LECTURE 2: INTRO TO STATISTICS

Two Schools of Statistics

- Frequentist Goal:

  Construct procedures with frequency guarantees (coverage): if we run a bunch of simulations we are guaranteed they fall within the quoted interval

- Bayesian Goal:

  Describe and update degree of belief in propositions

In this course we will follow Bayesian school of statistics, while also insisting we get frequentist guarantees

But first we must learn about probabilities

- Random Variable: x
- Outcomes: $S \equiv \{x_1, x_2, ...\}$

Discrete or continuous event E ⊂ S has a probability

$$p(E): \quad p_{\rm dice}(1) = \frac{1}{6}$$

$$p(E) \geq 0$$

$$p(A\&B) = p(A) + p(B)$$

$$p(S) = 1: \quad \sum_i p(x_i) = 1$$

Joint probability of x, y : $P(x, y)$

$x \in S_x, \;\; y \in S_y$ not necessarily independent

Marginal Probability $P(x)$:

$$P(x = x_i) = \sum_{y \in S_y} P(x = x_i, y)$$

Conditional Probability :

$$P(x = x_i \mid y = y_j) = \frac{P(x = x_i, y = y_j)}{P(y = y_j)}$$

Marginal

Prob. of $x = x_i$ given $y = y_j$

Independence:

x and y are independent if and only if $P(x, y) = P(x)P(y)$

Product rule (Chain Rule, giving conditional) :

$$P(x,y\,|\,\mathcal{H}) = P(x\,|\,y,\mathcal{H})P(y\,|\,\mathcal{H}) = P(y\,|\,x,\mathcal{H})P(x\,|\,\mathcal{H}).$$

where $\mathcal{H}$: Generative Model

Sum rule (giving marginal) :

$$\begin{aligned} P(x\,|\,\mathcal{H}) &= \sum_y P(x,y\,|\,\mathcal{H}) \\ &= \sum_y P(x\,|\,y,\mathcal{H})P(y\,|\,\mathcal{H}). \end{aligned}$$

Bayes Theorem :

$$\begin{aligned} P(y\,|\,x,\mathcal{H}) &= \frac{P(x\,|\,y,\mathcal{H})P(y\,|\,\mathcal{H})}{P(x\,|\,\mathcal{H})} \\ &= \frac{P(x\,|\,y,\mathcal{H})P(y\,|\,\mathcal{H})}{\sum_{y'} P(x\,|\,y',\mathcal{H})P(y'\,|\,\mathcal{H})}. \end{aligned}$$

Bayes Theorem
≠ Bayesian statistics

Example 2.3. Jo has a test for a nasty disease. We denote Jo's state of health by the variable a and the test result by b.

$$\begin{array}{ll} a = 1 & \text{Jo has the disease} \\ a = 0 & \text{Jo does not have the disease.} \end{array} \tag{2.12}$$

The result of the test is either 'positive' ($b = 1$) or 'negative' ($b = 0$); the test is 95% reliable: in 95% of cases of people who really have the disease, a positive result is returned, and in 95% of cases of people who do not have the disease, a negative result is obtained. The final piece of background information is that 1% of people of Jo's age and background have the disease.

OK – Jo has the test, and the result is positive. What is the probability that Jo has the disease?

From Information Theory, Inference and Learning Algorithms, by David MacKay

Step 1: Write down all probabilities

We are given conditional probabilities

$$P(b=1\,|\,a=1) = 0.95 \qquad P(b=1\,|\,a=0) = 0.05$$
$$P(b=0\,|\,a=1) = 0.05 \qquad P(b=0\,|\,a=0) = 0.95;$$

And marginal probability of a

$$P(a=1) = 0.01 \qquad P(a=0) = 0.99.$$

We want $P(a=1|b=1)$

Step 2: Deduce joint probability $P(a, b)$

$$P(a,b)=P(a|b)P(b)=P(b|a)P(a)$$

Step 3:

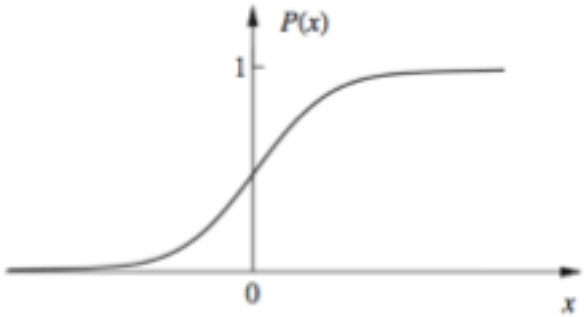
$$P(a=1\,|\,b=1) = \frac{P(b=1\,|\,a=1)P(a=1)}{P(b=1\,|\,a=1)P(a=1) + P(b=1\,|\,a=0)P(a=0)}$$
$$= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99}$$
$$= 0.16.$$

Lots of false positives!

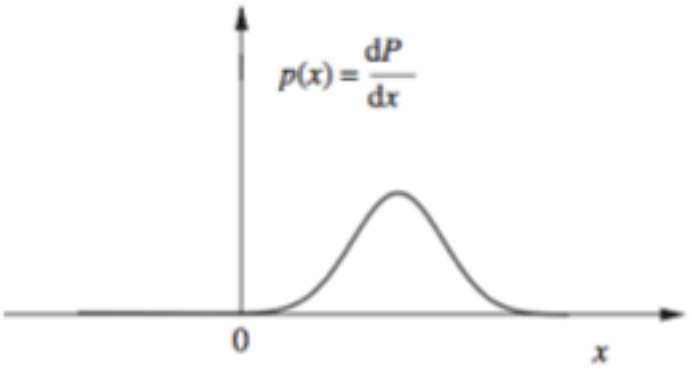
Continuous Variables

- Cumulative probability function (CPF)

$$P(x) = \int_{-\infty}^{x} p(x')dx', P(-\infty) = 0, \; P(+\infty) = 1$$



- Probability density function (PDF) $p(x) = \frac{dP(x)}{dx}$ has dimensions of $[x]^{-1}$



- Expectation value $\langle F(x)\rangle = \int_{-\infty}^{\infty} \mathrm{d}x\ p(x)F(x).$
- Moments $m_n \equiv \langle x^n \rangle = \int \mathrm{d}x p(x) x^n.$
- Characteristic function generates moments:

$$\tilde{p}(k) = \left\langle \mathrm{e}^{-\mathrm{i}kx} \right\rangle = \int \mathrm{d}x p(x)\ \mathrm{e}^{-\mathrm{i}kx}.$$

(Fourier Transform of the PDF)

Recovering the PDF from the characteristic function through the inverse F.T.:

$$p(x) = \frac{1}{2\pi} \int \mathrm{d}k \tilde{p}(k)\, \mathrm{e}^{+\mathrm{i}kx}.$$

Moments of the distribution:

$$\tilde{p}(k) = \left\langle \sum_{n=0}^{\infty} \frac{(-\mathrm{i}k)^n}{n!} x^n \right\rangle = \sum_{n=0}^{\infty} \frac{(-\mathrm{i}k)^n}{n!} \langle x^n \rangle .$$

PDF moments around $x0$:

$$e^{ikx_0}\tilde{p}(k) = \left\langle e^{-ik(x-x_0)} \right\rangle = \sum_{n=0}^{\infty} \frac{(-ik)^n}{n!} \langle (x-x_0)^n \rangle .$$

- Cumulant generating function defines cumulants $\langle x^n \rangle_c$

$$\ln \tilde{p}(k) = \sum_{n=1}^{\infty} \frac{(-ik)^n}{n!} \langle x^n \rangle_c .$$

We can obtain relations between moments and cumulants using

$$\ln(1+\epsilon) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{\epsilon^n}{n}.$$

We expand ε into a sum of moments and ln(1+ε) into a sum of cumulants and match powers of k to obtain

