Week 1 - HIV Dynamics

Course goals:

- Tools to understand biological processes.
- Draw conclusions from data
- Learn about control, noise, stochastic processes. Epidemics, cells, ecosystems, genes, measurements.

Homework: 40%, Final: 60%
Checked partially
Book - PMLS Nelson on website
Level - includes very basic level, the trick is in the synergy.

Why models?

- Organize data
- Check hypotheses quantitatively
- (Predict) - ML is better at “just” prediction.
- Science as an intelligence agency

HIV Dynamics

Los Alamos (1994) Alan Perelson

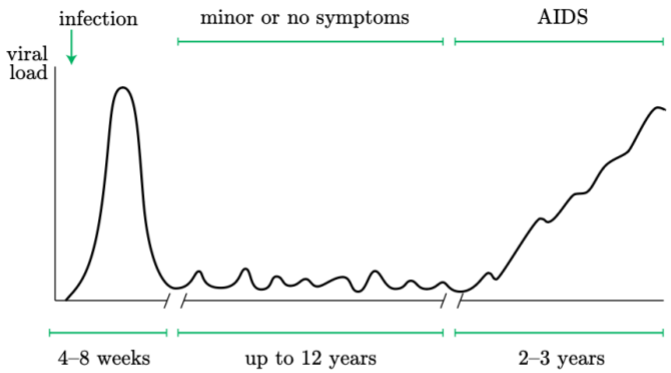
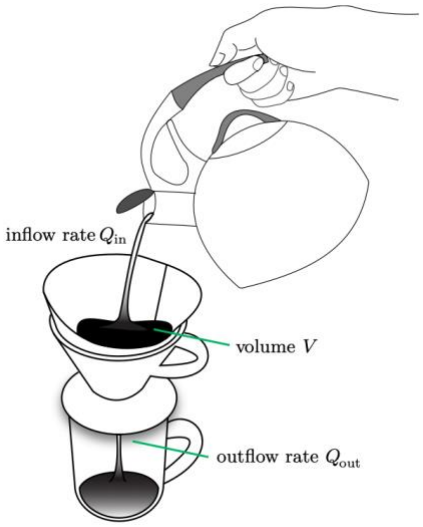


Figure 0.1: [Sketch graph.] **The time course of HIV infection,** representing the progression of the disease as it was understood in the early 1990s. After a brief, sharp peak, the concentration of virus particles in the blood (“viral load”) settled down to a low, nearly steady level for up to ten years. During this period, the patient showed no symptoms. Ultimately, however, the viral load increased and the symptoms of full AIDS appeared. [After Weiss, 1993.]

What's happening in the "latent" stage? Is the virus replicating very slowly?

Physical analogy

A leaky container that retains water at some level at steady state.



At steady state:

- The rate of flow out must equal the rate of flow in.
- The rate can be high or low, for as long as in and out are matched
- How fast the flow in = how fast is the virus replicating in the "latent" stage?

Hypothesis: The virus multiplies rapidly but after the initial episode the body clears it just as quickly. The steady state is possible for any rate in for as long as the rate out scales with the density.

- At higher volume: flow out more quickly because mgz+PV = const.
- "Death must win" the death rate is always a higher power than the generation rate for steady state. Baseline generation (no number dependence) means removal proportional to number. This way the blood cell population level stays constant in one's life despite producing 500 billion per day.
- Back to the virus: different people will have different state states, but for each person there will be a stable time window.

Question - how to make progress? Too many parameters, not enough data.

Ritonavir - Protease inhibitor

NYC 1994: David Ho, clinical trials. Works amazingly, after a few months, stops working. Does not kill the virus but stops the creation of new viruses.

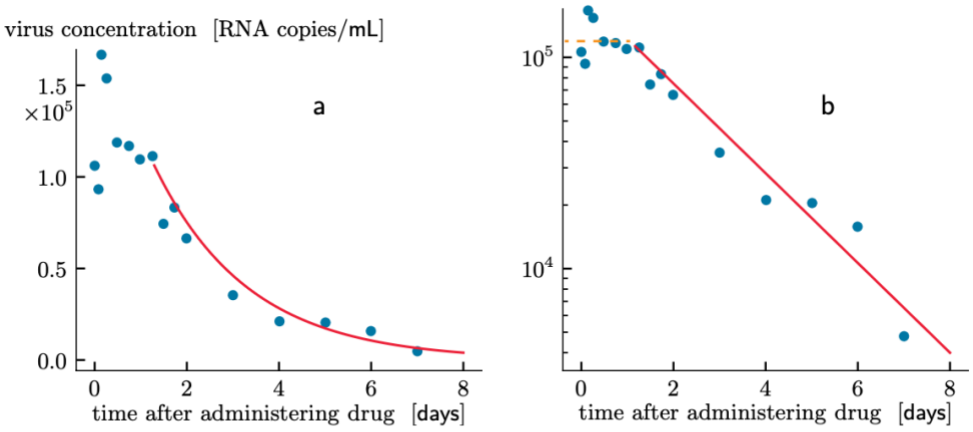


Figure 0.3: [Experimental data with preliminary fit.] **Virus concentration in one patient's blood ("viral load") after treatment with a protease inhibitor**, showing the rapid decline after treatment. (a) The *solid line* shows the time course corresponding to elimination of half the total viral population every 1.4 days. (b) In this semilog plot, the *dashed line* highlights a deviation from exponential behavior at early times (the "initial plateau"); see Chapter 1. [Data from Perelson, 2002, available in Dataset 1.]

Qout is fast!
(100 fold less in ~ 10 days or about 10 fold in 5-6 days.)
Quot large - this means that Qin the virus replicating like mad! (10^9 virions per day).

Implications:

- Fast replication - many mutations created → drug resistence
- In fact, we will learn, the drug resistance is already there and will simply take over
- The high replication rate is also why it can escape the immune system.

Hypothesis: if the probability of a mutation is p << 1, for two drugs you need p^2, for three drugs p^3 etc. Therefore the advantage of a cocktail!

What led to this discovery?

- Interdiscinary question / team.
- Simple physical idea (leaky container)
- Physics tools (differential equations modeling)
- New data, quantitatively analyzed, supports or refutes hypotheses
- Embodys the model (math) and attempts to fit the experimental data.

What does it mean to fit the data?

- To adjust the model parameters to best reflect the data. Words to watch: “Parametric model”, “Inverse problem”, “Infer / learn”.
- Is the fit good? If it is, the model is “promising”. Promising what: to answer a question. Here:
  Question: Why did the first antiviral drugs work for a while, then stop?
  Tools: simple dynamical model + experimental data to fit

What can we learn from the graph?

- There is actually a short plateau before falling off
- Not a perfect agreement with exponential decay. But how do we define agreement?

Modeling the dynamics

Let’s idealize the system by assuming that the antiviral drug completely stops new infections of T cells.We also simplify by assuming that each infected T cell has a fixed chance of being cleared in any short time interval.

Relevant processes:

- Infected T-cells produce free virions
- Virions infect new cells
- Infected T-cells die, and are also killed by the immune system = clearance of infected cells
- Immune system clears free virions
- $t<0$ before drug, quasi steady state, rate of new T-cell infections = T-cell clearance rate
- $t>0$ ideal drugs, completely stops new infections. Therefore, for $t>0$, the number of uninfected cells becomes irrelevant.
- Mutations happen mostly from reverse transcriptase therefore just once when a cell gets infected.
- Many mutations per unit time → Many infections per unit time
- Not the same as the new virions created
- With the drug, the uninfected cells become decoupled from the infected or from the virions since the virus cannot reproduce.

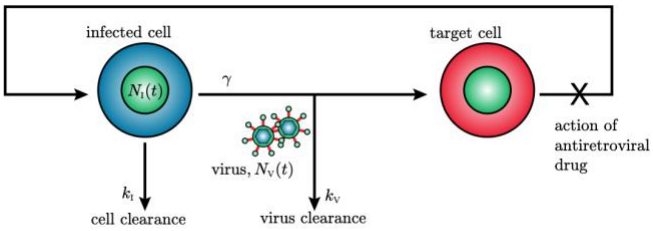


Figure 1.2: [Schematic.] **Simplified virus life cycle.** In this model, antiviral drug therapy works by halting new infections of T cells (*cross*). The constants k_I, k_V, γ introduced in the text are shown at the points of action of the associated processes.

- For short Δt the probability to get cleared is K*Δt for each of the N_I cells.

$$\frac{dN_I}{dt} = -k_I N_I \qquad \text{for } t \geq 0. \qquad \text{physical model infected cell count} \qquad (1.1)$$

Virions are produced at a rate ɣ and cleared at k_V

$$\frac{dN_V}{dt} = -k_V N_V + \gamma N_I. \qquad \text{physical model virus in patient} \qquad (1.2)$$

Hypotheses to test

- The virus evolves within a single patient
- Mutations most likely at the reverse-transcription step, which happens once per infection.
- Find the rate T-cells get infected in the quasi steady state.
- Data: $N_v(t)$, not the number of infected cells (at the time).

For reference, the following named quantities appear in our analysis:

t	time since administering drug
$N_I(t)$	population of infected T cells; its initial value is N_{I0}
$N_V(t)$	population of virions; its initial value is N_{V0}
k_I	clearance rate constant for infected T cells
k_V	clearance rate constant for virions
γ	rate constant for virion production per infected T cell
β	an abbreviation for γN_{I0}

$$N_I(t) = N_{I0} e^{-k_I t}.$$

Solving for N_I we have

Letting $\beta = \gamma N_{I0}$

In

$$\frac{dN_V}{dt} = -k_V N_V + \gamma N_{I0} e^{-k_I t}. \qquad (1.3)$$

For a real leaky container, the rate of outflow depends on the pressure at the bottom, and hence on the level of the water; similarly, Equation 1.2 specifies that the clearance (outflow) rate at time t depends on $N_V(t)$.

If $k_I \gg k_V$, then the inflow quickly shuts off, before much has had a chance to run out. After this **brief transient behavior,** N_v should therefore fall exponentially with time, in a way controlled by the decay rate constant k_v .

In the opposite extreme case, $k_V \gg k_I$, the water never drains completely, because in our model the rate of outflow goes to zero as the height goes to zero; instead, the water level simply tracks the rate of inflow. Thus, again the water level falls exponentially, but this time in a way controlled by the inflow decay rate constant k^I.

dNv/dt = -k_V Nv + c = 0 → Nv = C/k_V

With C slowly decaying controlled by k_I
Trial solution:

$$N_V(t) \stackrel{?}{=} X e^{-k_I t} + (N_{V0} - X) e^{-k_V t}, \qquad (1.4)$$

Works for

$$X = \beta/(k_V - k_I).$$

How can we have an initial plateau even with two decaying exponentials? We can have one of the terms X or the other negative.

We now need to extract parameters from the data.
Two integration constants: N_{I0} and N_{V0} - measured / known.
3 parameters unknown: kv, kI, beta

The model itself needs to be evaluated critically; all of the assumptions and approximations that went into it are, in principle, suspect.

The fits below are bad. Why are they bad?

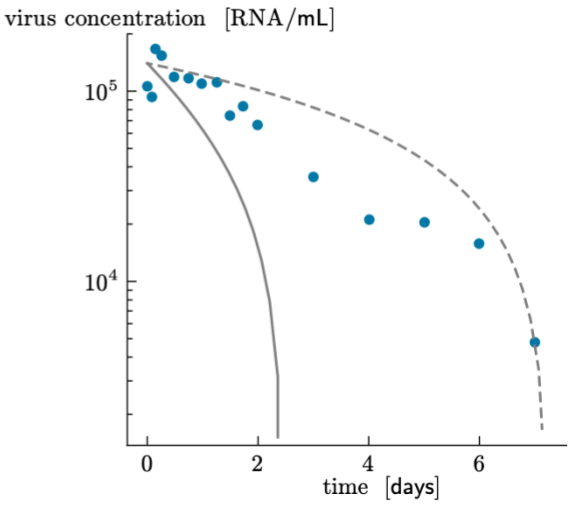


Figure 1.3: [Experimental data.] **Bad fits.** *Dots* on this semilog plot are the same experimental data that appeared in Figure 0.3 (page 3). The *solid curve* shows the trial solution to our model (Equation 1.4), with a bad set of parameter values. Although the solution starts out at the observed value, it quickly deviates from the data. However, a different choice of parameters does lead to a successful fit (see Problem 1.4). The *dashed curve* shows a fit to a different family (linear functions), not grounded in any physical mechanism. The curve starts and ends at the right values, but *no* choice of parameters for this unphysical model can fit all the data.

Overconstrained: More constraints than parameters: more than 3 data points
Overfit: More params then constraints - you can fit anything, you “memorize” the points.
Success: the theory is plausible. Even if there are hidden actors like the infected cells.
Blind fitting: If the data suggests a trend and we follow it without a theory. Good at summarizing the data, bad at extrapolating or explaining. Here we didn’t blind fit, we make a physical model.

Exit from latency

What happens there? The system gets exhausted, mutations increase, immunity breaks down.

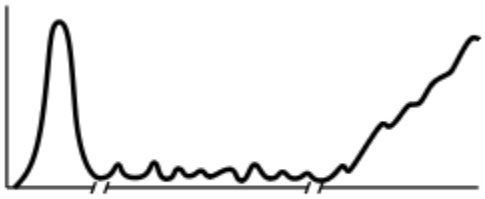


Fig. 0.1, p. 1

Informal criterion for falsification

The number of “features” in the data. 4 features, one already used in the solution. Therefore 4 features for 3 params

Nv(0), slope at plateau, sharpness of transition, slope after plateau, intercept of slope

There is no guarantee that any parameter combination will fit well.

Data is inconsistent with $1/k_I$ = 10 years - therefore virus not slow.

More realistic dynamics with drug (Perelson 2002)

$$\frac{dN_U}{dt} = \lambda - k_V N_U - \epsilon N_V N_U, \quad (1.5)$$

$$\frac{dN_I}{dt} = \epsilon N_V N_U - k_I N_I, \quad (1.6)$$

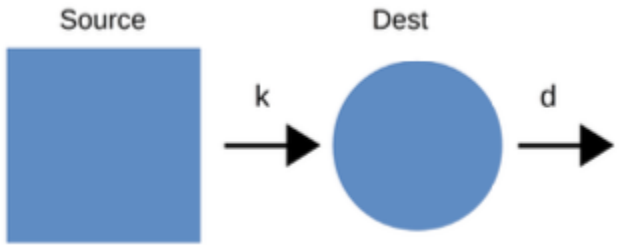
$$\frac{dN_V}{dt} = \epsilon' \gamma N_I - k_V N_V, \quad (1.7)$$

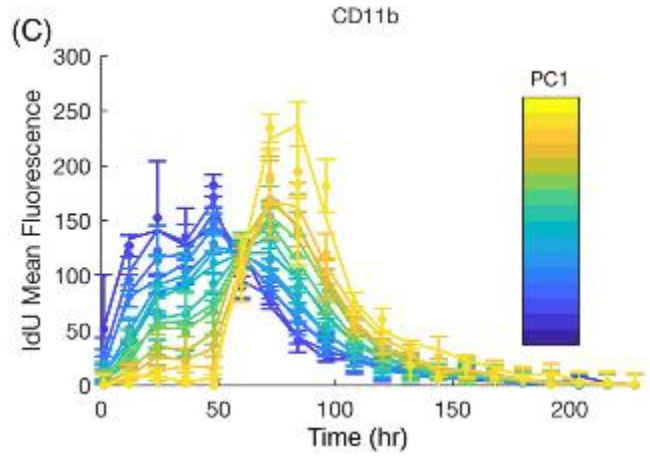
$$\frac{dN_X}{dt} = (1 - \epsilon')\gamma N_I - k_V N_X. \quad (1.8)$$

Nu - uninfected cell, Nx - inactive virions
ε - fraction remains infective
ε' = fraction of competent virions produced (can produce more virions)
Latently infected cells means clearing the infection is very difficult.

