Construct the model equations for the fraction of sites occupied by the superior and by the inferior species (x_{t+1} and y_{t+1}), respectively.

(b) As in the original Skellam model, the superior species has a positive equilibrium and therefore is said to be viable if $\alpha > 1$. Find the condition for the inferior species to be viable in the presence of the superior species, i.e., find the parameter region (α, β) where the two species coexist.

21. *The Nicholson-Bailey model of host-parasitoid systems.* For a parasitoid, a host individual is analogous to a living site for a seed in the Skellam model. Assume that each parasitised host can support the development of *B* parasitoid larvae, and parasitoid mothers deposit at least *B* eggs such that each attacked host will indeed release *B* parasitoids. Parasitised hosts die without reproduction, whereas non-parasitised hosts produce *F* offspring each. Both hosts and parasitoids are annual.

(a) Calculate the probability that a host individual avoids attack assuming that both the number of hosts (H_t) and the number of parasitoids (P_t) are large, and are of the same order of magnitude (i.e., the ratio of H_t and P_t is neither very large nor very small).

(b) Construct the model of population dynamics, i.e., write down the equations for H_{t+1} and P_{t+1} in terms of H_t and P_t.

> No one has been able prove what is the long-term behaviour of this model: it has an equilibrium which is always unstable, and periodic (or quasi-periodic)

solutions are not known. You may want to investigate the model numerically, by iterating the host and parasitoid densities from year to year and plotting P_t against H_t.

EXERCISES 22-25: MEAN AND VARIANCE

22. η is the average of 10 random numbers that we generate by casting a die. What is the expectation and the variance of η?

23. Let η_1 and η_2 denote the body weight we measure on a randomly chosen pair of identical twins. The body weight is the sum of a baseline weight, the effects of genes, and the effect of the environment. Identical twins have the same genes but have (partly) different environmental effects. Their phenotypes are therefore given by

$$\eta_1 = c + \xi + \varepsilon_1$$
$$\eta_2 = c + \xi + \varepsilon_2$$

where c is the constant baseline weight, ξ is the genetic value (same for both) and ε_1, ε_2 are the environmental deviations. ξ, ε_1, and ε_2 are independent, ε_1 and ε_2 are identically distributed, and the mean of ε_1 and ε_2 have been scaled to zero (see lecture).

(a) Calculate the covariance and the correlation coefficient between η_1 and η_2.

(b) The quotient $V(\xi)/V(\eta)$ is called the (broad-sense) *heritability*, which tells what fraction of the observable phenotypic variance ($V(\eta)$) is due to genetic effects. Based on (a), suggest a practical way to measure the heritability of body weight; and comment on whether this measurement is correct for the entire (non-twin) population.

24. *Variance vs the "accuracy" of measurement.* Suppose we would like to measure the frequency *p* of a certain trait (genotype, disorder, etc.) in a population. To this end, we take a sample of *n* individuals and count the number ξ of those in the sample who have the trait.

(a) Describe the conditions under which ξ is a binomially distributed random variable.

(b) One has the intuitive feeling that with larger sample size *n*, variation should somehow dampen and our measurement should become more accurate. Does the variance of ξ become smaller as *n* increases? Or the standard deviation, $D(\xi) = \sqrt{V(\xi)}$? Or the coefficient of variation, $c = D(\xi)/E(\xi)$? Or the variance of the estimated frequency, $V(\xi/n)$?

25. *Chemical reactions: how many molecules are "infinitely many"?* Suppose that there are *N* enzyme molecules and a large number *M* inhibitor molecules present in a well-mixed system. The corresponding concentrations are *x* for the free enzyme, *y* for the enzyme-inhibitor complex and *z* for the free inhibitor. The enzyme binds the

inhibitor at rate α and the inhibitor dissociates from the enzyme-inhibitor complex at rate β.

If *N* is sufficiently large, we can model this system with the differential equations

$$\frac{dx}{dt} = -\alpha z x + \beta y$$
$$\frac{dy}{dt} = \alpha z x - \beta y$$

We assume throughout that *z* is constant; if *M* is very much larger than *N*, then the change in the number of free inhibitor molecules is negligible even if all *N* enzyme molecules happen to bind an inhibitor.

Consider now the (realistic) case that *N* is not very large, and denote the number of free enzyme molecules (a random variable) with ξ. The number of enzyme-inhibitor complexes is then $N - \xi$.