Mean $\langle x \rangle_c = \langle x \rangle ,$

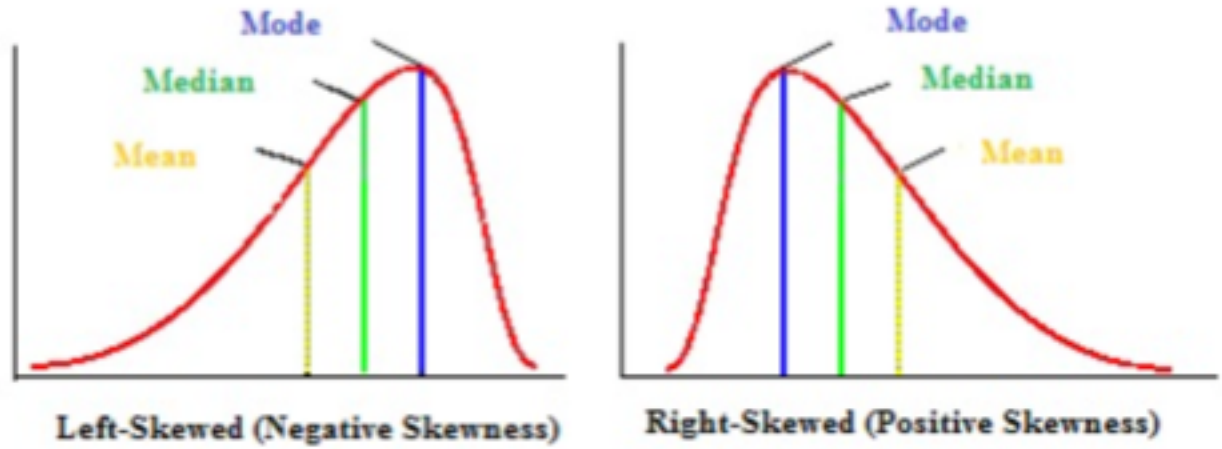
Variance $\langle x^2 \rangle_c = \langle x^2 \rangle - \langle x \rangle^2 ,$

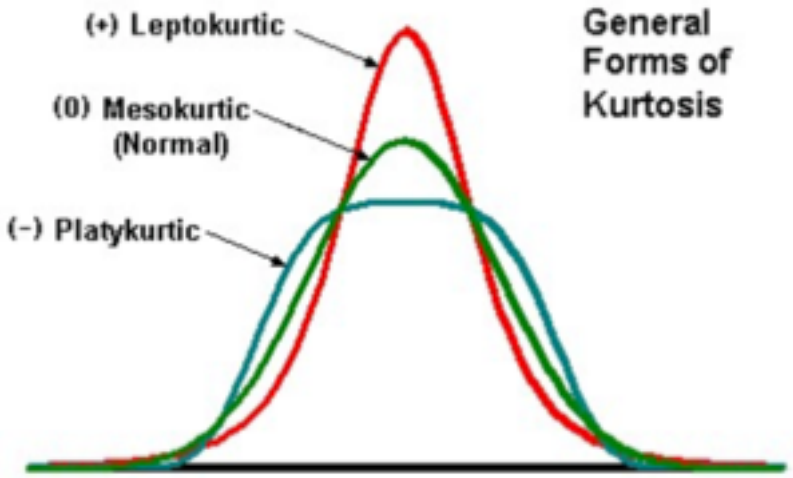
Skewness $\langle x^3 \rangle_c = \langle x^3 \rangle - 3\langle x^2 \rangle \langle x \rangle + 2\langle x \rangle^3 ,$

Curtosis $\langle x^4 \rangle_c = \langle x^4 \rangle - 4\langle x^3 \rangle \langle x \rangle - 3\langle x^2 \rangle^2 + 12\langle x^2 \rangle \langle x \rangle^2 - 6\langle x \rangle^4 .$

Mean, mode, median, skewness, curtosis

- Different ways to characterize the probability distribution function
- Variance describes the width





Moments as connected clusters of cumulants: useful pictorial that relates cumulants to moments. Cumulants are also called connected moments

$$\langle x \rangle = \bullet$$

$$\langle x \rangle = \langle x \rangle_c$$

$$\langle x^2 \rangle = \; :\!\bullet\bullet\!: \; + \; \bullet\bullet$$

$$\langle x^2 \rangle = \langle x^2 \rangle_c + \langle x \rangle_c^2$$

$$\langle x^3 \rangle = \; :\!\bullet\!\bullet\!: \; + 3 \; :\!\bullet\bullet\!: \; + \; \bullet\!\bullet$$

$$\langle x^3 \rangle = \langle x^3 \rangle_c + 3 \langle x^2 \rangle_c \langle x \rangle_c + \langle x \rangle_c^3$$

$$\langle x^4 \rangle = \; :\!\bullet\bullet\!: \; + 4 \; :\!\bullet\!\bullet\!: \; + 3 \; :\!\bullet\bullet\!: \; + 6 \; :\!\bullet\bullet\!: \; + \; \bullet\bullet\bullet\bullet$$

$$\langle x^4 \rangle = \langle x^4 \rangle_c + 4 \langle x^3 \rangle_c \langle x \rangle_c + 3 \langle x^2 \rangle_c^2 + 6 \langle x^2 \rangle_c \langle x \rangle_c^2 + \langle x \rangle_c^4$$

Many Random Variables

- Joint PDF $p(\mathbf{x})$ where $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$

$$p_{\mathbf{x}}(\mathcal{S}) = \int \mathrm{d}^N \mathbf{x} p(\mathbf{x}) = 1.$$ (*S* is the set of all outcomes)

- Joint characteristic function $\tilde{p}(\mathbf{k}) = \left\langle \exp\left(-\mathrm{i}\sum_{j=1}^{N} k_j x_j\right)\right\rangle$

Ex)

$$\langle x_1 x_2 \rangle = \overset{1}{\bullet}\ \overset{2}{\bullet} + \overset{1}{\circ}\ \overset{2}{\circ} \qquad \langle x_1 x_2 \rangle = \langle x_1 \rangle_c \langle x_2 \rangle_c + \langle x_1 * x_2 \rangle_c$$

$$\langle x_1^2 x_2 \rangle = \overset{1}{\bullet}\ \overset{1}{\bullet}\ \overset{2}{\bullet} + \overset{1}{\circ}\ \overset{1}{\circ}\ \overset{2}{\bullet} + 2\ \overset{1}{\circ}\ \overset{2}{\circ} + \overset{1}{\circ}\ \overset{1}{\circ}\ \overset{2}{\circ}$$

$$\langle x_1^2 x_2 \rangle = \langle x_1 \rangle_c^2 \langle x_2 \rangle_c + \langle x_1^2 \rangle_c \langle x_2 \rangle_c + 2 \langle x_1 * x_2 \rangle_c \langle x_1 \rangle_c + \langle x_1^2 * x_2 \rangle_c$$

$\langle x_\alpha * x_\beta \rangle_c$ is zero if x_α and x_β are independent (symbol * here means product)

Normal or Gaussian Distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\lambda)^2}{2\sigma^2}\right], \quad -\infty < x < \infty$$

- Characteristic Function

$$\tilde{p}(k) = \int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\lambda)^2}{2\sigma^2} - ikx\right] = \exp\left[-ik\lambda - \frac{k^2\sigma^2}{2}\right]$$

- Cumulants

From $\ln \tilde{p}(k) = -ik\lambda - k^2\sigma^2/2$, we can identify:

$$\langle x \rangle_c = \lambda, \quad \langle x^2 \rangle_c = \sigma^2, \quad \langle x^3 \rangle_c = \langle x^4 \rangle_c = \cdots = 0.$$

- Moments from cluster expansion

$$\langle x \rangle = \lambda, \qquad \langle x^3 \rangle = 3\sigma^2\lambda + \lambda^3,$$
$$\langle x^2 \rangle = \sigma^2 + \lambda^2, \qquad \langle x^4 \rangle = 3\sigma^4 + 6\sigma^2\lambda^2 + \lambda^4,$$
$$\cdots$$

Multi-variate Gaussian

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N \det[C]}} \exp\left[-\frac{1}{2}\sum_{mn}(C^{-1})_{mn}(x_m - \lambda_m)(x_n - \lambda_n)\right]$$

where C is the covariance matrix, and C^{-1} is its inverse

$$\tilde{p}(\mathbf{k}) = \exp\left[-ik_m\lambda_m - \frac{1}{2}C_{mn}k_mk_n\right]$$

Note: repeated indices are summed

- Cumulants

$$\langle x_m \rangle_c = \lambda_m, \quad \langle x_m * x_n \rangle_c = C_{mn}, \text{ rest are zero}$$

Wick's Theorem: 4 point function for a multi-variate gaussian with zero mean depends only on products of 2 point functions

$$\langle x_a x_b x_c x_d \rangle = C_{ab}C_{cd} + C_{ac}C_{bd} + C_{ad}C_{bc}$$

Sum of variables

$$X = \sum_{i=1}^{N} x_i$$

- PDF $$p_X(x) = \int \mathrm{d}^N \mathbf{x} p(\mathbf{x}) \delta\left(x - \sum x_i\right)$$
- Characteristic function

$$\tilde{p}_X(k) = \left\langle \exp\left(-\mathrm{i}k \sum_{j=1}^{N} x_j\right)\right\rangle = \tilde{p}\left(k_1 = k_2 = \cdots = k_N = k\right)$$