Aside: pulse-chase proliferation / differentiation

“Quantifying the Dynamics of Hematopoiesis by In Vivo IdU Pulse-Chase, Mass Cytometry, and Mathematical Modeling” Cytometry part A (2019)





Week 2 - Randomness

If each attempt at catching prey is an independent event, how many attempts are needed for a predator to succeed?

Physical idea: waiting distribution for the next event.

- Bernoulli trial - *s* - Succeeds with probability ξ.
- Two trials: $x = s_1 * ½ + s_2 * ¼$
  We get after many pairs of trials: (probability ½ “fair coin”)
  Similarly for 3 bits. A uniform distribution between 0 and 1. Similarly drawing from 0..9 for 1/10 + 1/100 + …

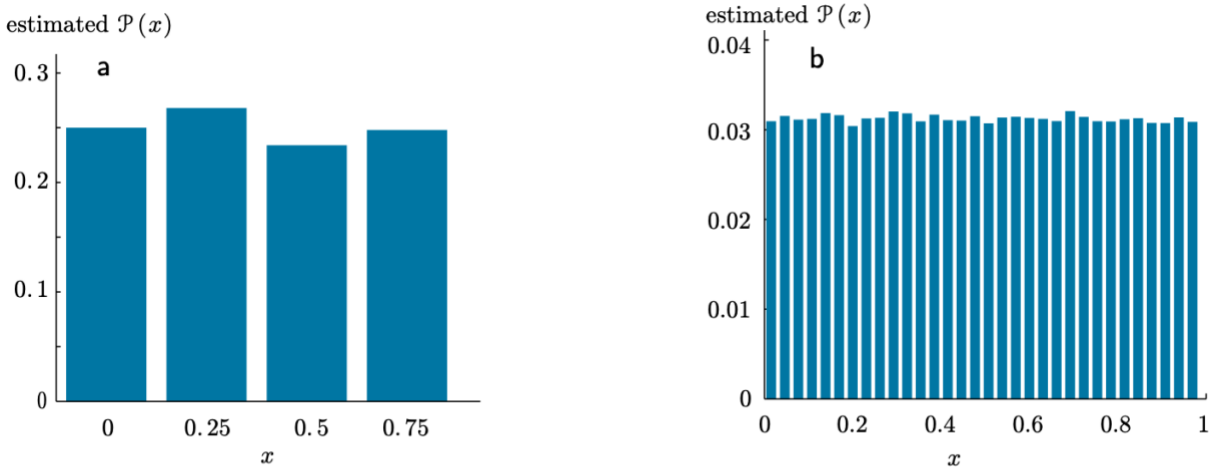


Figure 3.1: [Computer simulations.] **Uniformly distributed random variables.** Empirical distributions of (a) 500 two-bit random binary fractions (that is, $m = 2$ in Equation 3.1) and (b) 250 000 five-bit binary fractions ($m = 5$). The symbol $\mathcal{P}(x)$ refers to the probabilities of various outcomes; it will be defined precisely later in this chapter. Notice that the bar height in (b) is much lower than in (a), because 100% total probability has been subdivided into many more bins.

- Diploid organisms: two copies from every gene: one from male one from female. Together they form a germ cell. Inheritance: ½ to get a copy from grandma. (More complicated: transpositions, duplications, mutations…)

Probability mass function

Assume that the experiments are replicable.

$$\mathcal{P}(\ell) = \lim_{N_{\text{tot}} \to \infty} N_\ell / N_{\text{tot}}. \qquad (3.3)$$

- Note that $P(\ell)$ is always nonnegative
- Any discrete probability distribution function is dimensionless

$$\sum_\ell \mathcal{P}(\ell) = 1. \quad \text{normalization condition, discrete case} \qquad (3.4)$$

Addition of mutually exclusive events E_1, E_2
$P(E_1 \text{ or } E_2) = P(E_1)+P(E_2)$ [mutually exclusive]
When there is overlap: $P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2)$
P(not E) = 1-P(E)

Conditional probability

- A friend offers a bet on a die. You win if the die=5. How risky is the bet? Then someone tells you the die is an odd number. Now how risky? Somebody tells you that a coin flip is up - now how risky?

  P(5) = ⅙ < P(5 | roll is odd) = ⅓

  $$\mathcal{P}(\mathsf{E} \mid \mathsf{E}') = \lim_{N_{\text{tot}} \to \infty} \frac{N(\mathsf{E} \textbf{ and } \mathsf{E}')}{N(\mathsf{E}')}.$$

  Can can therefore,

  $$\mathcal{P}(\mathsf{E} \mid \mathsf{E}') = \lim_{N_{\text{tot}} \to \infty} \frac{N(\mathsf{E} \textbf{ and } \mathsf{E}')/N_{\text{tot}}}{N(\mathsf{E}')/N_{\text{tot}}}, \text{ or} \qquad (3.9)$$

  $$\mathcal{P}(\mathsf{E} \mid \mathsf{E}') = \frac{\mathcal{P}(\mathsf{E} \textbf{ and } \mathsf{E}')}{\mathcal{P}(\mathsf{E}')}. \quad \textbf{conditional probability} \qquad (3.10)$$

Similarly,

$\mathcal{P}(\mathsf{E}\ \textbf{and}\ \mathsf{E}') = \mathcal{P}(\mathsf{E} \mid \mathsf{E}') \times \mathcal{P}(\mathsf{E}').$ general **product rule** (3.11)

$\mathcal{P}(\mathsf{E}\ \textbf{and}\ \mathsf{E}') = \mathcal{P}(\mathsf{E}) \times \mathcal{P}(\mathsf{E}').$ statistically independent events (3.12)

The Geometric distribution describes the waiting until success in a series of independent trials

$\mathcal{P}_{\text{geom}}(j;\xi) = \xi(1-\xi)^{j-1}$, for $j = 1, 2, \ldots.$ **Geometric distribution** (3.13)

Normalized because

$$\sum_{j=1}^{\infty} P(j) = \xi \sum_{j=1}^{\infty} (1-\xi)^{j-1} = \xi \frac{1}{1-(1-\xi)} = 1$$

Joint distributions

$P_{XY}(X = x, Y = y)$

Marginal:

$P_X(X = x) = \sum_y P_{XY}(X = x, Y = y)$

Normalization:

$$\sum_x \sum_y P_{XY}(X = x, Y = y) = 1$$

Shannon entropy (in bits)

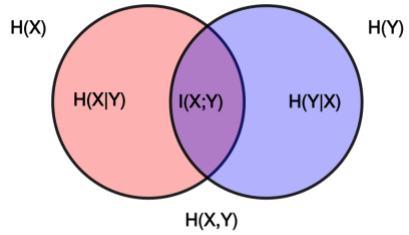
$$S_x = -\sum_x P_X(x) \log_2 P_X(x)$$

Mutual information

$$S_{XY} = -\sum_{x,y} P_{XY}(x,y) \log_2 P_{XY}(x,y)$$
$$I(X;Y) = S_X + S_Y - S_{XY}$$

$$I = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x) P_Y(y)}$$

The part of entropy of X explained by Y (or vice versa)



The Kullback-Leibler Divergence

Is a measure of how one probability distribution P is different from a second, reference probability distribution Q.

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

- Statistics: the expected log-likelihood ratio (expected under P)
- Coding: the expected number of extra bits required to code samples from P using a code optimized for Q rather than the code optimized for P
- Machine learning: the information gain achieved if P would be used instead of Q which is currently used. By analogy with information theory, it is called the relative entropy of P with respect to Q.
- Bayesian: a measure of the information gained by revising one's beliefs from the prior probability distribution Q to the posterior probability distribution P. In other words, it is the amount of information lost when Q is used to approximate P.

The mutual information is the KL divergence between Pxy and PxPy: how different is the joint distribution from the marginal, how much more information in Pxy vs. PxPy?

$$\mathrm{I}(X;Y) = D_{\mathrm{KL}}(P(X,Y) \parallel P(X)P(Y))$$

$$I = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$$

Medical tests - Prior knowledge

Imagine the following situation:

- Random screening, not feeling sick
- Test is positive. Doctor says test is 97% “accurate”.

Am I sick? For a yes/no question we have: P(sick | positive)

Sensitivity: Truly sick → Test positive = TP, small false negative

97% Sensitive → 3% false negative

P(sick, +) + P(sick, -) = P(sick)

$$\text{sensitivity} = \mathcal{P}(P \mid S) = \frac{\mathcal{P}(S \textbf{ and } P)}{\mathcal{P}(S)} = \frac{\mathcal{P}_{SP}}{\mathcal{P}_{SN} + \mathcal{P}_{SP}} = 97\%.$$

Selectivity: Healthy → Test negative = TN. Large TN = Small FP

TN + FP = P(healthy,-) + P(healthy,+) = P(healthy)
Assume the test has also 97% Selectivity, meaning 3% Positive and healthy (FP).

Am I sick?

$$\mathcal{P}(S \mid P) = \frac{\mathcal{P}(S \textbf{ and } P)}{\mathcal{P}(P)} = \frac{\mathcal{P}_{SP}}{\mathcal{P}_{SP} + \mathcal{P}_{HP}}.$$

I need to know P(Sick) = 0.9 % this is the prior. Then

P(Sick | Pos) = P(Pos | Sick) P(Sick) / P(pos)

But P(pos) = P(pos,sick) + P(pos, healthy) = P(pos|sick)P(sick) + P(pos|healthy)P(healthy) =
= 0.97*0.009 + 0.03*0.991 = 0.03846
Therefore, P(Sick | Pos) = 0.97*0.009 / 0.03846 = **0.227**
Therefore, even though the test is “97% accurate”, you have only ¼ probability of being sick!
Another way to think about it, is that before the test you had 0.9% Probability of being sick, and it increased 25-fold !

Bayes Theorem

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)},$$

Posterior of Hypothesis given Evidence = Likelihood of Hypothesis given a fixed Evidence * Prior Hypothesis / Marginal Evidence

- *H* stands for any *hypothesis* whose probability may be affected by data (called *evidence* below). Often there are competing hypotheses, and the task is to determine which is the most probable
- *P(H) the prior probability, is the estimate of the probability of the hypothesis H before the current evidence is observed.*
- *E, the evidence, corresponds to new data that were not used in computing the prior probability.*
- *P(H | E) = the posterior probability, is the probability of H after E is observed.*
- *P(E | H) = The likelihood - the probability of the evidence given the hypothesis. The likelihood is a function of the evidence, whereas the posterior is a function of the hypothesis.*
- *P(E) is sometimes termed the marginal likelihood or "model evidence". This factor is the same for all possible hypotheses being considered (as is evident from the fact that the hypothesis H does not appear anywhere in the symbol, unlike for all the other factors) and hence does not factor into determining the relative probabilities of different hypotheses.*

Expectation and other moments

$$\langle f^n \rangle = \sum_s f^n(s) P(s)$$

Var f = $< (f-<f>)^2 > = <f^2> - <f>^2$

Bernoulli: $<f> = 0*(1-\xi)+ 1*\xi = \xi$
$<f^2> = \xi$
Therefore, var(f) = $\xi - \xi^2 = \xi(1-\xi)$ which is maximized at $\xi=0.5$

For f,g independent: <fg> = <f><g> because

$$\sum_{ij} f(i)g(j)P_{ij} = \left(\sum_i f(i)P_i\right)\left(\sum_j (j)P_j\right)$$

When they are dependent this is no longer true.

For two independent f,g:

< f+g > = <f> + <g>

var(f+g) = < $(f+g)^2$> - (<f>+<g>)^2 = var(f) + var(g)
var(f-g) = < $(f-g)^2$> - (<f>-<g>)^2 = var(f) + var(g) —- Same as var(f+g) therefore

Coefficient of variation (CV) = Relative standard deviation (RSD) = root(var(f)) / |<f>|
"Mahalanobis distance"
Dimensionless

RSD(f+g) / RSD(f-g) = root(var(f) + var(g)) / |<f>+<g>| * |<f>-<g>| / root(var(f) + var(g)) =

|<f>-<g>| / |<f>+<g>| < 1 for non-negative random vars!

The difference between two noisy positive variables is a very noisy variable, noisier than their sum. Hard to calculate derivatives numerically!

The standard error of the mean improves with increasing sample size

We define the sample mean

$$\overline{f} = (f_1 + \cdots + f_M)/M. \qquad (3.23)$$

** This quantity is itself a random variable, because when we make another batch of M measurements and evaluate it, we won't get exactly the same answer.
** How good an estimate of the true expectation is fbar ?

$$\operatorname{var}(\overline{f}) = \operatorname{var}\Big(\frac{1}{M}(f_1 + \cdots + f_M)\Big).$$

The random variables fi are all assumed to be independent of one another, so

$$\operatorname{var}(\overline{f}) = \Big(\frac{1}{M^2}M\Big)\Big(\operatorname{var} f\Big) = \frac{1}{M}\operatorname{var} f.$$

Therefore, the sample mean becomes a better estimate of the true expectation as we average over more measurements. We call the square root of it the "standard error of the mean".

Bessel's correction: actually M-1 because when we have M=1 we don't know how good an estimator it is. Mathematically: the sample mean is known therefore only M-1 variables are independent.