(a) Show that ξ is binomially distributed and calculate its mean and variance in equilibrium.

(b) Say that the variation in ξ and in $N - \xi$ is negligible and the deterministic ODE model is applicable if the coefficient of variation of both ξ and of $N - \xi$ are less than 0.01. How large should *N* be to achieve this? (Recall that the coefficient of variation is $c(\xi) = \sqrt{V(\xi)} / E(\xi)$.)

EXERCISE 26: EXPONENTIAL DISTRIBUTION

26*. *How do stupid animals forage optimally?* (Based on Adler & Kotar (1999), Evolutionary Ecology Research 1:411-421.)

Assume that an animal arrives in a fresh patch of resource at time *t*=0, and the function $g(t) = 1 - e^{-\beta t}$ gives the amount of resource it has consumed by time *t* (the total resource content of a patch is scaled to 1). The animal leaves the patch at a constant rate α, such that the time spent in the patch *t* is exponentially distributed with probability density function $f(t) = \alpha e^{-\alpha t}$. (It is of course stupid to leave the patch too early when it still has a lot to consume; it is also stupid to stay too long when the patch is depleted. But a simple organism may be unable to judge time or resource level, and may leave at a constant rate.)

(a) Calculate the expected amount of resource eaten before leaving. (This is given by an integral that you can calculate explicitly.)

(b) Denote the expected amount of resources eaten (calculated above) by *A*. The long-term energy intake over many rounds of foraging in a patch and travelling to a new patch is *A* / (*S* + *T*), where *S* is the expected time spent in the patch and *T* is the

expected travel time to a new patch. *S* depends on the leaving rate α ; as we discussed in Part 1 of the course, $S = 1/\alpha$. *T* is independent of what the animal does within the patch and we assume it to be known.

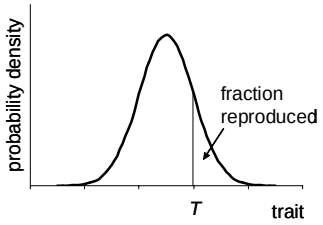
Find the optimal value of α , i.e., the value for which *A* / (*S* + *T*) is the greatest.

EXERCISES 27-31: NORMAL DISTRIBUTION

27. To be a pilot-astronaut, NASA requires that the candidate's height is between 163 cm and 193 cm. The height of people is normally distributed, the mean height of men in the US is 175 cm with variance 35 cm^2. What is the probability that a randomly chosen male US citizen meets NASA's height requirement?

28. A random variable ξ follows the standard normal distribution N(0,1). Find the number *u* such that ξ falls between -*u* and *u* with probability 95%.

29. *Truncation selection in animal breeding.* The distribution of body weight is normal. An animal breeder wants to select the heaviest animals for reproduction in a population where the mean weight is 10 units and the variance is 5 $unit^2$. This is commonly done by truncation selection: All animals above a threshold *T* of weight are bred, whereas those below *T* are not allowed to reproduce (see figure). The breeder, however, must reproduce a certain fraction of the population in order to maintain the number of animals. This fraction depends on fecundity (if one animal has a lot of offspring, then a few parents are enough to produce the next generation; otherwise more parents are needed to produce as many offspring as many animals the breeder had in the initial population).



Calculate *T* if the breeder has to select for reproduction

(a) 5% of the animals
(b) 60% of the animals

30. *Confidence interval of an average.* We measure a certain random variable ξ in a sample and calculate the sample average $\overline{\xi}$. Clearly, if we calculated the average in another sample, we would obtain a somewhat different result for $\overline{\xi}$, i.e., $\overline{\xi}$ itself is a random variable. The question is, how reliable the sample average $\overline{\xi}$ is as an estimate of the true mean.

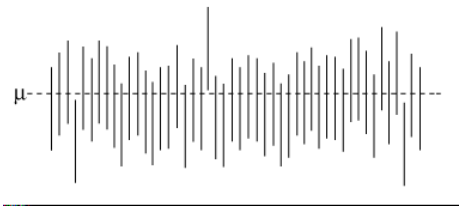
Denote the true mean with μ and the variance of ξ with V (here we assume that V is known). Let n denote the number of individuals in the sample. We assume that the sample is relatively large (n is not too small).

(a) Show that the transformed random variable $\frac{\bar{\xi} - \mu}{\sqrt{V/n}}$ follows the standard normal distribution N(0,1).

(b) Find the number u such that $P(-u \le z \le u) = 0.95$ (use the result of exercise 26 above).