- Cumulants

$$\ln \tilde{p}\left(k_1 = k_2 = \cdots = k_N = k\right) = -\mathrm{i}k \sum_{i_1=1}^{N} \left\langle x_{i_1}\right\rangle_c + \frac{(-\mathrm{i}k)^2}{2} \sum_{i_1,i_2}^{N} \left\langle x_{i_1} x_{i_2}\right\rangle_c + \cdots$$

$$\left\langle X\right\rangle_c = \sum_{i=1}^{N} \left\langle x_i\right\rangle_c, \quad \left\langle X^2\right\rangle_c = \sum_{i,j}^{N} \left\langle x_i x_j\right\rangle_c, \cdots$$

If variables are independent cross-cumulants vanish ->

$$\left\langle X^n\right\rangle_c = \sum_{i=1}^{N} \left\langle x_i^n\right\rangle_c \quad \Rightarrow \quad \left\langle X^n\right\rangle_c = N \left\langle x^n\right\rangle_c$$

If all drawn from p(x)

Central Limit Theorem

For large N $< x > \propto N,\ \langle (x - \langle x \rangle)^2 \rangle \propto N$

$$y = \frac{x - N\langle x \rangle_c}{\sqrt{N}},\ \langle y^n \rangle_c \propto N^{1-n/2}$$

$$N \to \infty,\ \langle y^n \rangle_c \to 0 \text{ for } n > 2$$

$$\lim_{N\to\infty} p\left(y = \frac{\sum_{i=1}^{N} x_i - N\langle x \rangle_c}{\sqrt{N}}\right) = \frac{1}{\sqrt{2\pi \langle x^2 \rangle_c}} \exp\left(-\frac{y^2}{2\langle x^2 \rangle_c}\right)$$

Gaussian Distribution

We have assumed cumulants of x are finite

Distribution of sum is Gaussian even if $p(x)$ very non-Gaussian

▷ Exercise 3.1.[2, p.59] A die is selected at random from two twenty-faced dice on which the symbols 1–10 are written with nonuniform frequency as follows.

Symbol	1	2	3	4	5	6	7	8	9	10
Number of faces of die A	6	4	3	2	1	1	1	1	1	0
Number of faces of die B	3	3	2	2	2	2	2	2	1	1

The randomly chosen die is rolled 7 times, with the following outcomes:

5, 3, 9, 3, 8, 4, 7.

What is the probability that the die is die A?

▷ Exercise 3.2.[2, p.59] Assume that there is a third twenty-faced die, die C, on which the symbols 1–20 are written once each. As above, one of the three dice is selected at random and rolled 7 times, giving the outcomes: 3, 5, 4, 8, 3, 9, 7.
What is the probability that the die is (a) die A, (b) die B, (c) die C?

Solution to 3.1

$$\frac{P(A\,|\,D)}{P(B\,|\,D)} = \frac{1}{2}\frac{3}{2}\frac{1}{1}\frac{3}{2}\frac{1}{2}\frac{2}{2}\frac{1}{2} = \frac{9}{32}$$

Solution to 3.2

$$P(D\,|\,A) = \frac{3}{20}\frac{1}{20}\frac{2}{20}\frac{1}{20}\frac{3}{20}\frac{1}{20}\frac{1}{20} = \frac{18}{20^7};$$

$$P(D\,|\,B) = \frac{2}{20}\frac{2}{20}\frac{2}{20}\frac{2}{20}\frac{2}{20}\frac{1}{20}\frac{2}{20} = \frac{64}{20^7};$$

$$P(D\,|\,C) = \frac{1}{20}\frac{1}{20}\frac{1}{20}\frac{1}{20}\frac{1}{20}\frac{1}{20}\frac{1}{20} = \frac{1}{20^7}.$$

$$P(A\,|\,D) = \frac{18}{18+64+1} = \frac{18}{83}; \qquad P(B\,|\,D) = \frac{64}{83}; \qquad P(C\,|\,D) = \frac{1}{83}.$$

Binomial Distribution

- Two outcomes A & B, $p_A = 1 - p_B$
- N trials, probability distribution: $p_N(N_A) = \binom{N}{N_A} p_A^{N_A} p_B^{N-N_A}$

where $\binom{N}{N_A} = \frac{N!}{N_A!(N-N_A)!}$ is the # of possible orderings of N_A in N

$$(p_A + p_B)^N = \sum \binom{N}{N_k} p_k^{N_A} p_B^{N-N_A}$$

- Use Stirling's approximation: $x! \approx x^x e^{-x} \sqrt{2\pi x}$

Multinomial Distribution

$$\{A, B, ..., M\},\ \{p_A, p_B, ..., p_M\},\ \sum_i p_i = 1$$

$$p_N(\{N_A, N_B, \cdots, N_M\}) = \frac{N!}{N_A! N_B! \cdots N_M!} p_A^{N_A} p_B^{N_B} \cdots p_M^{N_M}$$

- Characteristic Function

$$\tilde{p}_N(k) = \left\langle e^{-ikN_A} \right\rangle = \sum_{N_A=0}^{N} \frac{N!}{N_A!(N-N_A)!} p_A^{N_A} p_B^{N-N_A} e^{-ikN_A} = \left(p_A e^{-ik} + p_B\right)^N$$

For 1 trial allowed N_A are (0,1) with p_B and p_A (B=0, A=1)

- Cumulant

Cumulant generating function: N times cumulant for 1 trial

$$\ln \tilde{p}_N(k) = N \ln\left(p_A e^{-ik} + p_B\right) = N \ln \tilde{p}_1(k)$$

For 1 trial it is easier to compute moments

$$\left\langle N_A^\ell \right\rangle = p_A, \text{ for all } \ell$$

After N trials: we convert moments of 1 trial to cumulants of 1 trial and multiply by N to get cumulants of N trials. For l=1,2:

$$\langle N_A \rangle_c = N p_A$$

$$\left\langle N_A^2 \right\rangle_c = N\left(p_A - p_A^2\right) = N p_A p_B \qquad \left\langle x^2 \right\rangle_c = \left\langle x^2 \right\rangle - \langle x \rangle^2,$$

$$\sigma \propto \sqrt{N}$$

Poisson Distribution

Example: Radioactive decay

- Probability of one and only one event (decay) in [*t*, *t*+*dt*] is proportional to *dt* as *dt* -> 0.
- Probabilities of events are independent.

Poisson $p(M|T)$ M events in time interval T

- Limit of binomial: $N = \frac{T}{dt} >> 1$

$$p_1 = \alpha dt,\ p_0 = 1 - \alpha dt,\ p_2 \propto dt^2 \to 0$$

- We imagine many binomial events over time interval T
- Characteristic function: we use binomial and send N to infinity

$$\tilde{p}(k) = \left(pe^{-ik} + q\right)^n = \lim_{dt \to 0} \left[1 + \alpha dt \left(e^{-ik} - 1\right)\right]^{T/dt} = \exp\left[\alpha(e^{-ik} - 1)T\right]$$

- Do the inverse F.T. and get the Poisson PDF:

$$p(x) = \int_{-\infty}^{\infty} \frac{\mathrm{d}k}{2\pi} \exp\left[\alpha\left(\mathrm{e}^{-\mathrm{i}k} - 1\right)T + \mathrm{i}kx\right]$$

$$= \mathrm{e}^{-\alpha T} \int_{-\infty}^{\infty} \frac{\mathrm{d}k}{2\pi} \mathrm{e}^{\mathrm{i}kx} \sum_{M=0}^{\infty} \frac{(\alpha T)^M}{M!} \mathrm{e}^{-\mathrm{i}kM}$$

$$e^x = \sum_{M=0}^{\infty} \frac{x^M}{M!}$$

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}k}{2\pi} \mathrm{e}^{\mathrm{i}k(x-M)} = \delta(x - M)$$

$$p_{\alpha T}(x) = \sum_{M=0}^{\infty} \mathrm{e}^{-\alpha T} \frac{(\alpha T)^M}{M!} \delta(x - M)$$

$$p_{\alpha T}(M) = \mathrm{e}^{-\alpha T} (\alpha T)^M / M!$$

- Cumulants:

$$\ln \tilde{p}_{\alpha T}(k) = \alpha T(\mathrm{e}^{-\mathrm{i}k} - 1) = \alpha T \sum_{n=1}^{\infty} \frac{(-\mathrm{i}k)^n}{n!}, \quad \Longrightarrow \quad \langle M^n \rangle_c = \alpha T.$$

All cumulants are the same.

Moments:

$$\langle M \rangle = (\alpha T), \quad \langle M^2 \rangle = (\alpha T)^2 + (\alpha T),$$

$$\langle M^3 \rangle = (\alpha T)^3 + 3(\alpha T)^2 + (\alpha T)$$

Example: Assume stars randomly distributed around us with density n, what is probability that the nearest star is at distance R ?