Correlation and covariance

$$\mathrm{corr}(\ell, m) = \frac{\langle (\ell - \langle \ell \rangle)(m - \langle m \rangle) \rangle}{\sqrt{(\mathrm{var}\,\ell)(\mathrm{var}\,m)}}.$$

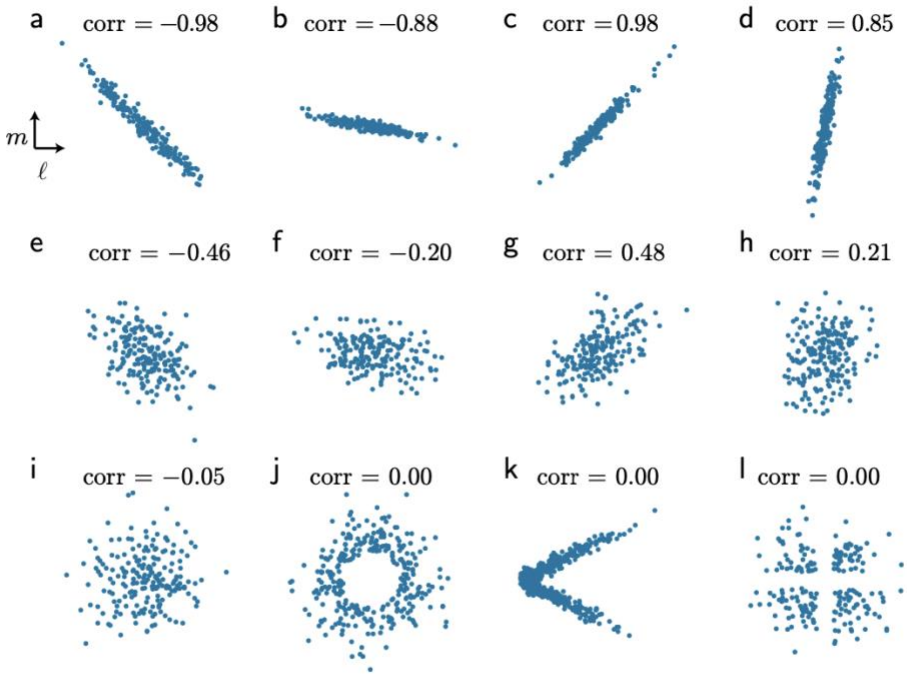


Figure 3.7: [Simulated datasets.] **Correlation coefficients of some distributions.** Each panel shows a cloud representation of a joint probability distribution, as a set of points in the ℓ-m plane; the corresponding value for $\mathrm{corr}(\ell, m)$ is given above each set. Note that the correlation coefficient reflects the noisiness and direction of a linear relationship (a–h), and it's zero for independent variables (i), but it misses other kinds of correlation (j–l). In each case, the correlation coefficient was estimated from a sample of 5000 points, but only the first 200 are shown.

Spearman correlation: rank order, then do Pearson correlation on the rank

Time correlation: C(j) = cov(f(i), f(i+j)) = 1/M sum_i cov(f(i), f(i+j)) → if > 0 then timeseries is correlated. Granger causality.

Week 3 – Using discrete distributions

- Some entropy and information measures
- Counting the number of fluorescent molecules in a cell
- The Luria-Delbrück experiment tested a model for resistance by checking a statistical prediction
- Continuous distributions

Some entropy and information measures

Say P(x) = 1/N for x_i in i=1…N
Then S(x) = - sum P ln P = ln N or $\log_2$ N
The "Effective number of species" Or "Hill number 1" is H_1 = exp(S) = N.

Renyi entropies:

The Rényi entropy of order α, where $0 < \alpha < \infty$ and $\alpha \neq 1$, is defined as[1]

$$\mathrm{H}_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n} p_i^\alpha \right).$$

With L'Hopital's rule giving

$$\mathrm{H}_1(X) \equiv \lim_{\alpha \to 1} \mathrm{H}_\alpha(X) = -\sum_{i=1}^{n} p_i \log p_i$$

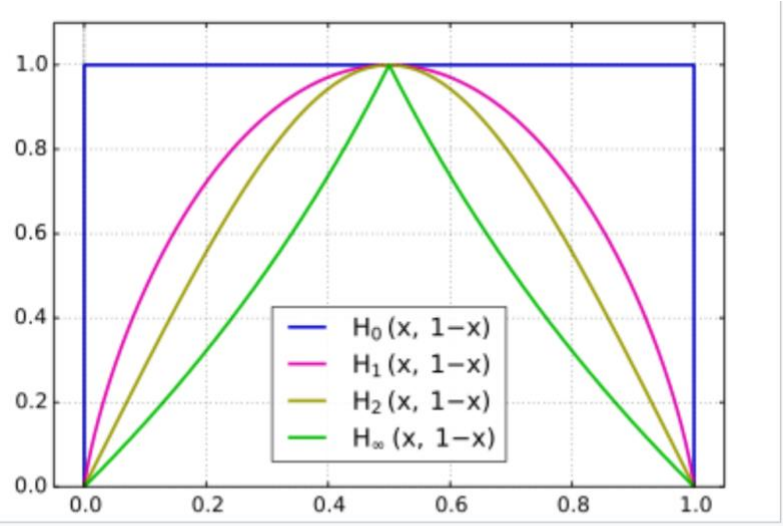
The Zero's Renyi entropy is ln sum P_i^0. Therefore exp(R_0)—number of different species, "species richness", unweighted. "Maximal entropy" or "Cardinality of the alphabet".

What's R_2 – The "Collision entropy". Related to λ, the Simpson index or (1-λ) the Gini-Simpson.

$$R_2 = -\log \sum_i P_i^2 \rightarrow H_2 = e^{R_2} = \frac{1}{\sum_i P_i^2} = \frac{1}{\lambda}$$

Similarly

$R_\infty = -\log P_i$, with i for the largest P_i, therefore, $H_\infty = P_i$



Rényi entropy of a random variable with two possible outcomes against p_1, where $P = (p_1, 1 - p_1)$. Shown are H_0, H_1, H_2 and H_∞, with the unit on the vertical axis being the shannon.

Binomial distribution

- Drawing a sample from a reservoir can be modeled via Bernoulli trials. Bernoulli trial - *s* - Succeeds with probability ξ. Mean: ξ. Variance: $\xi(1-\xi)$
- Suppose that you have 10 mL of solution containing just *four molecules* of a particular type, each of which is tagged with a fluorescent dye. Mix well and withdraw a 1 mL sample (an “aliquot”). How many of those four molecules will be in your sample?
- Find the *probability distribution* for the various values for ℓ, the number of molecules in the sample.
- The sum of several Bernoulli trials follows a Binomial distribution:

$$\mathcal{P}_{\text{binom}}(\ell; \xi, M) = \frac{M!}{\ell!(M-\ell)!}\,\xi^{\ell}(1-\xi)^{M-\ell} \quad \text{for } \ell = 0, \ldots, M \qquad \textbf{Binomial distribution} \qquad (4.1)$$

- Is normalized according to the sum rule.
- Independence of variables gives $< \ell > = M\,\xi$ and var(ℓ) = $M\,\xi\,(1-\xi)$

How to count the number of fluorescent molecules in a cell

- Fluorescence intensity, y, is proportional to the number of molecules M; that is, y = αM.
- The problem is that it is hard to estimate accurately the **normalization constant**, α, needed to convert the observable y into the desired quantity M. This constant depends on how brightly each molecule fluoresces, how much of its light is lost between emission and detection, and so on.
- Rosenfeld and coauthors found a method to *measure* α, by using a probabilistic argument.
- Cell division in bacteria splits the cell's volume into very nearly equal halves.
- Just prior to division, there are M_0 fluorescent molecules emitting light with total intensity
- y_0. After division, one daughter cell gets M_1 and the other gets $M_2 = M_0 - M_1$.
- M_1 is distributed binomially
  $\mathcal{P}_{\text{binom}}(M_1; M_0, 1/2)$.
- Therefore var(M_1) = 1/2 (1 – 1/2)M_0
- Defining the "error of partitioning" $\Delta M = M_1 - M_2$ then gives $\Delta M = M_1 - (M_0 - M_1) = 2M_1 - M_0$.
- Var(ΔM) $= 4\,\text{var}(M_1) = M_0$
- Since y = αM we have var(Δy) = α^2 var(ΔM) = α^2 $M_0 = \alpha\, y_0$, y_0 where is the fluorescence from the parent cell. Therefore,

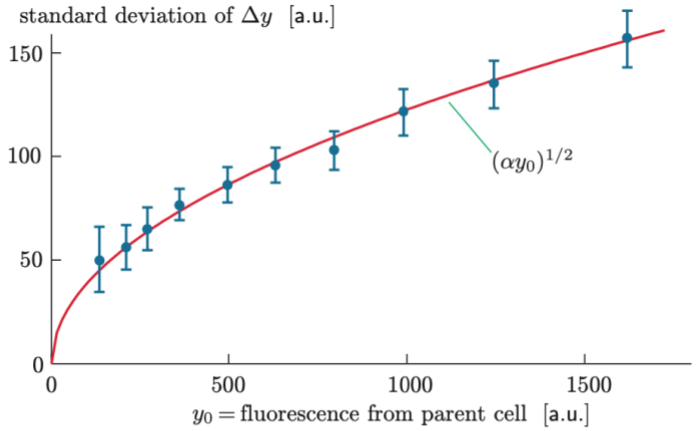


Figure 4.2: [Experimental data with fit.] **Calibration of a single-molecule fluorescence measurement.** *Horizontal axis:* Measured fluorescence intensity of cells prior to division. *Vertical axis:* Estimated standard deviation of the partitioning error of cell fluorescence after division, for cells with a particular y_0. *Error bars* indicate that this quantity is uncertain due in part to the finite number of cells observed. *Curve:* The predicted function from Idea 4.2. One global choice for the parameter α, corresponding to about 15 fluorescence units per tagged molecule, fits all the data. [Data from Rosenfeld et al., 2005.]

Note: an example of ξ not equal to ½: Asymmetric cell division. Good to make sure that at least on good one survives (such as in stem cells).

A generator from scratch, for *any* discrete distribution

- Suppose that we wish to simulate a variable ℓ drawn from $P_{binom}(\ell; M, \xi)$ with $M = 3$.
- We can do this by partitioning the unit segment into four bins of widths $(1 - \xi)^3$, $3\xi(1 - \xi)^2$, $3\xi^2(1 - \xi)$, and ξ^3, corresponding to $\ell = 0$, 1, 2, and 3 heads, respectively
- Then we draw uniformly from the line [0,1].

Poisson Distribution

- The formula for the Binomial distribution, Equation 4.1, is complicated. For example, it has two parameters, M and ξ.
- Often a simpler, approximate form of this distribution can be used instead with just *one* parameter.
  Say we draw many times, each with low probability, from an infinite reservoire. So, M->infinity and we have only one parameter.
  $$\lim_{M_* \to \infty} \mathcal{P}_{\text{binom}}(\ell; \xi, M_*),$$
  With $\mu = \xi$ M—the number of molecules captured—or otherwise $\xi = \mu / M_*$

- Substituting,

$$\lim_{M_* \to \infty} \left(\frac{\mu^\ell}{\ell!}\right)\left(1 - \frac{\mu}{M_*}\right)^{M_*}\left(1 - \frac{\mu}{M_*}\right)^{-\ell} \frac{M_*(M_* - 1)\cdots\left(M_* - (\ell - 1)\right)}{{M_*}^\ell}. \qquad (4.4)$$

The first factor of expression 4.4 doesn't depend on M_*, so it may be taken outside of the limit. The third factor just equals 1 in the large-M_* limit, and the last one is

$$(1 - M_*^{-1})(1 - 2M_*^{-1})\cdots\left(1 - (\ell - 1)M_*^{-1}\right).$$

Each of the factors above is approaching 1, and there are only a fixed number of them. Hence, in the limit the whole expression becomes another factor of 1, and may be dropped.

$$\lim_{M_* \to \infty}\left(1 - \frac{\mu}{M_*}\right)^{M_*} = \exp(-\mu). \qquad (4.5)$$

Giving,

$$\mathcal{P}_{\text{pois}}(\ell;\mu) = \frac{1}{\ell!}\,\mu^{\ell}\,\mathrm{e}^{-\mu}. \qquad \textbf{Poisson distribution} \qquad (4.6)$$

Now,

$$e^{\mu} = \sum_{l=1}^{\infty} \frac{\mu^{l}}{l!}$$

$$\frac{d}{d\mu}e^{\mu} = e^{\mu} = \frac{d}{d\mu}\sum_{l=1}^{\infty}\frac{\mu^{l}}{l!} = \sum_{l=1}^{\infty} l\,\frac{\mu^{l-1}}{l!} = \frac{1}{\mu}e^{\mu}\sum_{l=1}^{\infty} l\,e^{-\mu}\frac{\mu^{l}}{l!} = \frac{1}{\mu}e^{\mu}\langle l\rangle = e^{\mu} \rightarrow \langle l\rangle = \mu$$

Similarly, var(*l*) = μ from using the second derivative.

In short: the Poisson distribution: sum of a lot of Bernoulli trials, each with low probability giving a large number of yes-no events. Together, the total "yes" is not small. The total is Poisson distributed which we'll see looks like a Gaussian with its variance equal its mean.

THE JACKPOT DISTRIBUTION AND BACTERIAL GENETICS

- The key to understanding bacterial resistance
- Bacteria are killed by virus or AbX
- Few survive and transmit resistance to the next generation
- Even a colony made from one <u>non-resistant</u> bacterium will have some resistant survivors. How?

A colony descended from a single ancestor consists of identical individuals until a challenge to the population arises. When faced with the challenge, each individual struggles with it independently of the others, and most die. However, a small, randomly chosen subset of bacteria succeed in finding the change needed to survive the challenge, and are permanently modified in a way that they can transmit to their offspring.	***H1***

The "Darwinian" hypothesis amounted to

No mutation occurs in **response** to the challenge. Instead, the entire colony is always spontaneously mutating, whether or not a challenge is presented. Once a mutation occurs, it is heritable. The challenge wipes out the majority, leaving behind only those individuals that had previously mutated to acquire resistance, and their descendants.	***H2***

Luria-Delbrück Experiment

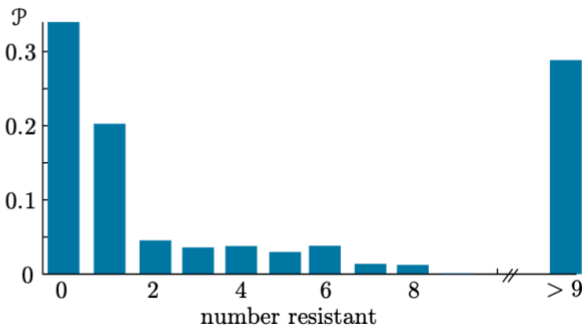


Figure 4.6: [Experimental data.] **Data from Luria and Delbrück's historic article.** The *solid bars* represent one of their trials, consisting of 87 cultures. The last bar lumps together all outliers; see text. Figure 4.8 gives a more detailed representation of the experimental data and fits to two competing models. [Data from Luria & Delbrück, 1943.]

- Used *Escherichia coli*.
- Each culture was given ample nutrients and allowed to grow for a time t_f , and then was challenged with a virus (now called "phage T1").
- "Plating" — To count the survivors, Luria and Delbrück spread each culture on a plate and continued to let them grow. Each surviving individual founded a colony, which eventually grew to a visible size.
- The survivors were few enough in number that the colonies were well separated, and so
- could be counted visually.
- Each culture had a different number m of survivors, so the experimenters reported not a single number but rather a histogram of the frequencies with which each particular value of m was observed .