(c) Find numbers c_1 and c_2 such that $P(c_1 \le \mu \le c_2) = 0.95$, if the sample average is $\bar{\xi} = 164$, the known variance of ξ is $V = 62$ and the size of the sample is $n = 30$. The interval $[c_1, c_2]$ is said to be the *confidence interval* for the unknown true mean μ, with confidence level 95%.

> There are two important points to mention here. First, can we say that "with 95% probability, the true mean is in the confidence interval"? It is a fact whether μ is in the calculated interval $[c_1, c_2]$ or not; only we do not know this fact. Strictly speaking, what we can say is this: if we repeated sampling and calculated $[c_1, c_2]$ from each sample, then 95% of these intervals will contain the true mean. This is illustrated with the figure below (from Wikipedia). Each sampling yields a confidence interval (vertical lines), which depend on the sample average and are thus generally different. Most of them cover the true mean, but an expected 5% does not.



> Second, in this example we assumed that the variance V is known, whereas in practice, we must estimate V also from the sample we have. The estimated variance is however a random variable, not a constant (just the same way as the sample average is a random variable). If we substitute the estimated variance for V in $\frac{\bar{\xi} - \mu}{\sqrt{V/n}}$, then it is not normally distributed (because it contains division by a random variable) but follows the so-called Student's t-distribution when ξ is normally distributed. The t-distribution is however very similar to the normal distribution if the sample size n is sufficiently large. For small samples, the calculation of the confidence interval goes similarly to this exercise but one has to use tables of the t-distribution (in part (b)) rather than the standard normal distribution.

31. *Hypothesis testing.* Suppose that in a large and well-known population, a normally distributed quantitative trait has mean $m = 143$ and variance $V = 441$. One culture derived from this population, however, appears to be different, because its average trait value is only $\bar{\xi} = 134$. This deviating average was calculated from measuring 25 individuals. Could the difference between the sample average $\bar{\xi}$ and the known population mean m be due to statistical fluctuations, or is there reason to suspect that something unusual happened to this culture?

Start with the *null hypothesis* that the 25 individuals represent a random sample from the population, and calculate the probability that the average trait value $\bar{\xi}$ of 25 randomly selected individuals differs from m by $|\bar{\xi} - m| = 143 - 134 = 9$ or more. If this probability is small (e.g. less than 5%), then we reject the null hypothesis and say that the difference between the measured average of the 25 individuals and the known mean trait value of the population is statistically significant, so that the culture is (probably) not just a random sample from the population.

(a) Argue that $\bar{\xi}$ is a normally distributed random variable. (Recall that the average is calculated as the sum of trait values divided with $n = 25$.)

(b) Determine the mean and the variance of $\bar{\xi}$ under the null hypothesis that the 25 individuals are a random sample from the large population.

(c) Calculate the probability $P(|\bar{\xi} - m|) \geq 9) = P(\bar{\xi} \leq 134) + P(\bar{\xi} \geq 152)$ under the null hypothesis and decide if we should reject the null hypothesis.

SOLUTIONS

1. The probability that a child is *aa* and therefore has the disorder is 1/4. One child is healthy with probability 1-1/4 = 3/4; the two children are independent and therefore both are healthy with probability 3/4 x 3/4 = 9/16.