All cumulants are the same. We can use cluster expansion to obtain Moments:

$$\langle M \rangle = (\alpha T), \quad \langle M^2 \rangle = (\alpha T)^2 + (\alpha T),$$

$$\langle M^3 \rangle = (\alpha T)^3 + 3(\alpha T)^2 + (\alpha T)$$

Example: Assume stars randomly distributed around us with density n, what is probability that the nearest star is at distance R ?

$$p = ndV, \text{ Poisson with } \alpha = n$$

$$p(R) = p_{nV}(0, V) \cdot p_{nV}(1, dV)$$

$$V = \frac{4\pi}{3} R^3 \qquad dV = 4\pi R^2 dR$$

$$p_{nV}(0) = e^{-\frac{4\pi}{3} R^3 n}$$

$$p_{nV}(1) = 4\pi n R^2 \cdot e^{-4\pi n R^2 dR} \cdot dR$$

$$p(R)dR = 4\pi n R^2 \cdot e^{-\frac{4\pi}{3} R^3 n} \cdot dR$$

Forward Probability

- Generative model describing a process giving rise to some data

Exercise 2.4.[2, p.40] An urn contains K balls, of which B are black and $W = K - B$ are white. Fred draws a ball at random from the urn and replaces it, N times.

(a) What is the probability distribution of the number of times a black ball is drawn, n_B?

(b) What is the expectation of n_B? What is the variance of n_B? What is the standard deviation of n_B? Give numerical answers for the

Solution : $f_B \equiv B/K$

a)

$$P(n_B \mid f_B, N) = \binom{N}{n_B} f_B^{n_B} (1 - f_B)^{N - n_B}$$

b)

$$\mathcal{E}[n_B] = N f_B$$

$$\mathrm{var}[n_B] = N f_B (1 - f_B).$$

NOTE: No Bayes Theorem used

Inverse Probability

- We compute probability of some unobserved quantity, given the observed variables. Use Bayes theorem

Example 2.6. There are eleven urns labelled by $u \in \{0, 1, 2, \ldots, 10\}$, each containing ten balls. Urn u contains u black balls and $10 - u$ white balls. Fred selects an urn u at random and draws N times with replacement from that urn, obtaining n_B blacks and $N - n_B$ whites. Fred's friend, Bill, looks on. If after $N = 10$ draws $n_B = 3$ blacks have been drawn, what is the probability that the urn Fred is using is urn u, from Bill's point of view? (Bill doesn't know the value of u.)

Solution. The joint probability distribution of the random variables u and n_B can be written

$$P(u, n_B \,|\, N) = P(n_B \,|\, u, N) P(u). \tag{2.20}$$

From the joint probability of u and n_B, we can obtain the conditional distribution of u given n_B:

$$P(u \,|\, n_B, N) = \frac{P(u, n_B \,|\, N)}{P(n_B \,|\, N)} \tag{2.21}$$

$$= \frac{P(n_B \,|\, u, N) P(u)}{P(n_B \,|\, N)}. \tag{2.22}$$

The marginal probability of u is $P(u) = \frac{1}{11}$ for all u. You wrote down the probability of n_B given u and N, $P(n_B \,|\, u, N)$, when you solved exercise 2.4 (p.27). [You *are* doing the highly recommended exercises, aren't you?] If we define $f_u \equiv u/10$ then

$$P(n_B \,|\, u, N) = \binom{N}{n_B} f_u^{n_B} (1 - f_u)^{N - n_B}. \tag{2.23}$$

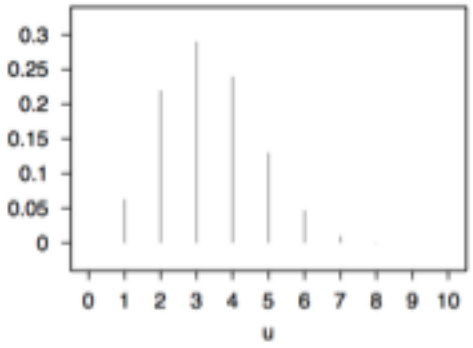
What about the denominator, $P(n_B \,|\, N)$? This is the marginal probability of n_B, which we can obtain using the sum rule:

$$P(n_B \,|\, N) = \sum_u P(u, n_B \,|\, N) = \sum_u P(u) P(n_B \,|\, u, N). \tag{2.24}$$

So the conditional probability of u given n_B is

$$P(u \,|\, n_B, N) = \frac{P(u) P(n_B \,|\, u, N)}{P(n_B \,|\, N)} \tag{2.25}$$

$$= \frac{1}{P(n_B \,|\, N)} \frac{1}{11} \binom{N}{n_B} f_u^{n_B} (1 - f_u)^{N - n_B}. \tag{2.26}$$



u	$P(u \,\|\, n_B = 3, N)$
0	0
1	0.063
2	0.22
3	0.29
4	0.24
5	0.13
6	0.047
7	0.0099
8	0.00086
9	0.0000096
10	0

$$P(n_B = 3 \,|\, N = 10) = 0.083.$$

Example 2.6 (continued). Assuming again that Bill has observed $n_B = 3$ blacks in $N = 10$ draws, let Fred draw another ball from the same urn. What is the probability that the next drawn ball is a black? [You should make use of the posterior probabilities in figure 2.6.]

Solution. By the sum rule,

$$P(\text{ball}_{N+1} \text{ is black} \,|\, n_B, N) = \sum_u P(\text{ball}_{N+1} \text{ is black} \,|\, u, n_B, N) P(u \,|\, n_B, N). \tag{2.29}$$

Since the balls are drawn with replacement from the chosen urn, the probability $P(\text{ball}_{N+1} \text{ is black} \,|\, u, n_B, N)$ is just $f_u = u/10$, whatever n_B and N are. So

$$P(\text{ball}_{N+1} \text{ is black} \,|\, n_B, N) = \sum_u f_u P(u \,|\, n_B, N). \tag{2.30}$$

Using the values of $P(u \,|\, n_B, N)$ given in figure 2.6 we obtain

$$P(\text{ball}_{N+1} \text{ is black} \,|\, n_B = 3, N = 10) = 0.333. \quad \square \tag{2.31}$$

Note: we have marginalized over all u, instead of evaluating at the best value of u

From Inverse Probability to Bayesian Inference

Example 2.7. Bill tosses a bent coin N times, obtaining a sequence of heads and tails. We assume that the coin has a probability f_H of coming up heads; we do not know f_H. If n_H heads have occurred in N tosses, what is the probability distribution of f_H? (For example, N might be 10, and n_H might be 3; or, after a lot more tossing, we might have $N = 300$ and $n_H = 29$.) What is the probability that the $N+1$th outcome will be a head, given n_H heads in N tosses?

- What is the difference between this problem and previous one?
- Before urn u was a random variable. Here coin bias f_H has a fixed, but unknown value.
- Before we were given $P(u)$, now we have to decide on $P(f_H)$: subjective prior
- If we choose uniform $P(f_H)$ and N=10, n_H=3 then we get the same answer as on slides 26/27.

The Meaning of Probability

1) Frequency of outcomes for repeated random experiments

2) Degrees of belief in propositions not involving random variables (quantifying uncertainty)

Example: What is probability that Mr. S killed Mrs. S given the evidence?

- He either was or was not the killer,
  but we can describe how probable it was.

This is Bayesian viewpoint: subjective interpretation of probability, since it depends on assumptions

We are imagining we are assigning probability to which "urn" it was drawn from even though there is only one urn

To be able to do so we also need to decide on imaginary prior probability distribution of different urns

- This is not universally accepted:
  20th century statistics dominated by frequentists (classical statistics).

- Main difference:
  Bayesians use probabilities to describe inferences

- It does not mean they view propositions (or hypotheses)
  as stochastic superposition of states

- There is only one true value and Bayesians use probabilities
  to describe beliefs about mutually exclusive hypotheses

- Ultimate proof of validity is its success in practical applications.
  Typically as good as the best classical method.