The results were surprising:

- In some ways, his data looked reasonable—the distribution had a peak near m = 0, then fell rapidly for increasing m.
- But there were also **outliers**, unexpected data points far from the main group.
- Worse, when he performed the same experiment a second and third time, the outliers, while always present, were quite different in number each time.
- It was tempting to conclude that this was just a bad, unreproducible experiment! In that case, the appropriate next step would have been to work hard to find what was messing up the results (contamination?), or perhaps to abandon the whole thing. Instead, Luria and Delbrück realized that hypothesis ***H2*** could explain their odd results.

- The empirical distribution in the Luria-Delbrück experiment is said to have a **long tail**; that is, the range of values at which it's nonnegligible extends out to very large m.
- The more colorful phrase **jackpot distribution** is also used, by analogy to a gambling machine that generally gives a small payoff (or none), but occasionally gives a large one.

Two competing models for the emergence of resistance

Luria and Delbrück reasoned as follows. At the start of each trial ("time zero"), a few non-resistant individuals are introduced into each culture. At the final time t_f, the population has grown to some large number $n(t_f)$; then it is subjected to a challenge, for example, an attack by phage.

- ***H1*** in the preceding section states that each individual either mutates, with low probability ξ, or does not, with high probability $1 - \xi$, and that this random decision is made by each individual independently of the others. We have seen that in this situation, the total number m of individuals that succeed is distributed as a Poisson random variable. The data in Figure 4.6 don't seem to be distributed in this way.
- ***H2*** states that every time an individual divides, during the entire period from time zero to t_f, there is a small probability that it will spontaneously acquire the heritable mutation conferring resistance. It matters *when* that mutation occurs: Early mutants generate many resistant progeny, whereas mutants arising close to t_f don't have a chance to do so. Thus, ***H2*** implies an amplification of randomness.

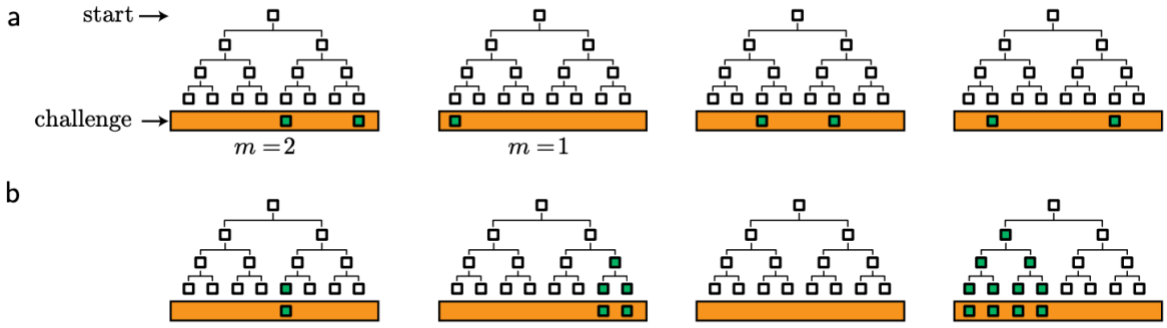
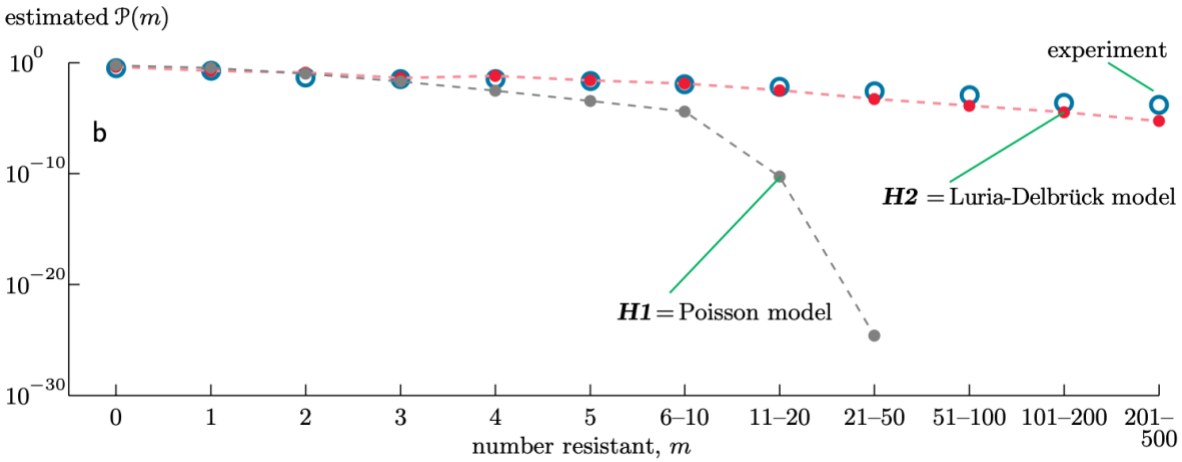


Figure 4.7: [Schematics.] **Two sets of imagined bacterial lineages relevant to the Luria-Delbrück experiment.** (a) The "Lamarckian" hypothesis ***H1*** states that bacterial resistance is created at the time of the challenge (*orange*). The number of resistant individuals (*green*) is then Poisson distributed. (b) The "Darwinian" hypothesis ***H2*** states that bacterial resistance can arise at any time. If it arises early (*last diagram*), the result can be very many resistant individuals.



Week 4 – Using continuous distributions

- The PDF. Uniform, differential entropy, Gaussian, von-Mises
- Multivariate Gaussian. Flow cytometry compensation.
- Principal Component Analysis
- Transformations of a PDF

Continuous distributions – The probability density function (PDF)

Look at the range $x_0 - \frac{1}{2}\Delta x$ to $x_0 + \frac{1}{2}\Delta x$

$$\wp_x(x_0) = \lim_{\Delta x \to 0}\left(\lim_{N_{\text{tot}} \to \infty} \frac{\Delta N}{N_{\text{tot}} \Delta x}\right). \quad \text{probability density function} \qquad (5.1)$$

- Note that the bin size is in the denominator! Dimensional.
- For $\Delta x<1$ you can have $P(x)>1$.
- Hard to define bins. Too large: lose resolution. Too small: Add noise.
- How to bin logarithmic data?
- Must check any analyses with different bin numbers to see if it changes
- It's HARD to estimate a pdf from data. A CDF is easier. Why?

Uniform Distibution:

$$\wp_{\text{unif}}(x) = \begin{cases} 1/(x_{\max} - x_{\min}) & \text{if } x_{\min} \le x \le x_{\max}; \\ 0 & \text{otherwise.} \end{cases} \qquad (5.6)$$

- All outcomes equally likely. "Microcanonical".
- How much uncertainty do we have over the outcome? Maximal (if it is bounded). It's a maximum entropy distribution.
- Differential entropy:

  P(x) = 1 for $0 \le x \le 1$. So $S = -\int P \log P = 0$
  But if $0 \le x \le {}^1/_2$ then $= -\int_0^{1/2} 2 \log 2 =$ - ln 2 < 0
  More generally, when P(x) = 1/N then S = ln N.
  The differential entropy is dangerous! It depends on the choice of coordinate.

Gaussian distribution

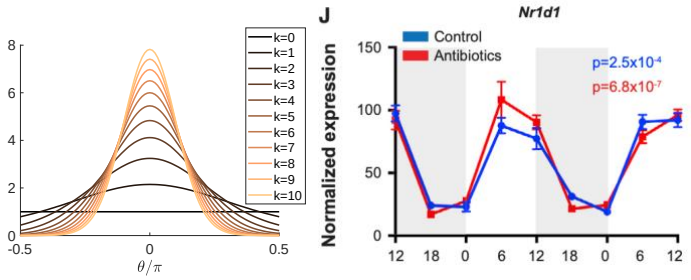
Described by two moments. <x> = μ and var(x) = σ^2.

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Entropy of a Gaussian is $\frac{1}{2}\log(2\pi e\sigma^2)$

von-Mises distribution:

Gaussian on a circle. Extends smoothly from Gaussian (small angle limit) to a straight line (k=0).



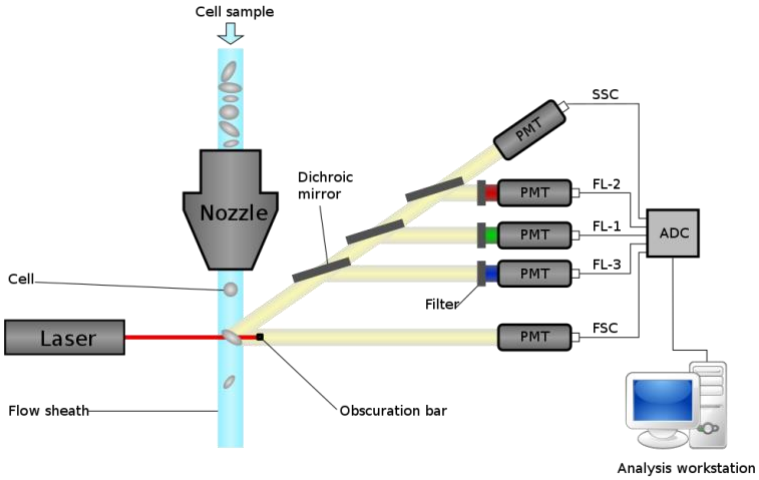
$$f(x \mid \mu, \kappa) = \frac{\exp(\kappa \cos(x-\mu))}{2\pi I_0(\kappa)}$$

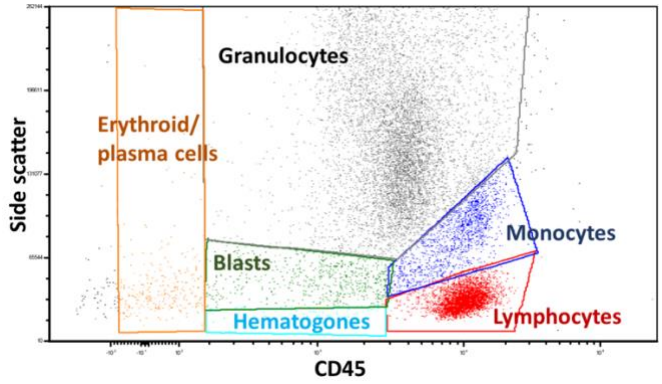
A clock gene, Thaiss,…, Elinav (2016)

Flow cytometry compensation

Brief Introduction to Flow Cytometry

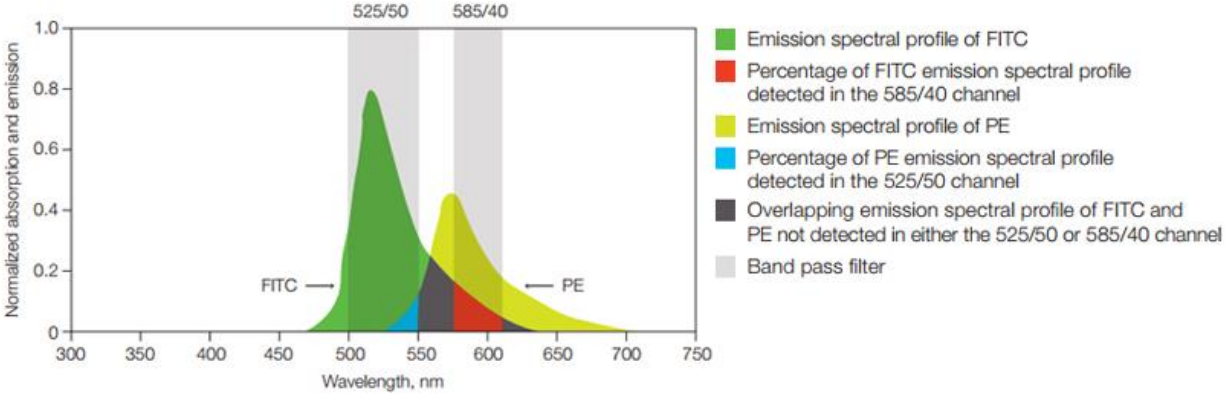
- A technique used to analyze the physical and chemical characteristics of particles in a fluid as it passes through at least one laser.
- Cells are tagged with fluorescent markers.





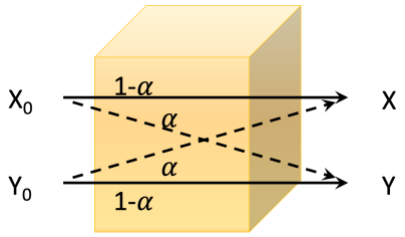
The Need for Fluorescence Compensation

- Fluorochromes have overlapping emission spectra.
- Without compensation, a single fluorochrome's emission can be detected in multiple detectors, leading to incorrect data interpretation.



How Fluorescence Compensation Works

- Adjusting the signal in each detector for the overlap.
- Essentially, it subtracts a portion of the detected signal in one channel that is attributable to another fluorochrome.
- **Can get negative values!**



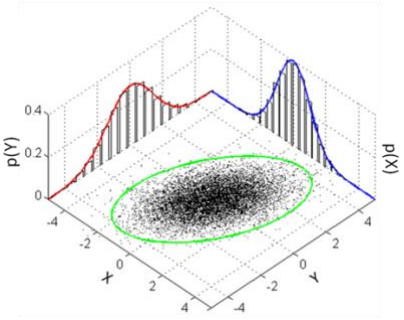
- What if $\alpha \ll 1$ but so is X_0 ?
- This is a major potential issue of false signal !
- X_0, Y_0, are fluorescence intensity, proportional to the number of antibodies. But the proportionality factor can differ widely.

$$X = (1-\alpha)X_0 + \alpha Y_0$$
$$Y = (1-\alpha)Y_0 + \alpha X_0$$
$$Cov(X,Y) = \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle$$
$$\langle X_0 Y_0 \rangle = \langle X_0 \rangle \langle Y_0 \rangle$$
$$Cov(X,Y) = \alpha(1-\alpha)(Var(X_0) + Var(Y_0)$$
$$Var(X) = (1-\alpha)^2 Var(X_0) + \alpha^2 Var(Y_0)$$

We have received the covariance matrix, a major hero of machine learning!

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

Multivariate Gaussian



$$(2\pi)^{-k/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)$$

Bivariate / Multivariate Gaussian

In the two-dimensional case, the covariance matrix:

$$f_{\mathbf{x}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}$$

Entropy: $\frac{k}{2}\ln(2\pi e) + \frac{1}{2}\ln\det(\boldsymbol{\Sigma})$

So the inverse covariance matrix gives

If $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ then $A^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

The bivariate normal distribution is the statistical distribution with probability density function

$$P(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{z}{2(1-\rho^2)}\right],$$

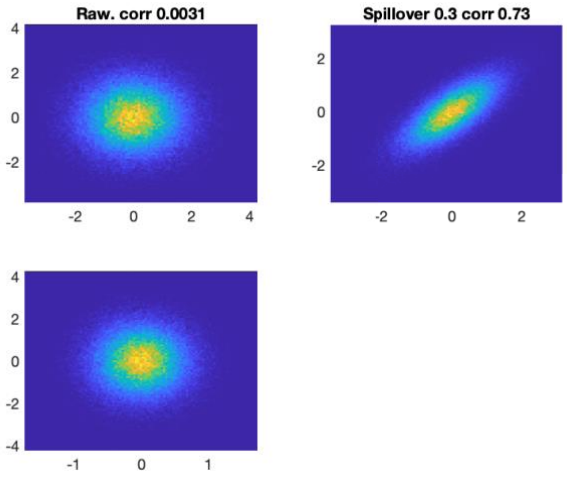
where

$$z \equiv \frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2},$$

In the bivariate case the expression for the mutual information is:

$$I(x;y) = -\frac{1}{2}\ln(1-\rho^2).$$

So – how to clean up flow cytometry data? If the input signals are Gaussian: (i) calculate the covariance matrix (ii) diagonalize it (iii) its eigenvectors are the “whitening” matrix to de-noise.