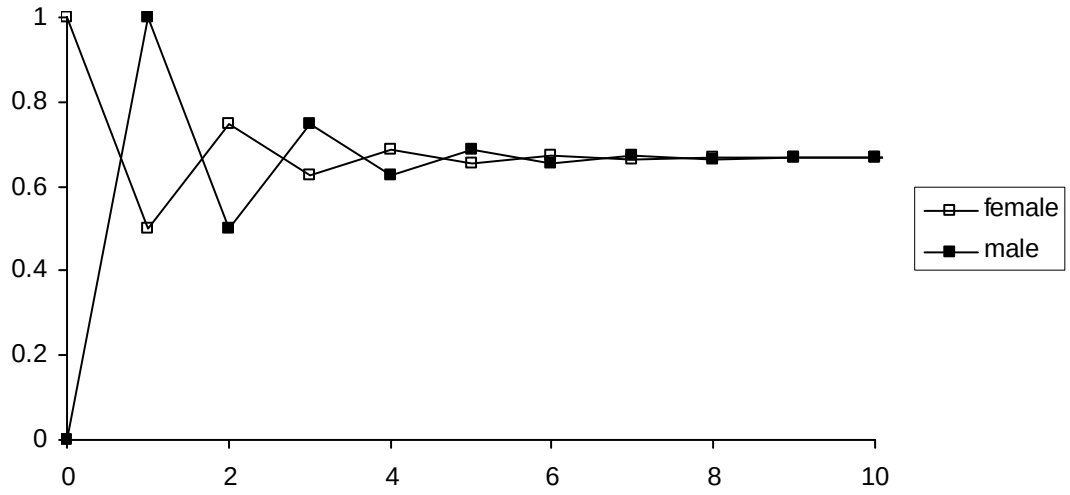
2. *a* is inherited from A to C with probability 1/2; if so, then it is inherited from C to E with probability 1/2; and if so, then G inherits it from E with probability 1/2. The probability that the paternally derived allele of G is *a* is therefore 1/8. The maternally derived allele of G is *a* also with probability 1/8, and the maternal line is independent of the paternal line. G is thus *aa* with probability 1/64, and healthy, concerning only the harmful allele *a*, with probability 63/64. The same calculation applies to any of the 6 harmful alleles in A and B, and the unlinked alleles are inherited independently. G is not a homozygote for any of the 6 harmful recessives with probability $(63/64)^6$ or 90.98%. Cousins are the closest relatives allowed to marry in most Western societies, but even these marriages carry a fairly high risk of conceiving a disordered child.

3. To have 7 loci on 7 different chromosomes, the first of the 7 loci can be on any chromosome; the next locus can be on any of the remaining 6 chromosomes, which has probability 6/7; the next locus can be on 5 chromosomes, which has probability 5/7; etc. The probability of having all 7 loci on different chromosomes is therefore

$$1 \cdot \frac{6}{7} \cdot \frac{5}{7} \cdot \frac{4}{7} \cdot \frac{3}{7} \cdot \frac{2}{7} \cdot \frac{1}{7} \approx 0.006$$

which is a very small probability. This simple argument started the rumour that Mendel might have cheated. More detailed analysis (taking into account that the chromosomes are not equally long, that loci on the same chromosome but far apart are inherited ca independently; and that Mendel did not perform every single cross) has however showed that the real probability of getting Mendel's data is not this low, and there is no ground of accusing Mendel of cheating.

4. Let p_f and p_m denote the frequency of the recessive "white" allele *w* in females and in males, respectively; at the beginning, $p_f = 1$ and $p_m = 0$. In the next generation, $p'_m = p_f$ because males inherit all their X chromosomes from females; and $p'_f = \frac{p_f + p_m}{2}$ because females inherit half their X chromosomes from females and half from males. The allele frequencies oscillate as shown below.



The overall frequency of allele *w* is $\frac{2}{3}p_f + \frac{1}{3}p_m$, because 2/3 of the X chromosomes are in females and 1/3 are in males. It is easy to see that $\frac{2}{3}p'_f + \frac{1}{3}p'_m = \frac{2}{3}p_f + \frac{1}{3}p_m$, i.e., this quantity remains the same in each generation and so remains the same as it was in the beginning, $\frac{2}{3}\cdot 1 + \frac{1}{3}\cdot 0 = \frac{2}{3}$. In equilibrium, $p'_m = p_m$ (frequencies do not change) and therefore $p_m = p_f$, i.e., females and males will have the same allele frequency. Because we still have that the overall allele frequency is $\frac{2}{3}$, both female and male allele frequencies must converge to this value. In equilibrium, therefore, the phenotypic frequencies are

white males: 2/3 of males
red males: 1/3 of males

white females: $\frac{2}{3}\cdot\frac{2}{3} = \frac{4}{9}$ of females

red females (all non-white): 5/9 of females

5. (a) The frequency of quadrats where plant A is present (either with B or alone) is P(A)=0.25+0.25=0.5; similarly, P(B)=0.25+0.05=0.3. If the two plants are present or absent independently of each other, then P(A and B) should equal P(A) times P(B). But this is not the case: the product of P(A) and P(B) is 0.15, whereas P(A and B)=0.25. Therefore, the plants occur together more often than they would if their spatial distribution is independent. (They might need the same type of environmental conditions, e.g. may both occur in wet sites; or may depend on each other as mutualists.)

(b) The same calculation shows that with these numbers, the plants occur independently of each other.