- Degrees of belief can be mapped onto probabilities (Cox's Axioms)

Let's apply Bayes Theorem to parameter testing:
A family of λ parameters we'd like to test

➢ We have data D and hypothesis space H

$$p(\lambda, D|\mathcal{H}) = p(\lambda|D,\mathcal{H})p(D|\mathcal{H}) = p(D|\lambda,\mathcal{H})p(\lambda|\mathcal{H})$$

$$p(\lambda|D,\mathcal{H}) = \frac{p(D|\lambda,\mathcal{H})p(\lambda|\mathcal{H})}{p(D|\mathcal{H})}$$

Posterior of λ

$P(D|\lambda,H)$: Likelihood of λ at fixed D, probability of D at fixed λ
$P(\lambda|H)$: Prior on λ
$P(D|H)$: Marginal or Evidence or Bayes factor
$P(\lambda|D,H)$: Posterior on λ

$$\text{Posterior} = \frac{\text{Likelihood x Prior}}{\text{Evidence}}$$

We can also apply it to families of hypotheses $\mathcal{H}$

$$p(\mathcal{H}|D,I) = \frac{p(D|\mathcal{H},I)p(\mathcal{H}|I)}{p(D|I)}$$

where I : space of all possible families

Once we have made the subjective assumption on prior $P(H \mid I)$ the inferences are unique

$$p(\lambda|D,\mathcal{H}) \propto p(D|\lambda,\mathcal{H}) \quad \text{if } p(\lambda) \propto \text{constant}$$

Uniform prior

1) Normalization $p(D|\mathcal{H})$ is needed and is the source of difficulty

2) Uniform prior not invariant under re-parametrization

$$\lambda \to F(\lambda) \qquad p(\lambda)d\lambda = p(F)dF$$

$$p(F) = p(\lambda)\left|\frac{d\lambda}{dF}\right|$$

Not uniform in general

Jacobian

- Priors are subjective, no inference is possible without assumptions

Non-informative priors try to be as agnostic as possible

How to choose informative priors?

- Conjugate prior: when posterior takes the same form as the prior it is conjugate to likelihood
- Example: beta distribution is conjugate to binomial (HW 2)
- Can be interpreted as additional data

- For Gaussian with known $\boldsymbol{\sigma}$: $p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)$
- Posterior: $p(\theta|y) \propto \exp\left(-\frac{1}{2}\left(\frac{(y-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right)$
- Completing the square: $p(\theta|y) \propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)$

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2}+\frac{1}{\sigma^2}$$

Non-informative priors

- If we take τ_0 to infinity we get non-informative prior
- No prior is truly non-informative, because the transformation of variable changes it
- Priors can be improper: do not integrate to 1. But posteriors must be proper (this must be checked)
- Jeffrey's prior based on minimal Fisher information matrix (to be discussed later): not a universal recipe

Non-informative priors

- Pivotal quantity has distribution independent of y and parameter λ: if this is x-λ then this is a location parameter: uniform prior. E.g. mean of a gaussian

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\lambda)^2}{2\sigma^2}\right], \quad -\infty < x < \infty$$

- Scale parameter: pivotal in x/λ. This leads to uniform prior in log λ. E.g. variance of a gaussian ($\lambda=\sigma$)
- Prior is rarely an issue in 1-d: either the data are good in which case prior does not matter or are not (so get more data!)
- Priors can become problematic in many dimensions, especially if we have more parameters than needed by the data: posteriors can be a projection of multi-dimensional priors without us knowing it: care must be taken to avoid this (we will discuss further throughout the course)

The Likelihood Principle

- Given generative model $\mathcal{H}$ for data D and model parameter λ, having observed D_1, all inferences should depend only on $p(D_1|\lambda,\mathcal{H})$

- Often violated in classical statistics (e.g. p-value)

- Built into Bayesian statistics

Alternative to Bayesian Statistics: Frequentist Statistics

Goal: Construct procedure with frequency guarantees
e.g. confidence interval with coverage

Coverage: An interval has coverage of 1-α if in the long run of experiments α fraction of true values falls out of the interval
(type I error, "false positive",
false rejection of a true null hypothesis)

Important: α has to be fixed ahead of time, cannot be varied
(Neyman-Pearson hypothesis testing also involves alternative hypothesis and reports type II error β, "false negative",
i.e. rate of retaining a false null hypothesis)

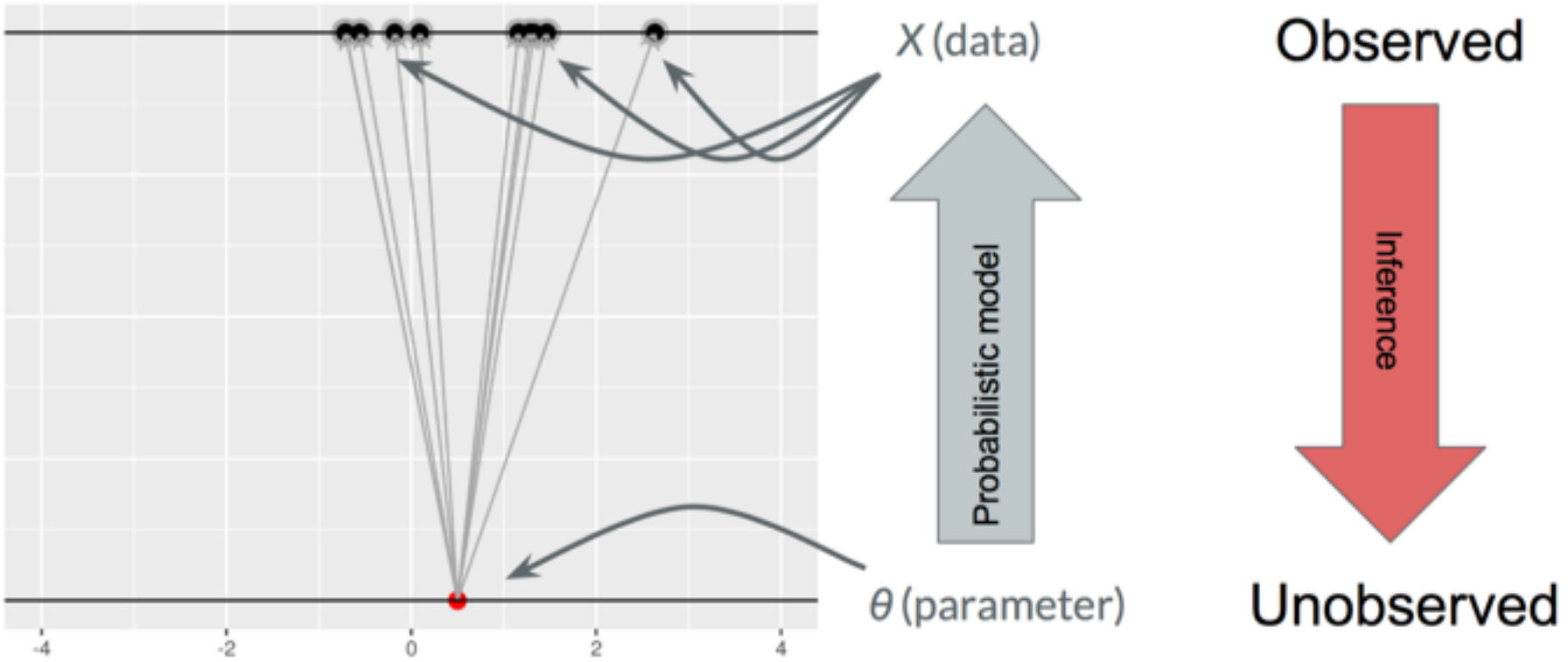
This guarantee of coverage even in the worst case is appealing, but comes at a price

Frequentists	Bayesians
Data are repeatable random sample, underlying parameters are unchanged	Data are observed from realized sample, parameters are unknown and described probabilistically
Parameters are fixed	Data are fixed
Studies (data) are repeatable	Studies are fixed
95% confidence intervals $\alpha = 0.05$ If $p(\text{data}\|\mathcal{H}_0) > \alpha$ accept, otherwise reject	Induction from posterior $p(\boldsymbol{\theta}\|\text{data})$ $p(\mathcal{H}_0\|\text{data})$: e.g. 95% credible intervals of posterior cover 95% of total posterior "mass"
Repeatability key, no use of prior information, alternative hypotheses yes (Neyman-Pearson school)	Assumptions are key element of inference, inference is always subjective, we should embrace it