Principal Component Analysis (PCA)

Let the multivariate Gaussian

$$Prob(\boldsymbol{X}) \propto e^{-\frac{1}{2}\boldsymbol{X}^T\hat{\Sigma}^{-1}\boldsymbol{X}}$$

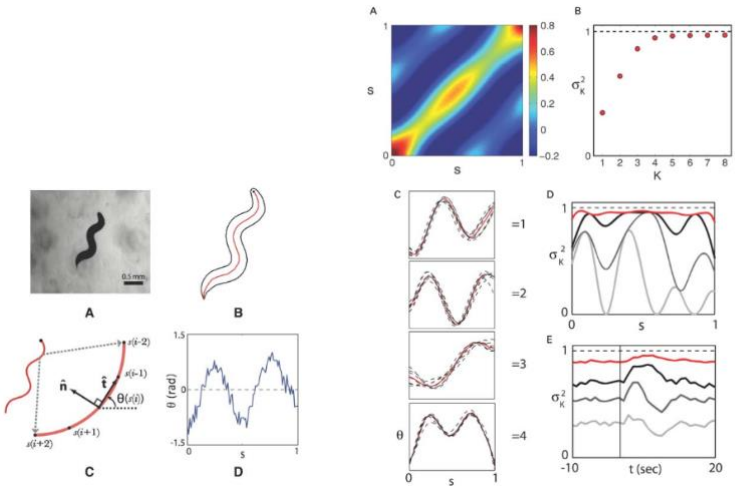
Let the covariance matrix

$$\hat{\Sigma} = cov(\boldsymbol{X})$$
$$\hat{\Sigma} = \hat{P}\tilde{\hat{\Sigma}}\hat{P}^{-1}$$
$$\boldsymbol{Y} = \hat{P}^{-1}\boldsymbol{X}$$

Then,

$$e^{-\frac{1}{2}\boldsymbol{X}^T\hat{\Sigma}^{-1}\boldsymbol{X}} = e^{-\frac{1}{2}\boldsymbol{X}^T\left(\hat{P}\tilde{\hat{\Sigma}}\hat{P}^{-1}\right)^{-1}\boldsymbol{X}} = e^{-\frac{1}{2}\boldsymbol{X}^T P\left(\tilde{\hat{\Sigma}}\right)^{-1}\hat{P}^{-1}\boldsymbol{X}} = e^{-\frac{1}{2}\boldsymbol{Y}^T\left(\tilde{\hat{\Sigma}}\right)^{-1}\boldsymbol{Y}} = \prod_{i=1}^{N} e^{-\frac{y_i^2}{2\sigma_i^2}}$$

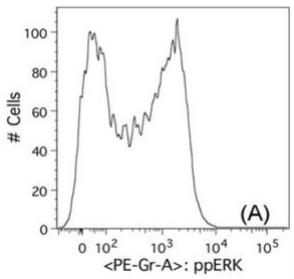
Vectors can be ranked by eigenvalue, keep only the largest eigenvalues to reduce dimension.

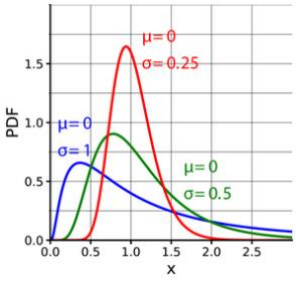


[From Stephens, Jonson-Kerner, Bialek, Ryu (2008)]

Transformations of a PDF (Erez *et al. 2018)*

Say, I measure flow cytometry data, it's over orders of magnitude! Makes sense to model as a log-normal distribution.





Let Z be a standard normal variable, and let μ and σ be two real numbers, with $\sigma > 0$.
Then, the distribution of the random variable

$$X = e^{\mu+\sigma Z}$$

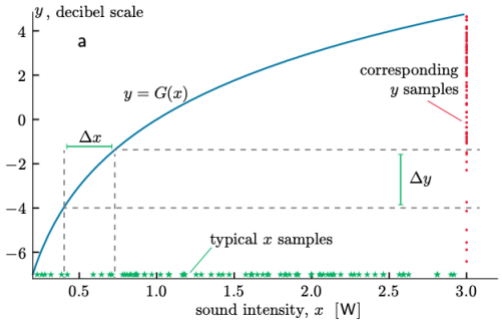
A positive random variable X is log-normally distributed (i.e., $X \sim \text{Lognormal}(\mu_x, \sigma_x^2)$), if the natural logarithm of X is normally distributed with mean μ and variance σ^2 :

$$\ln(X) \sim \mathcal{N}(\mu, \sigma^2)$$

- It appears everywhere in biology: protein counts, gene expression, tissue size, particle size distribution
  Why? Perhaps because

The geometric or multiplicative mean of n independent, identically distributed, positive random variables X_i shows, for $n \to \infty$ approximately a log-normal distribution with parameters $\mu = E[\ln(X_i)]$ and $\sigma^2 = \text{var}[\ln(X_i)]/n$, assuming σ^2 is finite.

How to transform variables in a probability distribution?



$$\left[\wp_y(G(x_0))\right]\left[(\Delta x)\frac{dG}{dx}\Big|_{x_0}\right] = \wp_x(x_0)(\Delta x).$$

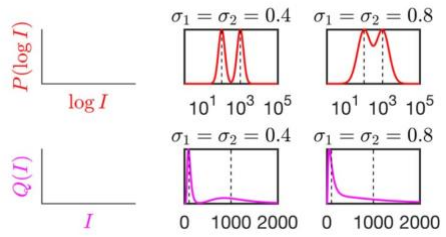
Dividing both sides by $(\Delta x)(dG/dx|_{x_0})$ gives the desired formula for $\wp_y$:

$$\wp_y(y_0) = \wp_x(x_0)\Big/\frac{dG}{dx}\Big|_{x_0} \quad \text{for monotonically increasing } G.$$

(In fact the derivative is with an absolute value if you think about the other direction)

P(log x) = P(x) / (dlogx/dx) = x P(x)

As a result, you can have two peaks in P(log I) but only one peak in Q(I) !



To simulate a specific PDF, P_x

If Y = G(x) = ln(x) then dG/dx = 1/x and Py = Px * x
If Py is uniform, the Px = | dG/dx |

To simulate a random system with a specified PDF $\wp_x$, find a function $G(x)$ whose derivative equals $\pm\wp_x$ and that maps the desired range of x onto the interval $[0, 1]$. Then apply the inverse of G to a Uniformly distributed variable y; the resulting x values will have the desired distribution. (5.22)

Ex. The probability density function $\wp(x) = e^{-x}$, where x lies between zero and infinity, will be important in later chapters. Apply Idea 5.22 to simulate draws from a random variable with this distribution.

Solution: To generate x values, we need a function G that solves $|dG/dx| = e^{-x}$. Thus,

$$G(x) = \text{const} \pm e^{-x}.$$

Applying functions of this sort to the range $[0, \infty)$, we see that the choice e^{-x} works. The inverse of that function is $x = -\ln y$.

Try applying $-\ln$ to your computer's random number generator, and making a histogram of the results.

Week 5 – Model Selection and Parameter Estimation

- Viewpoints

- Parameter Estimation. Maximum Likelihood.

- Localization microscopy
- Curvature of the likelihood function

Introduction

- model predicts not only the average value of some experimental observable taken over many trials, but its full pdf.
- We can just compare graphs of the predicted versus experimental distributions. Can't we find a more objective way to evaluate a model?
- Each model is really a *family* of models, depending on a parameter. How do we find the "right" value of the parameter?
- Suppose that one value of the parameter makes a prediction that's better on one region of the data, while another value succeeds best on a different region. Which value is better overall?
- This chapter will build on our earlier discussion of the Bayes formula to answer questions like these using the notion of *likelihood*.
- *Biological question:* Light microscopes blur everything smaller than about two hundred nanometers; how, then, can we see individual molecular motor steps?
  Physical idea: The location of a single spot can be measured to great accuracy, if we collect enough photons.

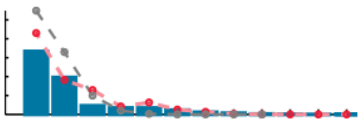


Fig. 4.8a, p. 88

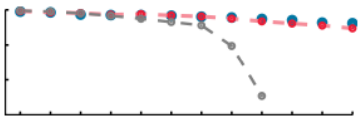


Fig. 4.8b, p. 88

Probability: we may know that a deck of cards contains 52 cards with particular markings, and model a good shuffle as one that disorganizes the cards, leaving no discernible relevant structure. Then we can ask about certain specified outcomes.

Statistics or **statistical inference:** We have already measured an outcome (or many), but we don't know the underlying mechanism that generated it. Reasoning backward from the data to the mechanism.

Viewpoints

Viewpoint 1: unknown parameter α. Given experimental data, and now I want to choose between the models, or between different parameter values in one family. So I want to find $\mathrm{P}(\mathrm{model}\alpha \mid \mathrm{data})$ and find the model or parameter value that maximizes this quantity.

Viewpoint 2: But the "probability of a model" is meaningless, because it doesn't correspond to any replicable experiment. The mutation probability α of a particular strain of bacteria under particular conditions is a definite number. We don't have a large collection of universes, each with a different value of α. We have *one* Universe.

What's meaningful is $\mathrm{P}(\mathrm{data} \mid \mathrm{model}\alpha)$. It answers the question, "If we momentarily assume that we know the true model, then how likely is it that the data we did observe *would have been* observed?" If it's unacceptably low, then we should reject the model. If we can reject all but one reasonable model, then that one has the best chance of being right.

Viewpoint 1: When you said "reject all but one model," you can always construct a highly contrived model that predicts *exactly those data*, and so will always win, despite having no foundation! Presumably you'd say that the contrived model is not "reasonable," but how do you make that precise? I'm sorry, but I really do want $\mathrm{P}(\mathrm{model}\alpha \mid \mathrm{data})$, which is different from the quantity that you proposed.

In real life, we do not know a priori the probabilities of hypotheses, nor even the sample space of all possible outcomes. Nevertheless, each of us constantly estimates *our degree of belief* in various propositions, assigning each one a value near 0 if we are sure it's false, near 1 if we are sure that it's true, and otherwise something in between.

We also constantly *update* our degree of belief in every important proposition as new information arrives. Can we systematize this process?

The Bayes formula gives a consistent approach to updating our degree of belief in the light of new data

We consider each possible model as a proposition. We wish to quantify our degree of belief in each proposition, given some data.

If we start with an initial estimate for $\wp(\text{model}_\alpha)$, then obtain some relevant experimental data, we can update the initial probability estimate by using the Bayes formula, which in this case says

$$\wp(\text{model}_\alpha \mid \text{data}) = \frac{\wp(\text{data} \mid \text{model}_\alpha)\wp(\text{model}_\alpha)}{\wp(\text{data})}. \qquad (7.1)$$

But science is supposed to be objective! It's unacceptable that your formula should depend on your initial estimate of the probability that model is true. Why should *I* care about *your* subjective estimates?

Without declaring P(model) one assumes a particular prior distribution, namely, the **uniform prior**, or $\mathrm{P}(\text{model}_\alpha)$ = constant. This sounds nice and unbiased, but really it isn't: If we re-express the model in terms of a different parameter (for example, $\beta = 1/\alpha$), then the probability density function for α must transform. In terms of the new parameter β, it generally will no longer appear uniform!

A pragmatic approach to likelihood

- We use our knowledge of the world to put together one or more physical models and attribute some prior belief to each of them. This is the step that Nora calls restricting to "reasonable" models; it seeks to eliminate the "contrived" models that Nick worries about.

- Instead of attempting an absolute statement that any model is "confirmed," we can limit our ambition to *comparing* the posterior probabilities of the set of models selected in the first step. Because they all share the common factor $1/\mathrm{P}(\text{data})$ (see Equation 7.1), we needn't evaluate that factor when deciding which model is the most probable. All we need for comparison are the **posterior ratios** for all the models under consideration, that is,

$$\frac{\wp(\text{model} \mid \text{data})}{\wp(\text{model}' \mid \text{data})} = \frac{\wp(\text{data} \mid \text{model})}{\wp(\text{data} \mid \text{model}')} \times \frac{\wp(\text{model})}{\wp(\text{model}')}. \qquad (7.2)$$

- *If the likelihood function* P(data | model) *strongly favors one model, or is very sharply peaked near one value of a model's parameter(s), then our choice of prior doesn't matter much when we compare posterior probabilities.*

- Often it suffices to compute likelihood ratios when choosing between models. This procedure is aptly named ***maximum likelihood estimation****, or "the* ***MLE*** *approach." Using an explicit prior function, when one is available, is called "****Bayesian inference****."*

Parameter Estimation – Maximum likelihood

- Suppose that a strain of laboratory animal is susceptible to a particular cancer: 17% of individuals develop the disease. Now, a test group of 25 animals is given a suspected carcinogen, and six of them develop the disease. The quantity 6/25 is larger than 0.17—but is this a significant difference?
- the best we can do is to suppose that each individual is an independent Bernoulli trial and the environment can be summarized by a single number ξ, the probability to get the disease. We wish to assess the hypothesis that the experimental group can be regarded as being drawn from a distribution with the same value of ξ as the control group.
- We will evaluate P(modelα | data), a probability distribution in α, and ask what range of α values *contains most of the posterior probability.*

We've got a model for this random system (it's a Bernoulli trial), but the model has an unknown parameter (the fairness parameter ξ), and we'd like to know whether ξ = 1/2. We'll consider three situations:
a. We observed ℓ = 6 heads out of M = 10 flips.
b. We observed ℓ = 60 heads out of M = 100 flips.
c. We observed ℓ = 600 heads out of M = 1000 flips.

Intuitively, in situation **a** we could not make much of a case that the coin is unfair: Fair coins often do give this outcome. But we suspect that in the second and third cases we could make a much stronger claim that we are observing a Bernoulli trial with $\xi \neq 1/2$.

<u>The maximally likely value for a model parameter can be computed on the basis of a finite dataset</u>

- If we have no other prior knowledge of ξ, then we use the Uniform distribution on the allowed range from ξ = 0 to 1 as our prior.
- Before we do our experiment (that is, make M flips), both ξ and the actual number ℓ of heads are unknown. After the experiment, we have some data, in this case the observed value of ℓ. Because ℓ and ξ are not independent, we can learn something about ξ from the observed ℓ. To realize this program, we compute the posterior distribution and maximize it over ξ, obtaining our best estimate of the parameter from the data.