6. Denote the event of having high blood pressure with HB; African American with A and Caucasian with C. Then the data are

$$P(HB|A) = 0.33$$
$$P(HB|C) = 0.25$$
$$P(A) = 0.4,\ \ P(C) = 0.6$$

The total probability of HB is $P(HB|A)P(A) + P(HB|C)P(C) = 0.282$.

7. If the recipient is of blood group A, than it receives the wrong blood with probability P(B)+P(AB)=0.247. By the analogous calculation for each blood group, we obtain the conditional probabilities

$$P(\,fatal|A) = P(B) + P(AB) = 0.247$$
$$P(\,fatal|B) = P(A) + P(AB) = 0.523$$
$$P(\,fatal|AB) = 0$$
$$P(\,fatal|0) = P(A) + P(B) + P(AB) = 1 - P(0) = 0.688$$

The total probability of a fatal transfusion is

$$P(\,fatal) = P(\,fatal|A)P(A) + P(\,fatal|B)P(B) + P(\,fatal|AB)P(AB) + P(\,fatal|0)P(0),$$

which yields $P(\,fatal) = 0.4106$.

8. $\ln 0.01/\ln(1-q) \approx 4.6/q$

9. 160

10. $\mu = 5.98 \cdot 10^{-8}$

11. The probability of disease after the positive test is only 1.96%.

12. (a) prior probability: 0.15; posterior probability: 0.2609
(b) the posterior probability decreases when the prior decreases

13. (a)

$$P\left(q=0.5\middle|\ data\right)=0.60494$$

$$P\left(q=0\middle|\ data\right)=0$$

$$P\left(q=1\middle|\ data\right)=0.39506$$

(b)

$$P\left(q=0.5\middle|\ data\right)=0.015385$$

$$P\left(q=0\middle|\ data\right)=0$$

$$P\left(q=1\middle|\ data\right)=0.984615$$

(c) If we exclude a possibility in the prior, the corresponding posterior probability will be zero. (For example, if we use P(tail coin)=0 then $P\left(q=1\middle|\ data\right)=0$ even if there are many tails in the data.)

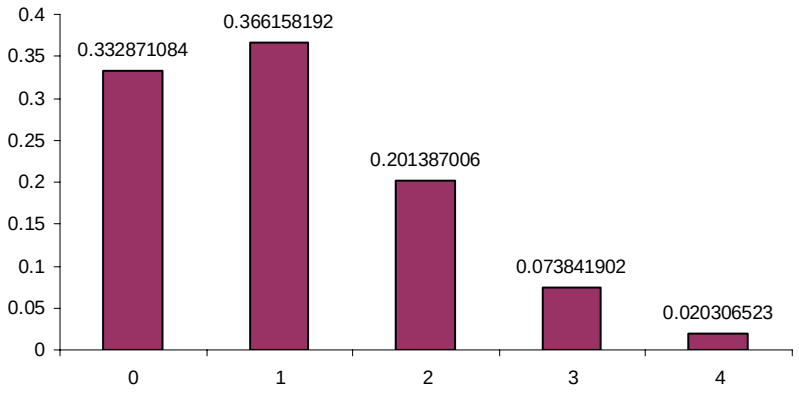
If we assume P(fair coin)=1 and hence P(tail coin)=P(head coin)=0, then all posterior probabilities other than the fair coin's probability will be zero.

14. (a)