Bayesian vs frequentist simulations

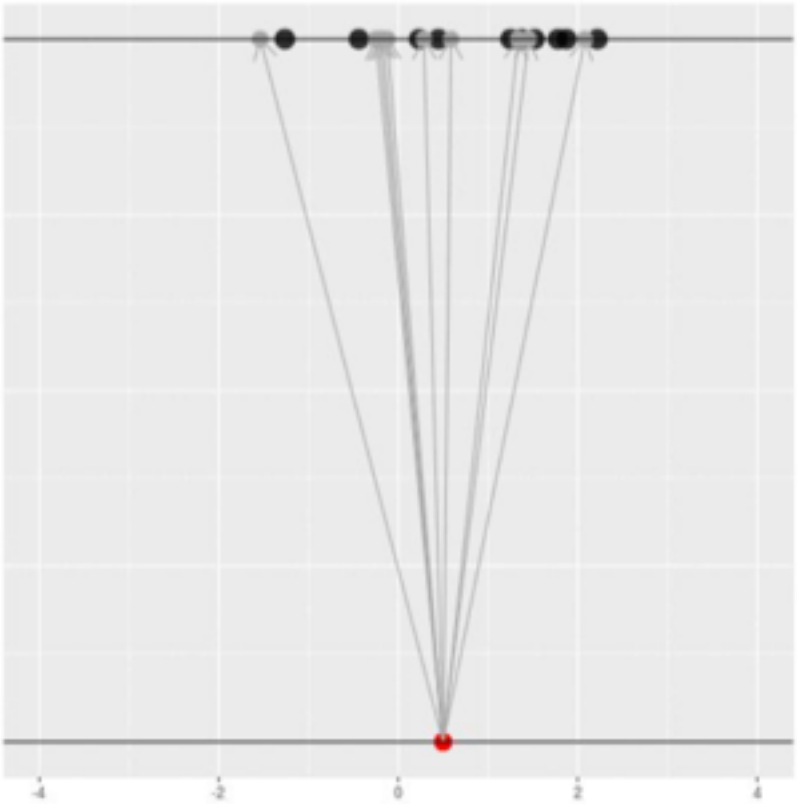
- Frequentist approach: choose a summary statistic. Often, but not always, this is the maximum likelihood estimator, corrected if needed for proper frequentist guarantees (for example, Bessel correction for variance). Apply it to the data, get an estimate of the parameter. Use the parameter to generate mock simulated data generated from that parameter. Use their distribution to determine the errors and confidence intervals.
- Alternative: use data subsamples in place of simulations (bootstrap)
- Bayesian approach: generate random parameters realizations consistent with the prior and the likelihood. The result is a posterior distribution. We will look at these sampling ("Monte Carlo") approaches later in the course.

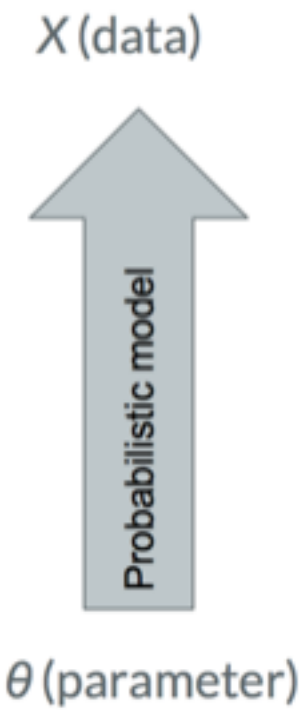
Parameters and data



Slides credit: Ryan Giordano

Frequentist approach



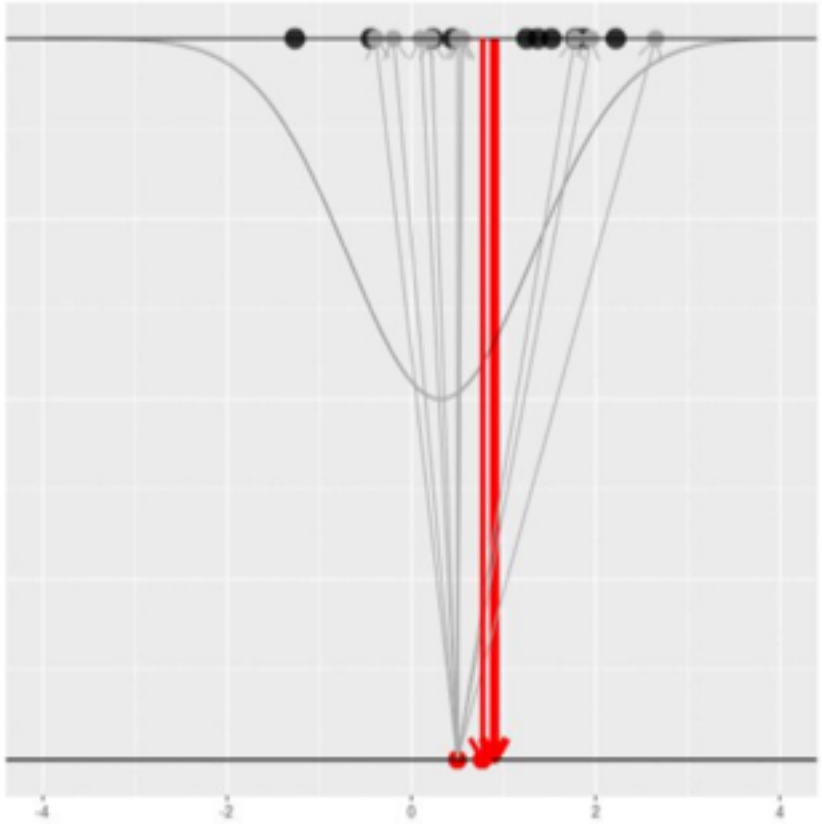


Frequentist idea:

We got the parameter indicated by the red dot and saw the dataset in black.

But the same parameter could have given us lots of other datasets.

Frequentist approach



X (data)

Probabilistic model

θ (parameter)

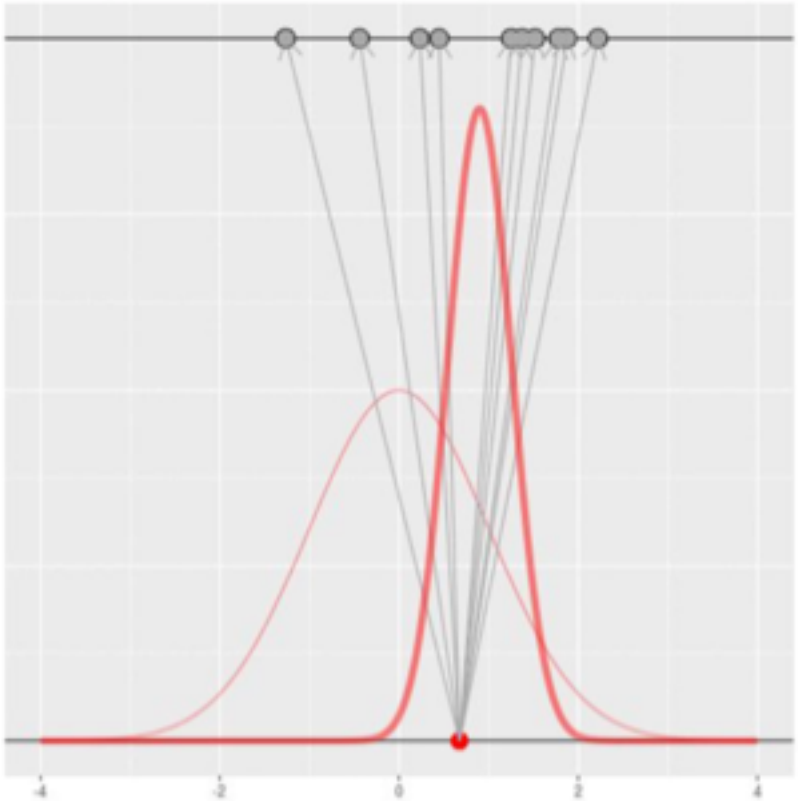
Frequentist idea:

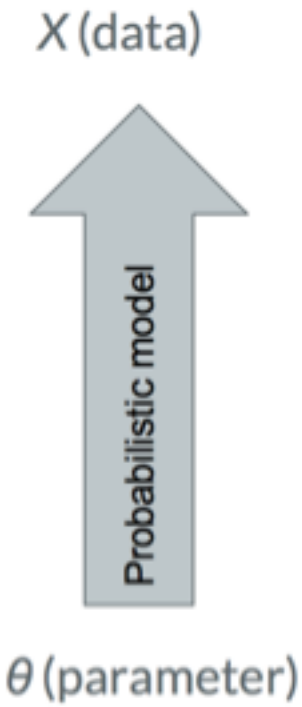
For each dataset, we might pick some summary function and call it an "estimate".
It will be different each time because the data will be different each time.

A typical estimate -- but not the only one -- is the value that maximizes the likelihood of the data.

We hope the estimate is usually near the true parameter in some sense.