- When we do the maximization, we *hold the observed data fixed*. The experimental data (ℓ) are frozen there in our lab notebook while we entertain various hypotheses about the value of ξ. So the factor P(ℓ), which depends only on ℓ, is a constant for our purposes; it doesn't affect the maximization. We are assuming a Uniform prior, so $\mathrm{P}(\mathrm{model}\xi)$ also doesn't depend on ξ, and hence does not affect the maximization problem. $\mathrm{P}(\mathrm{model}\xi \mid \ell) = A\mathrm{P}(\ell \mid \mathrm{model}\xi)$.

$$\mathcal{P}(\ell \mid \mathsf{model}_\xi) = \mathcal{P}_{\mathrm{binom}}(\ell; \xi, M) = \frac{M!}{\ell!(M-\ell)!}\xi^\ell(1-\xi)^{M-\ell}.$$

$$\wp(\mathsf{model}_\xi \mid \ell) = A' \times \xi^\ell(1-\xi)^{M-\ell}. \qquad (7.5)$$

We wish to maximize $\mathrm{P}(\mathrm{model}\xi \mid \ell)$, holding ℓ fixed, to find our best estimate for ξ. Equivalently, we can maximize the logarithm:

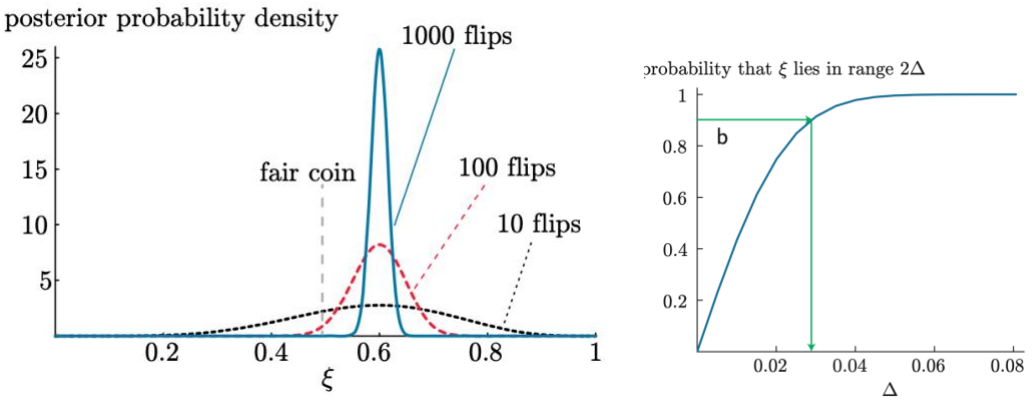
$$0 = \frac{\mathrm{d}}{\mathrm{d}\xi}\ln\wp(\mathsf{model}_\xi \mid \ell) = \frac{\mathrm{d}}{\mathrm{d}\xi}\big(\ell\ln\xi + (M-\ell)\ln(1-\xi)\big) = \frac{\ell}{\xi} - \frac{M-\ell}{1-\xi}.$$

- The maximum is at $\xi* = \ell/M$
- if a person we trust tells us that the coin is fair, then we use a prior with a maximum near $\xi = 1/2$; our best estimate of ξ then accounts for both the prior and the experimental data.

<u>The credible interval expresses a range of parameter values consistent with the available data</u>

- What's the error bar on ξ? How sharply is it peaked? Can it be a fair coin?

- The posterior distribution $\wp(\mathrm{model}\xi \mid \ell)$ is a probability density function for ξ. So we can find the prefactor A' in Equation 7.5 by requiring that $\int_0^1 \mathrm{d}\xi\ \wp(\mathrm{model}\xi \mid \ell) = 1$. The integral is not hard to compute.



Localization Microscopy

FIONA - Fluorescence imaging at one nanometer accuracy

We've discussed the use of flourescent dyes and filters to pick up only the emission wavelength. But then we have Abbe diffraction:

$$d = \frac{\lambda}{2n \sin\theta} = \frac{\lambda}{2\mathrm{NA}}$$

S for NA = 1.4-1.6 and light in 500nm we have d > 250nm. Which is small compared to most biological cells (1 μm to 100 μm), but large compared to viruses (100 nm), proteins (10 nm) and less complex molecules (1 nm).

Unfortunately, a subwavelength object, such as an individual macromolecule, appears as a blur, indistinguishable from an object a few hundred nanometers in diameter.

To break through this impasse, first note that for some problems, we do not need to form a full image. For example, molecular motors are devices that convert "food" (molecules of ATP) into mechanical steps. In order to learn about the stepping mechanism in a particular class of motors, it's enough to label an individual motor with a fluorophore. For that problem, we don't really need to resolve two nearby points. Instead, we have one point source of light (a fluorophore attached to the motor), and we wish to determine its position accurately enough to detect and measure individual steps.

Although the pixels fire at random, they have a definite probability distribution, called the **point spread function** of the microscope (Figure 7.3b). If we deliberately move the sample by a tiny, known amount, the smeared image changes only by a corresponding shift. So we need to measure the point spread function only *once*; thereafter, we can think of the true location of the fluorophore, (μx,μy), as a pair of parameters describing a family of hypotheses. Each hypothesis is described by a known likelihood function—the shifted point spread function. Maximizing the likelihood over the parameters then tells what we want to know: Where is the source?

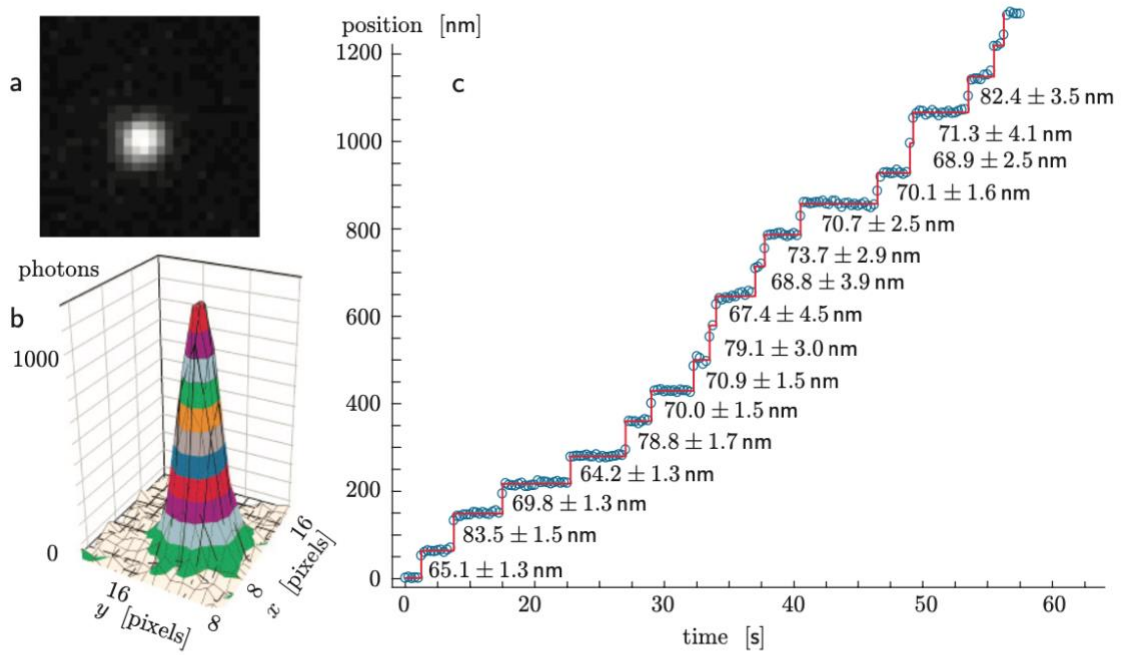


Figure 7.3: [Experimental data.] **FIONA imaging.** (a) One frame from a video micrograph of the movement of a single fluorescent dye attached to the molecular motor protein myosin-V. Each camera pixel represents 86 nm in the system, so discrete, 74 nm steps are hard to discern in the video (see Media 7). (b) Another representation of a single video frame. Here the number of light blips collected in each pixel is represented as height. The center of this distribution can be determined to accuracy much better than the value suggested by its spread. (c) The procedure in the text was applied to each frame of a video. A typical trace reveals a sequence of ≈ 74 nm steps. The horizontal axis is time; the vertical axis is position projected onto the line of average motion. Thus, the steps appear as vertical jumps, and the pauses between steps as horizontal plateaux. [Courtesy Ahmet Yildiz; see also Media 7 and Yildiz et al., 2003.]

$$\boxed{\wp(x_1, \ldots, x_M \mid \mu_x) = \wp_{\text{gauss}}(x_1; \mu_x, \sigma) \times \cdots \times \wp_{\text{gauss}}(x_M; \mu_x, \sigma)}$$

physical model, localization microscopy (simplified)

or

$$\ln \wp(x_1, \ldots, x_M \mid \mu_x) = \sum_{i=1}^{M} \left[-\tfrac{1}{2} \ln(2\pi\sigma^2) - (x_i - \mu_x)^2/(2\sigma^2) \right]. \qquad (7.6)$$

Therefore $\frac{d \ln P}{d\mu} = 0 = \sum_i \frac{x_i - \mu}{\sigma^2} \rightarrow \sum_i x_i = M\,\mu \rightarrow \mu = \frac{1}{M}\sum_i x_i$. Perhaps not surprising. But how good is this estimate? Rearranging Equation 7.6 gives the log likelihood as

$$\begin{aligned} \text{const} - \tfrac{1}{2\sigma^2} \sum_i ((x_i)^2 - 2x_i\mu_x + (\mu_x)^2) &= \text{const}' - \tfrac{M}{2\sigma^2}(\mu_x)^2 + \tfrac{1}{2\sigma^2} 2M\bar{x}\mu_x \\ &= \text{const}'' - \tfrac{M}{2\sigma^2}(\mu_x - \bar{x})^2. \qquad (7.7) \end{aligned}$$

The constant in the second expression includes the term with the sum of x^2. This term does not depend on μ , so it is “constant” for the purpose of optimizing over that desired quantity.

Exponentiating the third form of Equation 7.7 now shows that the full posterior PDF is in fact a Gaussian. Its variance equals σ^2/M, agreeing with our earlier result.

Also, from Eq. 7.6, $\frac{d \ln P}{d\,\sigma^2} = 0 \rightarrow \sum_{i=1}^{M} \frac{(x_i-\mu)^2}{(2\,\sigma^4} - \frac{M}{2\sigma^2} = 0 \rightarrow \sigma^2 = \frac{1}{M}\sum_{i}^{M}(x_i - \mu)^2$

Yildiz and coauthors applied this method to the successive positions of the molecular motor myosin-V, obtaining traces like those in Figure 7.3c. Each such "staircase" plot describes the progress of a single motor molecule. The figure shows the motion of a motor that took a long series of rapid steps, of length always near to 74 nm. Between steps, the motor paused for various waiting times. Chapter 9 will study those waiting times in greater detail.

Complete images: PALM/FPALM/STORM

- More generally, we'd like to get an *image,* that is, a representation of the positions of *many* objects. For example, we may wish to see the various architectural elements in a cell and their spatial relationships.
- We mark the objects of interest with many fluorophores (in this context called "tags") and model the light distribution as the sum of multiple point spread functions, each centered on a different unknown location
- In the mid-1990s, E. Betzig outlined a fruitful approach: to arrange that each emitter be somehow different from its neighbors.

- R Dickson found that GFP has a long-lived "dark" conformation that does not fluoresce. Unexpectedly, they found that they could pop individual molecules from this dark state to a fully fluorescent state by activation with light of wavelength 405 nm.
- Can turn on individual molecules in a time-resolved way. (See images from book).

Curvature of the likelihood function

Model fitting

Suppose that we make trials at each of several typical x values and find that, for each fixed x, the observed y values have a Gaussian distribution about some expectation $\bar{y}(x)$. Suppose further that the variances of each of these distributions are all the same constant value σ^2 independent of x, as in Figure 7.6b. If we have reason to believe that $\bar{y}(x)$ depends on x via a linear function, that is, $\bar{y}(x) = Ax + B$, then we'd like to know the values of A and B that are best supported by the data.

$$\wp(y \mid \text{model}_{A,B}; x) = \wp_{\text{gauss}}(y; Ax + B, \sigma),$$

generic physical model for linear relation

$$\wp(y_1, \ldots, y_M \mid \text{model}_{A,B}; x_1, \ldots, x_M) = \wp_{\text{gauss}}(y_1; Ax_1 + B, \sigma) \times \cdots \times \wp_{\text{gauss}}(y_M; Ax_M + B, \sigma),$$

or

$$\ln \wp(y_1, \ldots, y_M \mid \text{model}_{A,B}; x_1, \ldots, x_M) = \sum_{i=1}^{M} \left[-\tfrac{1}{2} \ln(2\pi\sigma^2) - (y_i - Ax_i - B)^2/(2\sigma^2) \right].$$

To find the optimal fit, we maximize this quantity over A and B holding the xi and yi fixed. We can neglect the first term in square brackets, because it doesn't depend on A or B. In the second term, we can also factor out the overall constant (2σ2)−1. Finally, we can drop the overall minus sign and seek the *minimum* of the remaining expression:

> ***Under the assumptions stated, the best-fitting line is the one that minimizes the* chi-square statistic** $\sum_{i=1}^{M} (y_i - Ax_i - B)^2/\sigma^2$. (7.8)

Because we have assumed that every value of x gives y values with the same variance, we can even drop the denominator when minimizing this expression.

Idea 7.8 suggests a more general procedure called **least-squares fitting**. Even if we have a physical model for the experiment that predicts a *non*linear relation $y(x) = F(x)$, we can still use it, simply by substituting $F(x)$ in place of $Ax + B$ in the formula.

- We assumed that, for fixed x, the variable y was Gaussian distributed. Many experimental quantities are not distributed in this way.
- We assumed that var $y(x) = \sigma$2 was independent of x. Often this is not the case.

If we assume different variance for different points, then

$$P(\{x_i\}|\mu_i, \sigma_i) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\sum_i \frac{(x_i-\mu)^2}{2\sigma_i^2}}$$

$$\frac{d \ln P}{d\mu} = 0 \to \frac{d}{d\mu} \sum_i \frac{(x_i - \mu)^2}{2\sigma_i^2} = 0 \to \sum_i \frac{(x_i - \mu)}{\sigma_i^2} = 0 \to \mu = \left. \sum_i \frac{x_i}{\sigma_i^2} \middle/ \sum_i \frac{1}{\sigma_i^2} \right.$$

Using Cramers-Rao:

$$E\left[\frac{d \ln P}{d\,\mu^2}\right] = -\sum_i \frac{1}{\sigma_i^2} \to Var[\mu] \geq \frac{1}{\sum_i \frac{1}{\sigma_i^2}}$$

If all sigmas are the same we have $\sigma_\mu^2 = \frac{\sigma^2}{M}$

Cramér–Rao bound

$$\operatorname{var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

where the Fisher information $I(\theta)$ is defined by

$$I(\theta) = n \operatorname{E}_{X;\theta}\left[\left(\frac{\partial \ell(X;\theta)}{\partial \theta}\right)^2\right]$$

and $\ell(x;\theta) = \log(f(x;\theta))$ is the natural logarithm of the likelihood function for a single sample x and $\operatorname{E}_{x;\theta}$ denotes the expected value with respect to the density $f(x;\theta)$ of X. If not indicated, in what follows, the expectation is taken with respect to X.