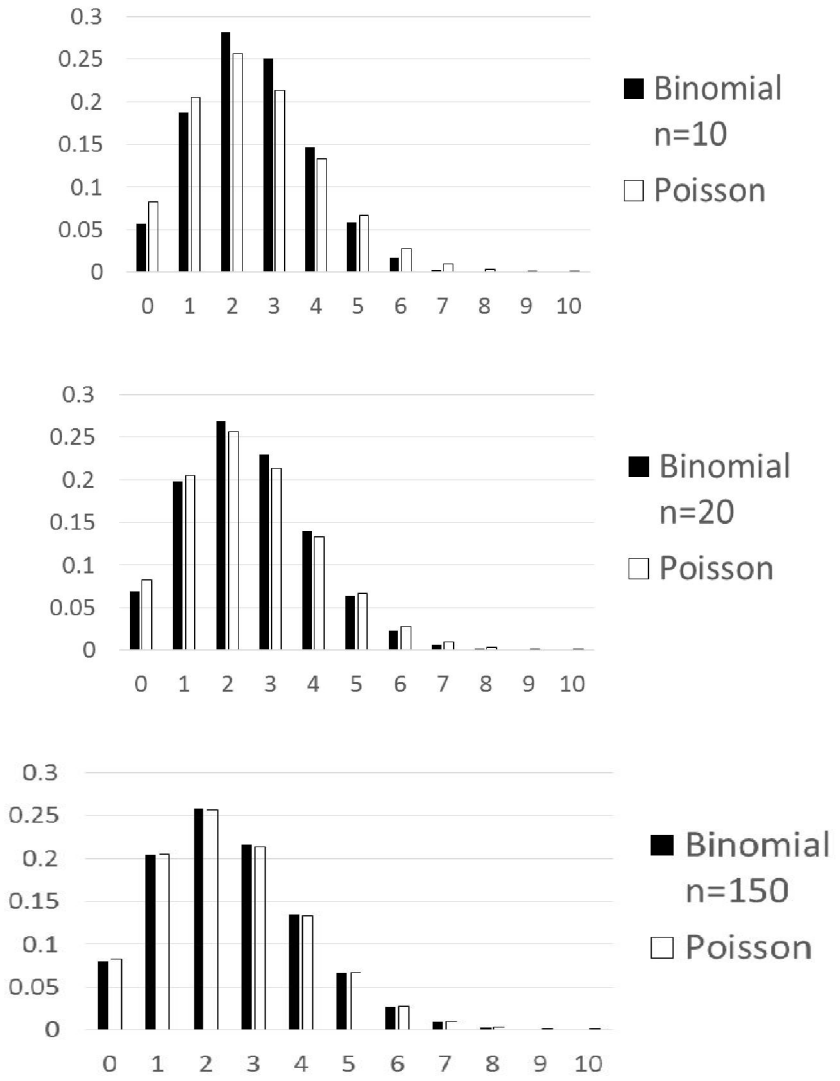
# alive	t=1	t=3	t=10
0	0.00243	0.12241	0.86652
1	0.02835	0.31954	0.12594
2	0.1323	0.33365	0.00732
3	0.3087	0.17419	0.00021
4	0.36015	0.04547	$3.09\cdot10^{-6}$
5	0.16807	0.00475	$1.8\cdot10^{-8}$

(b) 13 years

15.



16. The black columns of the following three charts show the histogram of the binomial distribution for n = 10, 20 and 150, and p such that np = 2.5. The white columns are the histogram of the Poisson distribution for $\lambda = 2.5$ for comparison. For large n, the two distributions are very similar.

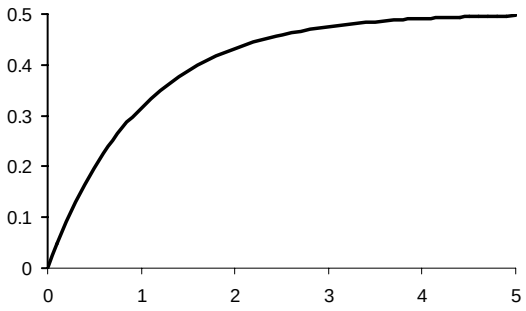


17. (a) $\hat{x} = \dfrac{\beta}{\alpha + \beta}$; with $\beta = 2\alpha$, this yields $\hat{x} = \dfrac{2\alpha}{\alpha + 2\alpha} = \dfrac{2}{3}$. Notice that the equilibrium does not depend on the absolute speed of opening and closing (i.e., on the values of β and α), only on their ratio. When opening is twice as fast as closing, the number of open channels is twice as high as the number of closed channels (2/3 vs 1/3) so that the number of all closing events is the same as the number of all opening events ($(2/3)\alpha = (1/3)\beta$).

(b) The channels are independent of each other and each channel is open with probability $\hat{x} = 2/3$. The number of open channels is therefore binomially

distributed. $P(k=12)=\binom{18}{12}\left(\frac{2}{3}\right)^{12}\left(\frac{1}{3}\right)^{6}=0.19627$

18. (a) $RF=\frac{1}{2}(1-e^{-\lambda})$

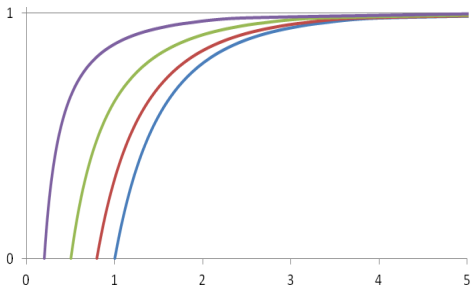


(c) $\lambda=-\ln(1-2RF)$ and therefore $RF\approx(\lambda/2)\cdot 100\%$ as long as RF is small. Hence for short distances, $(\lambda/2)\cdot 100$ gives the distance in traditional map units.