Bayesian approach





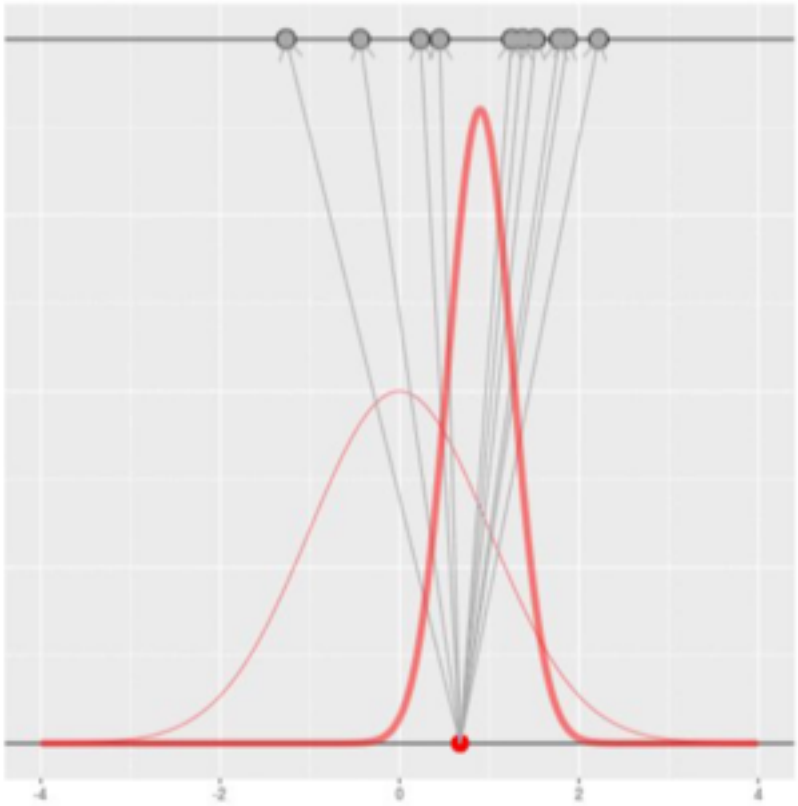
Bayesian idea:

Suppose we draw a bunch of parameters and datasets, and then throw out every pair where the data doesn't match what we observed.

The distribution of the parameters that are left represents which parameters could have given us the dataset we saw.

We hope the prior is reasonable and the model is accurate.

Bayesian approach



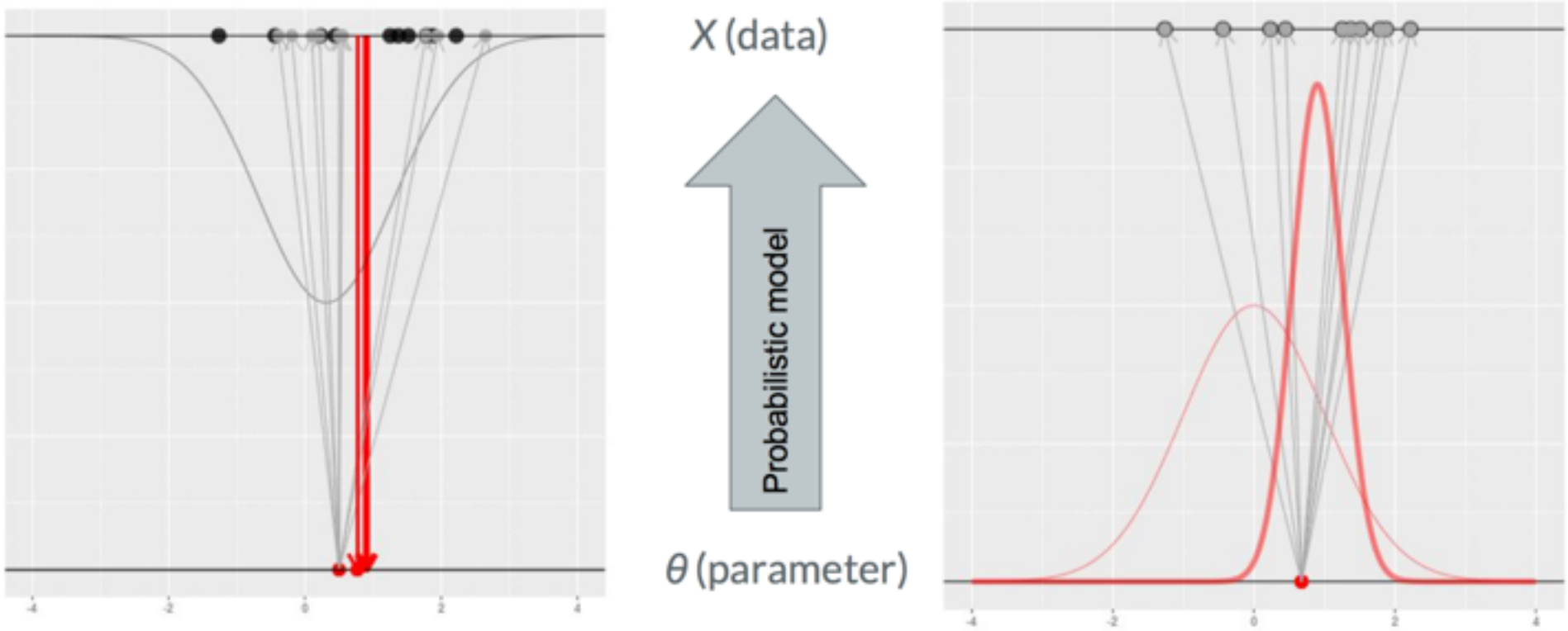
Of course, in practice you don't usually generate parameters and data hoping to get your original dataset.

Instead, you use Bayes' rule:

$$p(\theta|x) = \frac{p(x|\theta)\,p(\theta)}{p(x)}$$

This is intractable in general (the denominator is a problem). Turn to approximations schemes like MCMC, variational Bayes, &c.

Frequentist: the parameter is fixed, data is random.
Bayes: the data is fixed, the parameter is random.

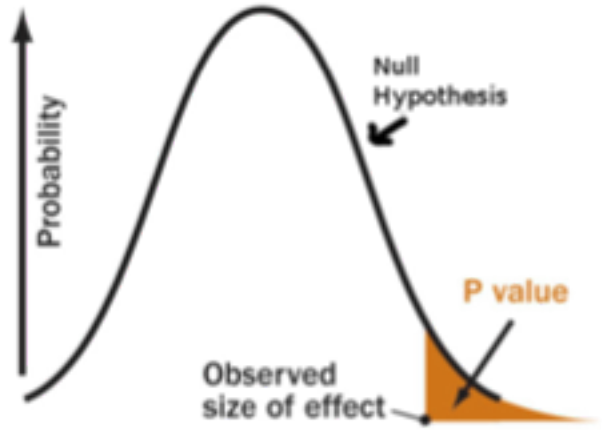


What if we do not know the probability distribution (likelihood)?

- We can still create posteriors with the following procedure, called Approximate Bayesian Computation (ABC): from a given prior we generate a realization of both the parameters and the data.
- We throw away the data realizations that are not consistent with the observed data.
- The remaining realizations will give the correct posterior of parameters.
- In practice this can only be done with some summary statistics for the data (curse of dimensionality), accepting some finite "distance" between these realizations and the data.
- While this is how ABC is normally defined as likelihood free, in a well defined sense if we know how to generate data given the parameters we have the likelihood (implicitly, as a function of latent variables needed to generate simulated data)

p-value for Hypothesis Testing

- Probability of finding the observed, or more extreme (larger or smaller), when H_o, null hypothesis is true



If $p < \alpha$, H_o rejected
$p > \alpha$, H_o accepted

often $\alpha = 0.05$

Example: We predict $H_o = 66$, but we observe $H_o = 73 \pm 3$. So H_o is more than 2-sigma away $p < 0.05 \rightarrow H_o$ rejected

Gaussian distribution: ± 1 sigma $p = 0.32$
± 2 sigma $p = 0.045$
± 3 sigma $p = 0.0017$

Criticisms of p-value

1) Discrete: If $p < \alpha$ rejected, $p > \alpha$ accepted.
Only α is reported in N-P testing, and this guarantees coverage.
So if we measure $H_o = 72 \pm 3$ we accept $H_o = 66$ ($p>0.05$),
if we measure $H_o = 72.1 \pm 3$ we reject it ($p<0.05$).

This makes little sense: the data is almost the same

2) Decision depends only on H_o, not on alternative hypotheses.
Some viewed it as a good thing (Fisher), but it can also be bad:

Sherlock Homes: once we reject all alternatives, the remaining one, no matter how improbable, is the correct one (Sherlock was using the likelihood principle!)

3) The p-value cannot be interpreted as error distribution
all that matters is whether $p < \alpha$, regardless of value of p.
E.g. $H_o = 72.1 \pm 0.3$ is just as good as $H_o = 72.1 \pm 3$ to reject 66.

Criticisms of p-value

this long series of tests, half of the null hypotheses are initially true, then, among the subset of tests for which the p-value is near 0.05, at least 22%—and typically over 50%—of the corresponding null hypotheses will be true. As another illustration, Sterne and Davey Smith (2001) estimated that roughly 90% of the null hypotheses in the epidemiology literature are initially true; the applet shows that, among the subset of such tests for which the p-value is near 0.05, at least 72%—and typically over 90%—of the corresponding null hypotheses will be true. The harm from the common misinterpretation of $p = 0.05$ as an error probability is apparent.