If $\ell(x;\theta)$ is twice differentiable and certain regularity conditions hold, then the Fisher information can also be defined as follows:[9]

$$I(\theta) = -n \operatorname{E}_{X;\theta}\left[\frac{\partial^2 \ell(X;\theta)}{\partial \theta^2}\right]$$

→ *The Fisher information is the curvature of the log-likelihood function*

If $\log f(x;\theta)$ is twice differentiable with respect to θ, and under certain regularity conditions, then the Fisher information may also be written as[8]

$$\mathcal{I}(\theta) = -\operatorname{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(X;\theta) \;\middle|\; \theta\right],$$

since

$$\frac{\partial^2}{\partial \theta^2} \log f(X;\theta) = \frac{\frac{\partial^2}{\partial \theta^2} f(X;\theta)}{f(X;\theta)} - \left(\frac{\frac{\partial}{\partial \theta} f(X;\theta)}{f(X;\theta)}\right)^2 = \frac{\frac{\partial^2}{\partial \theta^2} f(X;\theta)}{f(X;\theta)} - \left(\frac{\partial}{\partial \theta} \log f(X;\theta)\right)^2$$

and

$$\operatorname{E}\left[\frac{\frac{\partial^2}{\partial \theta^2} f(X;\theta)}{f(X;\theta)} \;\middle|\; \theta\right] = \frac{\partial^2}{\partial \theta^2} \int_{\mathbb{R}} f(x;\theta)\, dx = 0.$$

Week 6 – Poisson Processes and Their Simulation

- How does the molecular motor of the myosin-V

THE KINETICS OF A SINGLE-MOLECULE MACHINE

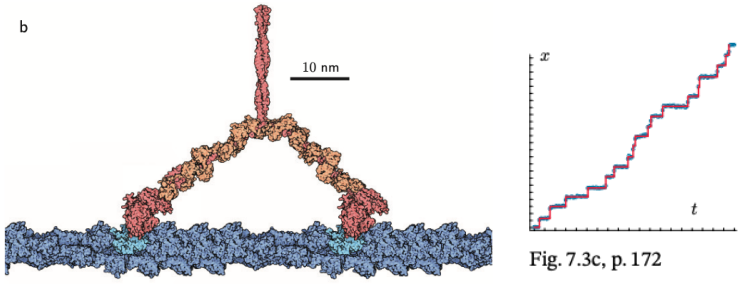


Fig. 7.3c, p. 172

(b) The myosin-V molecule has two “legs,” which join its “feet” to their common “hip,” allowing it to span the 36 nm separation between two binding sites (*light blue*) on an actin filament (*blue*). [Art by David S Goodsell.]

Two “feet” walk on a “track” e.g. actin or tubulin. The feet are subunits recognizing binding sites regularly spaced on the track. ATP → ADP energy is used to disconnect the foot and jump to the next site (ratchet). In a large muscle there are about 10^{19} myosin molecules pulling together. (For comparison: 10^{11} galaxies, the milky way has about that number of stars). 1 ATP produces about 6nN of force.

- Figure 7.3c: Chapter 7 introduced myosin-V and described how Yildiz et al. visualized its individual steps. As shown in, the motor’s position as a function of time looks like a staircase.

- The figure shows an example with rapid rises of nearly uniform height, corresponding to 74 nm steps. But the *widths* of the stairs in that figure, corresponding to the waiting times between steps, are quite nonuniform.
- The motor’s progress consists of sudden steps, spread out between widely variable pauses. And yet, the overall trend of the trace in Figure 7.3c does seem to be a straight line of definite slope. We need to make this intuition more precise.
- Each step requires that the motor bind an ATP molecule. ATPs are available, but they are greatly outnumbered by other molecules, such as water.
- So the motor’s ATP-binding domain is bombarded by molecular collisions at a very high rate, but almost all collisions are not “productive”; that is, they don’t lead to a step. Even when an ATP does arrive, it may fail to bind, and instead simply wander away.
- We also assume that after a productive collision, the internal state resets; the motor has no memory of having just taken a step.
- The output of each trial is not a single number but is an entire *time series of steps* (the staircase plot).

- Each step advances the molecule by about the same distance; thus, to describe any particular trial, we need only state the list of times $\{t_1, t_2, \ldots, t_N\}$ when steps occurred.
- A random system with this sort of sample space is called a **random process**.
- Because the motor is assumed to have no memory of its past, we fully specify the process when we state the collision interval Δt and productive-step probability ξ.
- This Markov property greatly simplifies the analysis.
- We are considering a physical model of molecular stepping that idealizes each collision as independent of the others, and also supposes them to be simple Bernoulli trialss

Let E_* denote the event that a step happened at time slot i. Then to characterize the discrete-time stepping process, we can find the probability that, given E_*, the *next* step takes place at a particular time slot $i + j$, for various positive integers j. Call this proposition "event E_j." We seek the conditional probability $P(E_j \mid E_*)$.

- More explicitly, $P(E_*)$ is the probability that a step occurred at slot i, *regardless of what happened on other slots.*
- To find the conditional probability $P(E_j \mid E_*) = P(E_j \textbf{ and } E_*)/P(E_*)$
- In an interval of duration T, there are $N = T/\Delta t$ time slots. Each outcome of the random process is a string of N Bernoulli trials (step/no-step in time slot 1, ... , N). E_* is the subset of all possible outcomes for which there was a step at time slot i. Because they are independent they can be "integrated out" then $P(E_*) = \xi$.
- Similarly, $P(E_j \textbf{ and } E_*) = \xi(1-\xi)^{j-1}\xi$
- Therefore, $P(E_j \mid E_*) = \xi(1-\xi)^{j-1} = P_{\text{geom}}(j;\xi)$, for $j = 1,2, \ldots$.
- Thus, like the Binomial, Poisson, and Gaussian distributions, the Geometric distribution also has its roots in the Bernoulli trial.

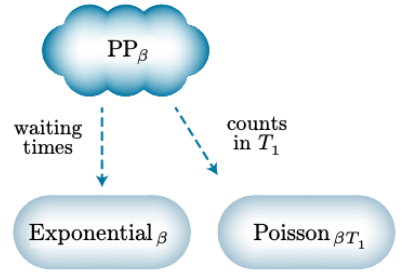
A POISSON PROCESS CAN BE DEFINED AS A CONTINUOUS-TIME LIMIT OF REPEATED BERNOULLI TRIALS

- Often it's not appropriate to treat time as discrete. For example, as far as motor stepping is concerned, nothing interesting is happening on the molecular collision time scale Δt.
- We consider a *limit*, $\Delta t \to 0$. If such a limit makes sense, then our formulas will have one fewer parameter (the irrelevant Δt will disappear) .
- We now show that the limit does make sense, and gives rise to a one-parameter family of continuous-time random processes called Poisson processes.
- Poisson processes arise in many contexts, so from now on we will use the word "blip" referring to a sudden event.
- The total number of time slots in a fixed interval T is $T/(\Delta t)$, which approaches infinity as $\Delta t \to 0$. If we were to hold ξ fixed, then the total number of blips expected in the interval T, that is, $\xi T/\Delta t$, would also become infinite. To get a reasonable limit, then, we must imagine a series of models in which ξ is *also* assumed to be small.
- *$\xi = \beta\Delta t$, independent of what is happening in any other interval, and we take the continuous-time limit $\Delta t \to 0$ holding β fixed.*

- The constant β is called the **mean rate** (or simply "rate") of the Poisson process; it has dimensions $1/\mathbb{T}$. The components ξ and Δt are irrelevant.

Example: Taking Δt = 1 μs, you conclude that the is satisfied with β = 5⁄s. But your friend takes Δt = 2 μs. Will you agree?
Answer: The probability for an event in my time window is tiny, so even at a double window the probability for two events is small enough to be neglected, leading to the same statistics.

Poisson *process* vs. Poisson *distribution:* Each draw from the Poisson distribution is a single integer; each draw from the Poisson process is a sequence of real numbers $\{t\alpha\}$. However, there is a connection. The distribution of counts of a Poisson process is a Poisson distribution.



Waiting times are Exponentially distributed:

- The interval between successiveblips iscalled the **waiting time**, t_w
- The PDF of the waiting time is the discrete distribution divided by Δt:

$$\wp(t_w) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathcal{P}_{\text{geom}}(j; \xi). \qquad (9.4)$$

In this formula, $t_w = (\Delta t)j$ and $\xi = (\Delta t)\beta$, with t_w and β held fixed as $\Delta t \to 0$. To simplify Equation 9.4, note that $1/\xi \gg 1$ because Δt approaches zero.

$$\wp(t_w) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \xi (1-\xi)^{(t_w/\Delta t)-1} = \lim_{\Delta t \to 0} \frac{\xi}{\Delta t} \left((1-\xi)^{(1/\xi)}\right)^{(t_w \xi/\Delta t)} (1-\xi)^{-1}.$$

Taking each factor in turn,

- $\xi/\Delta t = \beta$.
- The middle factor involves $(1-\xi)^{(1/\xi)}$. The compound interest formula[10] says that this expression approaches e^{-1}. It is raised to the power $t_w\beta$.
- The last factor approaches 1 for small ξ.

With these simplifications, we find a family of continuous PDFs for the interstep waiting time:

The waiting times in a Poisson process are distributed according to the **Exponential distribution** $\wp_{\text{exp}}(t_w; \beta) = \beta e^{-\beta t_w}$. (9.5)

The mean waiting time, $\langle t_w \rangle$ follows,

$$\langle t_w \rangle = \int_0^\infty t_w \beta e^{-\beta t_w} dt_w = -\int_0^\infty t_w \frac{d\, e^{-\beta t_w}}{dt_w} dt_w = -t_w e^{-\beta t_w} \Big|_0^\infty + \int_0^\infty e^{-\beta t_w} dt_w$$
$$= \frac{1}{\beta}$$

Similarly, $\langle t_w^2 \rangle = \frac{2}{\beta^2}$ and therefore $Var[t_w] = \frac{2}{\beta^2} - \frac{1}{\beta^2} = \frac{1}{\beta^2}$

We can use the waiting time distribution to fit experimental data!

Counts are Poisson distributed:

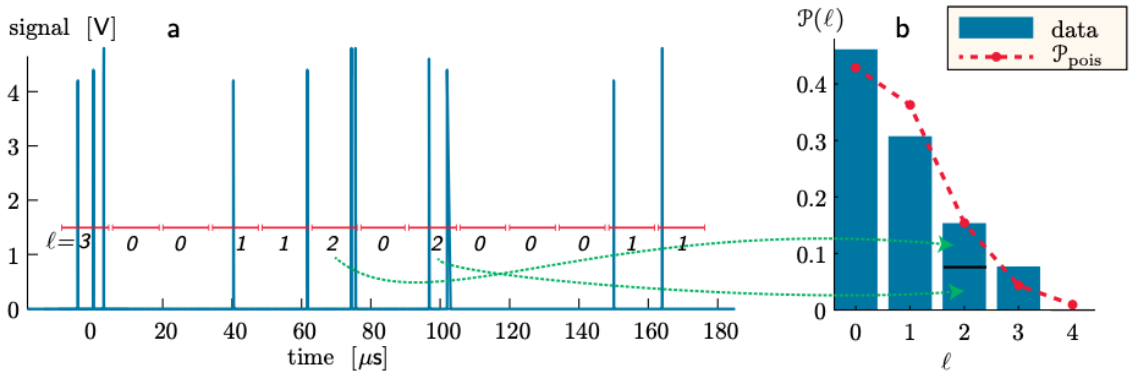


Figure 9.6: [Experimental data with fit.] **The count distribution of a Poisson process over fixed intervals (Idea 9.6).** (a) The same 11 blips shown in Figure 9.4a. The time interval has been divided into equal bins, each of duration $T_1 = 13$ s (*red*); the number of blips in each bin, ℓ, is given beneath its bin indicator. (b) On this graph, *bars* indicate estimates of the probability distribution of ℓ from the data in (a). *Dotted arrows* connect the instances of $\ell = 2$ with their contributions to the bar representing this outcome. The *red dots* show the Poisson distribution with expectation equal to the sample mean of the observed ℓ values. [Data courtesy J F Beausang (Dataset 12).]

How many blips will we observe in a fixed, finite time interval T?

To approach the question, we again begin with the discrete-time process, regarding the interval T as a succession of $M = T/\Delta t$ time slots. The total number of blips, ℓ, equals the sum of M Bernoulli trials. For a Poisson process with mean rate β, the probability of getting ℓ blips in any time interval T is $P_{pois}(\ell; \beta T)$.

With $< \ell > = \beta T$

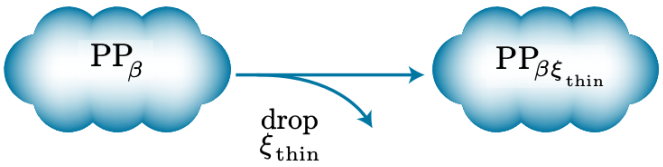
This can explain why the staircase plot has a *mean slope* despite the random waiting times.

Thinning a Poisson process results in another Poisson process

Suppose that we have a Poisson process with mean rate β. We now create another random process. Each time we draw a blip sequence from the original process, we accept or reject individual blips based on independent Bernoulli trials with probability ξ_{thin}, reporting only the

times of the accepted blips. The **thinning property** states that the new process is also Poisson, but with mean rate reduced from β to $\xi_{\text{thin}}\beta$.

To prove this result, again, divide time into slots Δt so small that there is negligible probability to get two or more blips in a slot



A regularly spaced timeseries like this would not remain regularly spaced, but a Poisson process would remain a Poisson process.

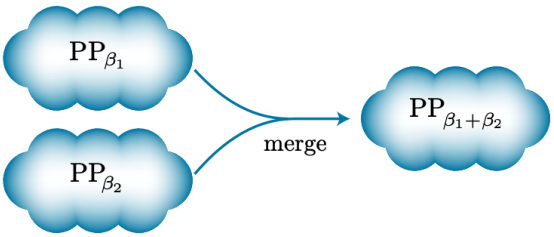
Merging two Poisson processes also results in another Poisson process

Suppose that we have two independent Poisson processes, generating distinct types of blips with mean rate β_1 and β_2. We can define a "merged process" that reports the arrival times of *either* kind of ball. The **merging property** states that the merged process is itself Poisson, with mean rate $\beta_{\text{tot}} = \beta_1 + \beta_2$.

To prove it, again divide time into small slots Δt. Then $(\beta_1\Delta t)+(\beta_2\Delta t) = \beta_{\text{tot}}\Delta t$

(From $P(E_1 \text{ and } E_2) = P(E_1) + P(E_2)$ for exclusive events).