19. (a) $x_{t+1}=px_t+(1-px_t)(1-\exp(-\alpha x_t))$.

(b) At equilibrium, $\alpha=-\frac{1}{\hat{x}}\ln\left(\frac{1-\hat{x}}{1-p\hat{x}}\right)$

The chart below shows $\hat{x}$ as a function of α for p = 0.8, 0.5, 0.2, 0 (from left to right). The rightmost curve (p = 0) corresponds to the original Skellam model for annual plants. The positive equilibrium exists when the population is viable, i.e., for annual plants, when $\alpha>1$.



20. (a)

$$x_{t+1}=1-\exp(-\alpha x_t)$$
$$y_{t+1}=\exp(-\alpha x_t)\left(1-\exp(-\beta y_t)\right)$$

(b) The equilibrium equations

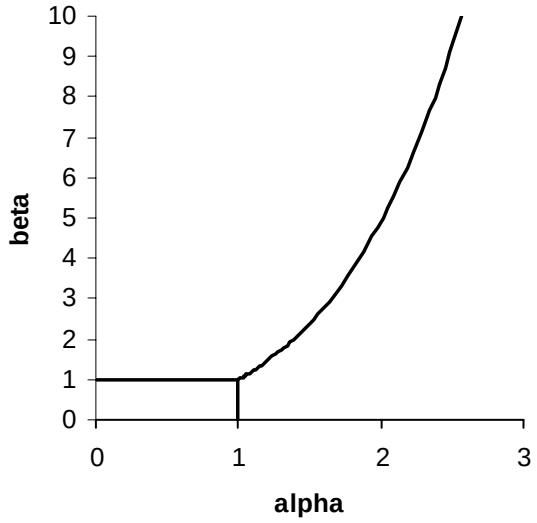
$$x = 1 - \exp(-\alpha x)$$
$$y = \exp(-\alpha x)(1 - \exp(-\beta y))$$

cannot be solved for *x* and *y* directly. Solving for the parameters (α, β) yields

$$\alpha = -\frac{\ln(1-x)}{x}$$
$$\beta = -\frac{1}{y}\ln\left(1 - \frac{y}{1-x}\right)$$

One can vary *x* and *y* between 0 and 1 and calculate the corresponding range of (α, β) numerically.

In particular, if *y* is close to zero, then $\ln\left(1 - \frac{y}{1-x}\right) \approx -\frac{y}{1-x}$ and $\beta \approx \frac{1}{1-x}$ so that $\beta > \frac{1}{1-x}$ is necessary for the inferior species to be viable. Vary *x* between 0 and 1 and calculate the corresponding pairs of parameter values $\alpha = -\ln(1-x)/x$ and $\beta = 1/(1-x)$. Plot the resulting β's against the α's to arrive at the curve in the figure below.



$\alpha < 1$ implies that the superior species is not viable, and therefore the inferior species is viable if and only if $\beta > 1$. When $\alpha > 1$, the superior species is present and the inferior species is viable when it's fecundity is sufficiently large to compensate for the loss of seeds to unsuccessful competition with the superior species; i.e., the inferior species is present and the two species coexist when $\alpha > 1$ and β is above the curve.

21. (a) The expected number of attacks on a given host is proportional to the parasitoid density, i.e., $\lambda = \alpha P_t$. The number of attacks is Poisson distributed, and therefore the probability of avoiding attack is $e^{-\alpha P_t}$.

(b)

$$H_{t+1} = H_t \exp(-\alpha P_t) F$$
$$P_{t+1} = H_t (1 - \exp(-\alpha P_t)) B$$

22. $E(\eta) = 3.5$, $V(\eta) = 0.29167$

23. (a) $COV(\eta_1, \eta_2) = V(\xi)$, $r = \dfrac{V(\xi)}{V(\eta)}$

(b) The correlation coefficient can be measured from the twin data. From (a), this equals the heritability.

A problem with this measurement is that twins share more than their genes; part of the environmental effects (important childhood effects, for example) are also the same, which means that we underestimate the variance of environmental effects. This can be corrected if we consider also fraternal twins, who share the environment ca to the same extent as identical twins but share less of their genes.

24. (b) As *n* increases,

$V(\xi) = np(1-p)$ increases

$D(\xi) = \sqrt{np(1-p)}$ increases

$c = D(\xi)/E(\xi) = \sqrt{\dfrac{1-p}{np}}$ decreases (*c* is a measure of variability *relative* to "typical" values, such as the mean; with increasing *n*, variation indeed decreases *on the scale of typical values*)

$V(\xi/n) = (1/n)^2 V(\xi) = p(1-p)/n$ decreases (with large *n*, we are getting the *frequency* more precisely)

25. (a) In equilibrium, $\beta y = \alpha z x$ and therefore the fraction of free enzymes is $\dfrac{x}{x+y} = \dfrac{x}{x + (\alpha z/\beta)x} = \dfrac{\beta}{\beta + \alpha z}$. The enzyme molecules are independent of each other (as long as the number of inhibitor molecules is much higher than the number of enzyme molecules). Therefore ξ is binomial with parameters *N* and $p = \dfrac{\beta}{\beta + \alpha z}$ (note that *z* is considered constant).

$$E(\xi) = Np = N\frac{\beta}{\beta+\alpha z};\ E(N-\xi) = N(1-p) = N\frac{\alpha z}{\beta+\alpha z}$$

$$V(\xi) = V(N-\xi) = Np(1-p) = N\frac{\beta\alpha z}{(\beta+\alpha z)^2}$$

(b) N should be at least $\frac{10000\alpha z}{\beta}$ to have the coefficient of variation of ξ to be at most 0.01; and N should be at least $\frac{10000\beta}{\alpha z}$ to have the same for $N-\xi$. Hence $N \geq 10000\max\left(\frac{\alpha z}{\beta}, \frac{\beta}{\alpha z}\right)$.

26. (a) $A = \int_0^\infty g(t)\,f(t)dt = \frac{\beta}{\alpha+\beta}$

(b) $\alpha = \sqrt{\frac{\beta}{T}}$

27. Let η denote the transformed random variable $\eta = \frac{\xi - 175}{\sqrt{35}}$. η follows the standard normal distribution, so that we can use the table of the standard normal distribution to obtain the probability that η is less than a given value.

$$P(\xi < 193) = P\left(\eta < \frac{193-175}{\sqrt{35}}\right) = P(\eta < 3.04) = 0.9988$$

$$P(\xi < 163) = P\left(\eta < \frac{163-175}{\sqrt{35}}\right) = P(\eta < -2.03)$$

$P(\eta < -2.03)$ is not listed in the table. Because the normal distribution is symmetric, $P(\eta < -z) = P(\eta > z) = 1 - P(\eta < z)$. Look up $P(\eta < 2.03) = 0.9788$ from the table, and obtain $P(\eta < -2.03)$ as $P(\eta < -2.03) = 1 - 0.9788 = 0.0212$

$$P(163 < \xi < 193) = P(\xi < 193) - P(\xi < 163) = 0.9988 - 0.0212 = 0.9776$$

28. 1.96

29. (a) T=13.6895, (b) T=9.4186 (your result may be somewhat different due to different precision of the calculation)

30. (c) $c_1 = \bar{\xi} - 1.96\sqrt{V/n} = 161.18$, $c_2 = \bar{\xi} + 1.96\sqrt{V/n} = 166.82$

31. (a) The sum of 25 independent, identically distributed random variables is approximately normal.
(b) $E(\bar{\xi}) = m = 143$;

$$V(\bar{\xi}) = V\left(\frac{\xi_1 + \ldots + \xi_n}{n}\right) = \frac{1}{n^2} \cdot n \cdot V(\xi) = \frac{V(\xi)}{n} = \frac{441}{25} = 17.64$$

(c) If $\bar{\xi}$ is indeed normally distributed with mean 143 and variance 17.64, then $\eta = \dfrac{\bar{\xi} - 143}{\sqrt{17.64}}$ follows the standard normal distribution.

$$P(\bar{\xi} < 134) = P\left(\eta < \frac{134 - 143}{\sqrt{17.64}}\right) = P(\eta < -2.14) = 1 - P(\eta < 2.14) = 0.0162$$

$P(\bar{\xi} < 134) + P(\bar{\xi} > 152) = 2 \cdot 0.0162 = 0.0324$, and because this is less than 5%, we reject the null hypothesis.