Probability the blip is type 1: $\beta_1 / \beta_{\text{tot}}$



Some biological contexts

- Section 9.2 imagined the stepping of myosin-V as a result of two sequential events: First an ATP molecule must encounter the motor's ATP-binding site, but then it must also bind and initiate stepping. It's reasonable to model the first event as a Poisson process, because most of the molecules surrounding the motor are not ATP and so cannot generate a step. It's reasonable to model the second event as a Bernoulli trial, because even when an ATP does encounter the motor, it must overcome an activation barrier to bind; thus, some fraction of the encounters will be nonproductive. The thinning property leads us to expect that the complete stepping process will itself be Poisson, but

with a mean rate lower than the ATP collision rate. We'll see in a following section that this expectation is correct.

- Photons (particles of light) arrive at your eye in a Poisson process, but many are randomly "lost" by being scattered or by being absorbed by things other than your photoreceptors. Nevertheless, photon absorptions by photoreceptors follow a Poisson process, by the thinning property. This fact lets us apply simple models to visual reception.
- Suppose that two or more identical enzyme molecules exist in a cell, each continually colliding with other molecules, a few of which are substrates for a reaction that the enzymes catalyze. Each enzyme then emits product molecules in a Poisson process, just as in the motor example. The merging property leads us to expect that the *combined* production will also be a Poisson process.

MULTISTEP PROCESSES AND CONVOLUTION

Myosin-V is a processive molecular motor whose stepping times display a dual character

- Two feet rarely detach together, letting it take many steps.
- The graph shows each step advancing the motor by sudden jumps of roughly 74 nm. Interestingly, however, only about one quarter of the individual myosin-V molecules studied had this character.
- The others *alternated* between short and long steps; the *sum* of the long and short step lengths was about 74 nm. This division at first seemed mysterious—were there two distinct kinds of myosin-V molecules? Was the foot-over-foot mechanism wrong?
- A. Yildiz and coauthors proposed a simpler hypothesis to interpret their data:
  All the myosin-V molecules are in fact stepping in the same way along their actin tracks. In this experiment, the anomalous subpopulation differed only in where on the myosin-V molecule the fluorescent marker was attached.

Poisson process, with mean rate β depending on the concentration of ATP.
Hence, the PDF of interstep waiting times should be an Exponential distribution.
In fact, the subpopulation of myosin-V motors with alternating step lengths really did obey this prediction (see Figure 9.10a), as do the kinetics of many other chemical reactions. But for the other subpopulation (the motors that took 74 nm steps), the prediction failed badly (Figure 9.10b).

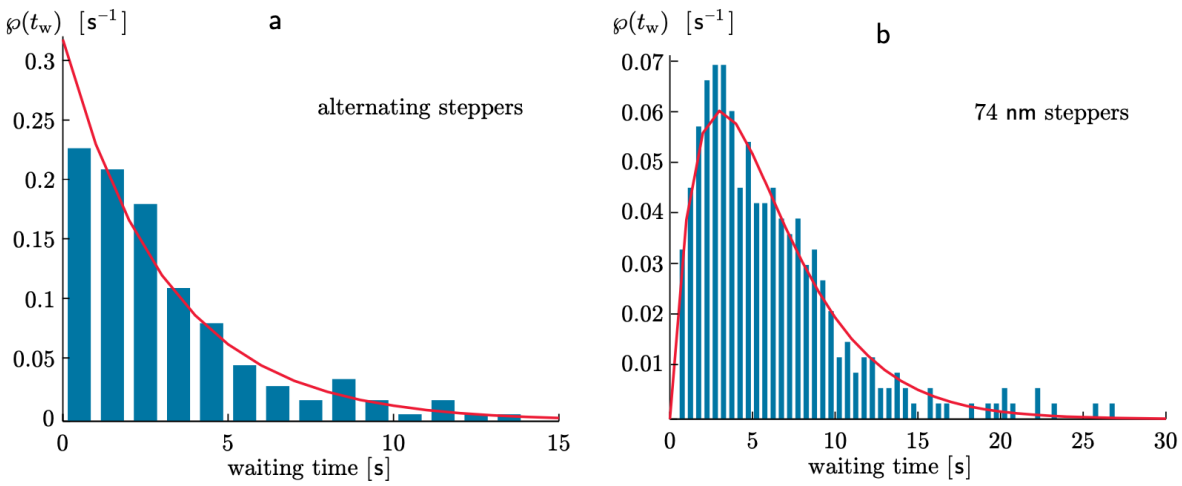


Figure 9.10: [Experimental data with fits.] **The stepping of molecular motors.** (a) Estimated PDF of the waiting times for the subpopulation of myosin-V molecules that displayed alternating step lengths, superimposed on the expected Exponential distribution (Equation 9.9). (b) Similar graph for the other subpopulation of molecules that displayed only long steps, superimposed on the two-step distribution derived in Equation 9.9. The shape of the curve in (b) is the signature of a random process with two alternating types of substep. Each type of substep has Exponentially distributed waiting times with the same mean rate as in (a), but only one of them is visible. [Data from Yildiz et al., 2003.]

In the subpopulation of 74 nm steppers, the first, third, fifth, . . . steps are *not visible*. Therefore, what appears to be the αth interstep waiting time, t' , is actually the *sum* of two consecutive waiting times:
$t_{w',\alpha} = t_{w,2\alpha} + t_{w,2\alpha-1}$.

Even if the true waiting times are Exponentially distributed, we will still find that the apparent waiting times t' have a different distribution: the convolution. Thus,

$$\wp_{t'_w}(t'_w) = \int_0^{t'_w} dx\, \wp_{exp}(x;\beta) \times \wp_{exp}(t'_w - x;\beta), \qquad (9.8)$$

where x is the waiting time for the first, invisible, substep.

$$\beta^2 \int_0^{t'_w} dx\, \exp(-\beta x - \beta(t'_w - x)) = \beta^2 e^{-\beta t'_w} \int_0^{t'_w} dx = \beta^2 t'_w e^{-\beta t'_w}. \qquad (9.9)$$

In fact, fitting the histogram in Figure 9.10a leads to a value for the mean rate β, and hence to a definite prediction (no further free parameters) for the histogram in Figure 9.10b.

Note that in general, the Gamma distribution

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

Here with alpha=2 for a two-step process.

Relative standard deviation can be used to reveal substeps in a kinetic scheme

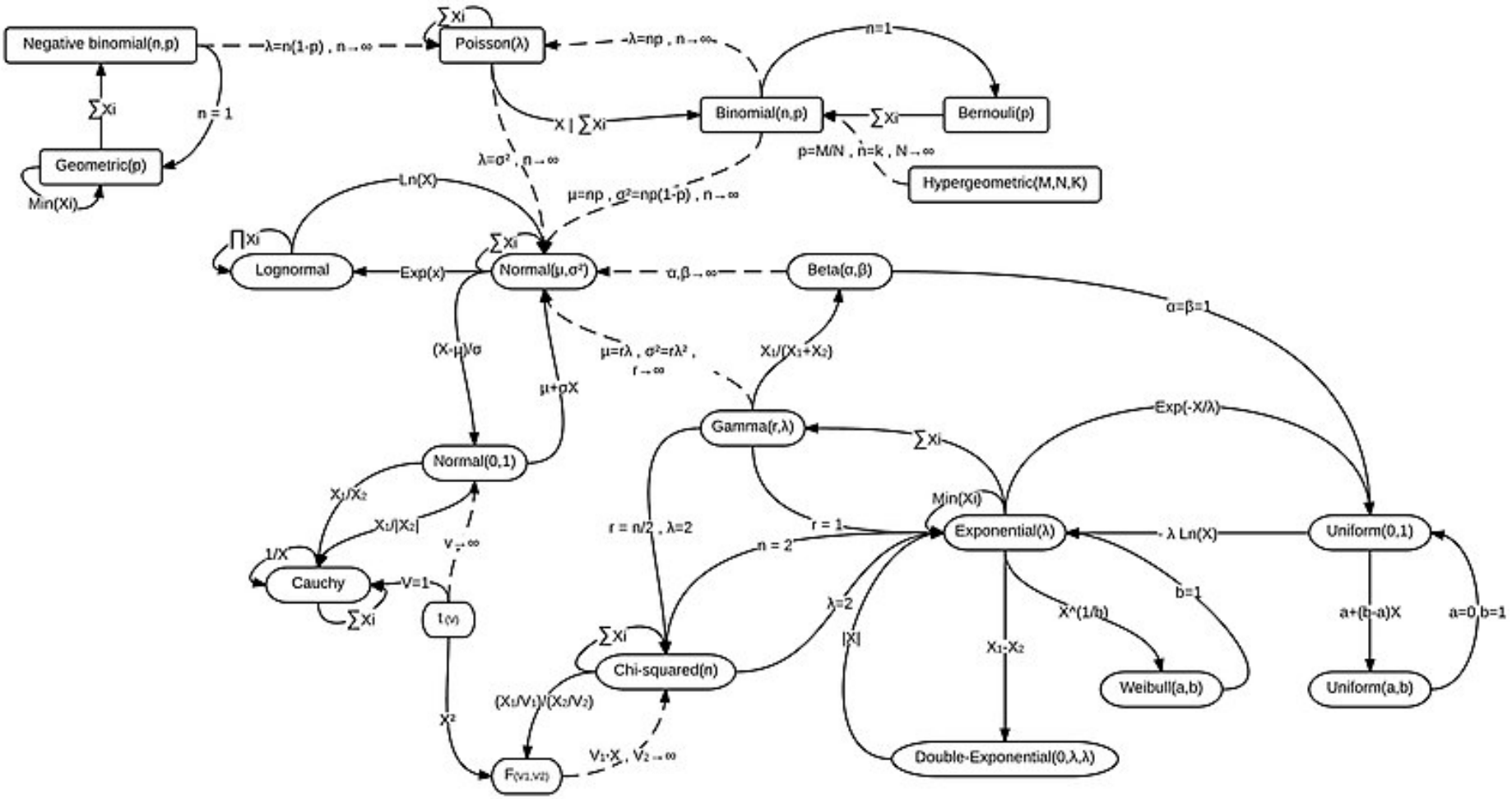
Note that $E[t_w] = 2/\beta$ and $\mathrm{Var}[t_w] = 2/\beta^2$. So $\mathrm{root(var)/mean} = 1/\mathrm{root}(2)$

But for an exponential distribution $\mathrm{root(var)/mean} = 1$.

This is a statistical test of the number of steps the motor takes without finding beta, just by computing the mean and variance.

Simulating simple Poisson processes

- We can avoid stepping through the vast majority of time slots in which nothing happens. We just generate a series of Exponentially distributed intervals $t_{w,1}$, ..., then define the time of blip α to be $t_\alpha = t_{w,1} + \cdots + t_{w,\alpha}$, the accumulated waiting time.
- A computer's basic random-number function has a Uniform, not an Exponential, distribution. However, we can convert its output to get what we need, by adapting the Example on page 113. This time the transformation function is $G(t_w) = e^{-\beta t_w}$, whose inverse gives $t_w = -\beta^{-1} \ln y$.
- Suppose that we wish to simulate a process consisting of two types of blip. Each type arrives independently of the other, in Poisson processes with mean rates β_a and β_b, respectively. We could simulate each series separately and merge the lists, sorting them into a single ascending sequence of blip times accompanied by their types (*a* or *b*).
- There is another approach, however, that runs faster and admits a crucial generalization that we will need in Chapter 10. We wish to generate a single list $\{(t_\alpha , s_\alpha)\}$, where t_α are the event times (continuous), and $s\alpha$ are the corresponding event types (discrete).


Negative binomial(n,p)
λ=n(1-p) , n→∞
Poisson(λ)
∑Xi
λ=np , n→∞
n=1
∑Xi
n = 1
Geometric(p)
Min(Xi)
X | ∑Xi
Binomial(n,p)
∑Xi
Bernouli(p)
λ=σ² , n→∞
p=M/N , n→∞
Hypergeometric(M,N,K)
Ln(X)
μ=np , σ²=np(1-p) , n→∞
∏Xi
∑Xi
Lognormal
Exp(x)
Normal(μ,σ²)
α,β→∞
Beta(α,β)
α=β=1
(X-μ)/σ
μ=rλ , σ²=rλ² , r→∞
X1/(X1+X2)
μ+σX
Gamma(r,λ)
Exp(-X/λ)
∑Xi
Normal(0,1)
X1/X2
X1/|X2|
Min(Xi)
r = n/2 , λ=2
r = 1
Exponential(λ)
- λ Ln(X)
Uniform(0,1)
1/X
V→∞
n = 2
Cauchy
V=1
∑Xi
t(v)
λ=2
b=1
X^(1/b)
a+(b-a)X
a=0,b=1
∑Xi
|X|
X1-X2
Chi-squared(n)
Weibull(a,b)
Uniform(a,b)
(X1/V1)/(X2/V2)
X²
V1·X , V2→∞
F(V1,V2)
Double-Exponential(0,λ,λ)

Maximum entropy distributions

1) Max Ent for $x \in [a,b]$ is the uniform

2) Max Ent for $x \in [0,\infty)$ with finite mean μ is exponential

3) Max Ent for $x \in (-\infty,\infty)$ with finite μ, σ^2 is Gaussian

4) Max Ent for $x \in \mathbb{N} \cup \{0\}$ with finite mean is Poisson

5) Max Ent $[0,\infty)$ with $E(\ln x)=\mu$, $\mathrm{Var}(\ln x)=\sigma^2$ → log normal.

Proofs

1) We maximize $S(p) = -\int_a^b p(x) \ln p(x)\, dx$

given $p(x) \geq 0$ and $\int_a^b p(x)\, dx = 1$

→ + Gaussian add $\lambda_2 \left[\int (x-\mu)^2 p(x)\, dx - \sigma^2\right]$

Using Lagrange multipliers, we define the functional

$$J[p(x)] = -\int_a^b p(x) \ln p(x)\, dx + \lambda_0 \left(\int_a^b p(x)\, dx - 1\right)$$

$\frac{\delta J}{\delta p} = 0$ $\quad p \to p + \delta p$, $\ln(p+\delta p) = \ln\left[p\left(1+\frac{\delta p}{p}\right)\right] \approx \ln p + \frac{\delta p}{p}$

$$J[p+\delta p] = -\int_a^b (p+\delta p)\underbrace{\left(\ln p + \frac{\delta p}{p}\right)}\, dx + \lambda_0\left[\int_a^b (p+\delta p)\, dx - 1\right]$$

$$= \underbrace{\left(-\int_a^b p \ln p + \lambda_0\left[\int_a^b p\, dx - 1\right]\right)}_{J[p]} - \int_a^b \delta p \left[\ln p + 1 - \lambda_0\right] dx$$

$\frac{\delta J}{\delta p} = 0 \to \ln p + 1 - \lambda_0 = 0 \to p = e^{\lambda_0 - 1} = \text{const}$ $\quad \int_a^b p(x)\, dx = (b-a)\cdot e^{\lambda-1} = 1 \to p = \frac{1}{b-a}$