It is natural (and common) in these sciences to fault the statistics profession for the situation, pointing out that common textbooks teach frequentist testing and then p-values, without sufficient warning that these are completely different methodologies (e.g., without showing that a p-value of 0.05 often corresponds to a frequentist error probability of 0.5, as indicated by the mentioned applet and conditional frequentist developments).

In contrast, the statistics profession mostly holds itself blameless for this state of affairs, observing that the statistical literature (and good textbooks) does have appropriate warnings. But we are not blameless in

J. Berger: http://www2.stat.duke.edu/~berger/applet2/pvalue.html

In a setting where we have two (or more) hypotheses, the probability of rejecting a valid null hypothesis when p is close to 0.05 is high. Note that there are many more cases with p > 0.05, which are inconclusive (we do not reject either).

One approach: move from 2 sigma to 5 sigma

- P value for 5 sigma is $3\text{x}10^{-7}$, vs 0.045 for 2 sigma.
- Even if this cannot be interpreted as the error rate it is clear that the rate will be very very small. For example, likelihood ratio is $e^{(-25/2)} = 3\text{x}10^{-6}$
- Experimental particle physics has decided, through many repeated experiments, that 5 sigma provides good protection against false positives and negatives. It ”only” needs 4 times more data than 2.5 sigma
- 5 sigma may be an impossible goal in some fields where more data cannot easily be taken
- Who wants to wait for 4 times more data when we can continuously update our posteriors as new data come in

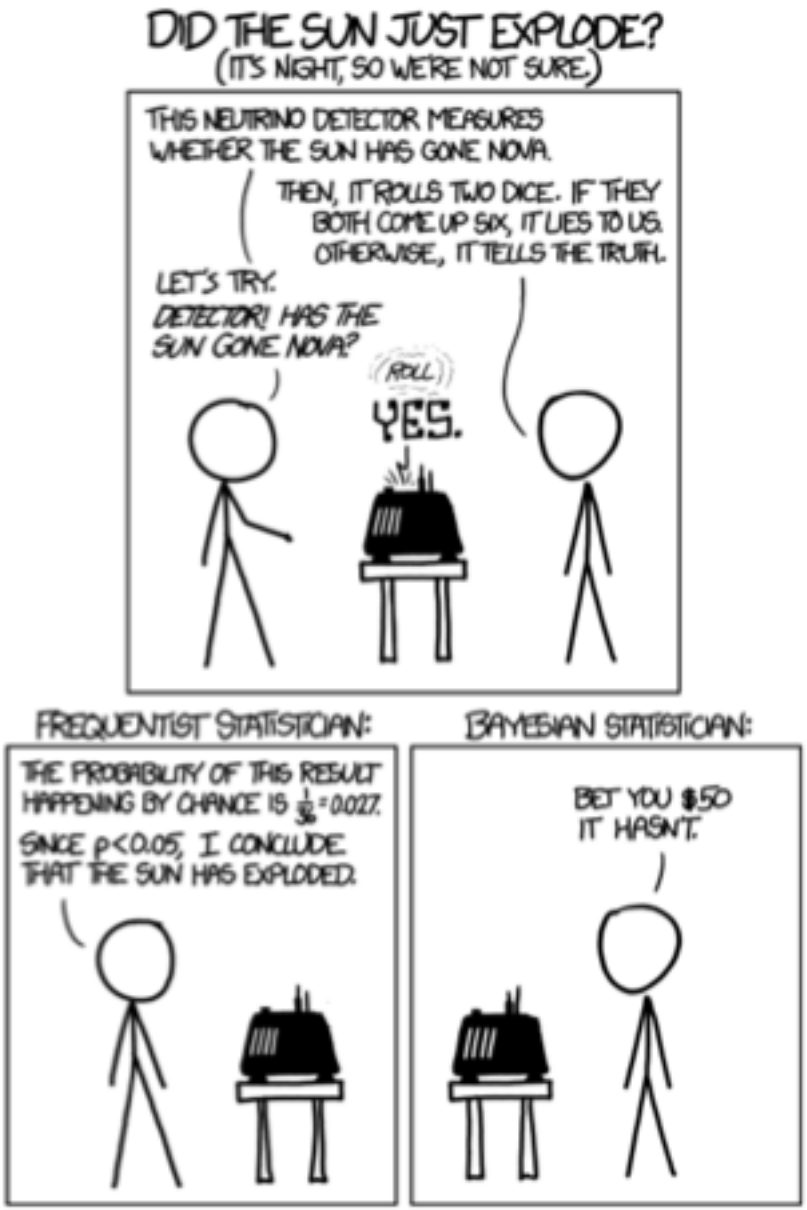


Figure 1: From xkcd.com

Third school of hypothesis testing: **Fisher's p-value**

- Fisher's significance testing: use p-values without the frequentist concept of coverage, but also without priors and without alternative hypotheses.

  Best or worst of both worlds?

  Note that this is what is being done in today's practice: we report p, not just a=0.05, and we attach some sense to the validity of null hypothesis from its value.

  Fisher was not a Bayesian, but was also not a frequentist.

- Main argument: p value is useful since it can be defined without alternatives (goodness of fit test).

  We will return to this later.

Classical Statistics: automated, cookbook recipes (very fat books).

Can be a good thing (many options to try) or a bad thing (need to know them; only one (at best) is optimal…)

Why it persisted for so long as the only option?

Slow computers (or unavailable): Bayesian requires high computing power (we will discuss methods later)

Worst case scenario (coverage for any truth) favors frequentism,
Average scenario favors Bayes

> 2) When interpreting results of Bayesian estimates, there is one more important issue - all such statements are conditional on chosen prior. Thus Bayesian methods do provide more natural interpretation, provided that we are willing to summarize our beliefs over the parameter uncertainty in precise probabilistic terms. I think this is not such a trivial assumption!
>
> On the other hand, frequentist methods (and CIs in particular) perhaps do not provide straightforward interpretation, but they don't require choosing a prior either. They simply say: here is an algorithm to provide an interval estimate from data, and what you do with it is up to you. We can't guarantee that it will work in every case, but if you use it lot of times, it will be correct in such and such fraction of cases, whatever the true parameter is.

> Some texts interpret confidence intervals as follows: if I repeat the experiment over and over, the interval will contain the parameter 95 percent of the time. This is correct but useless since we rarely repeat the same experiment over and over. A better interpretation is this: On day 1, you collect data and construct a 95 percent confidence interval for a parameter $\theta 1$.

>On day 2, you collect new data and construct a 95 percent confidence interval for an unrelated parameter $\theta 2$. On day 3, you collect new data and construct a 95 percent confidence interval for an unrelated parameter $\theta 3$. You continue this way constructing confidence intervals for a sequence of unrelated parameters $\theta 1$, $\theta 2$, . . . Then 95 percent of your intervals will trap the true parameter value. There is no need to introduce the idea of repeating the same experiment over and over.

Still an issue in that a frequentist approach does not answer the question of what is the best possible reduction of uncertainty given the data at hand

The two schools are likely to agree to disagree on the language of statistics, but often lead to the same conclusions: we will show this explicitly in the asymptotic limit where priors don't matter.

They both want the best possible results in practical applications, hence should not be viewed as competing, but complementary.

What are shortcomings of Bayesian statistics?

- 1) Computationally intractable in high dimensions: approximate Bayes is no longer guaranteed to be optimal.
- 2) Likelihood is not always known, so needs to be approximated
- 3) Specially in high dimensions even non-informative priors can be tricky and lead to a bias
- Our approach will be: use Bayesian methodology since it is optimal in a well defined sense (minimizes Bayes risk) but verify bias and confidence intervals of the final result (which may arise because of the reasons above) with simulations (frequentist guarantees)

A Socratian dialog

- Jason Eisner, computer science professor at Johns Hopkins
- Read this and you will understand why this is a Bayesian course. You may not understand all the points immediately: reread this as we go through the course
- *Q: How many frequentists does it take to change a light bulb?*
  A: Well, there are various defensible answers ...
- *Q: How many Bayesians does it take to change a light bulb?*
  A: It all depends on your prior!
- **Narrator:** Let p be an unknown probability distribution. An *estimator* is a function that attempts to answer a question about p, given a dataset that was sampled from p.
- Statistics is largely about the design and analysis of estimators ...
- **Freddy the Frequentist:** Here's an estimator I just made up! And I can prove that for *any* p in this family of distributions, my estimator "works well."
- **Narrator:** Please explain "works well" to our audience. You may wave your hands.
- **Freddy (waving hands):** No matter what p is, my estimator will generally give pretty accurate answers on datasets sampled from p. Of course, it will fail on the occasional *unrepresentative* dataset, but there's no avoiding bad luck.
- **Basia the Bayesian**: Congratulations! Is it optimal at "working well"? I always want optimal.
- **Freddy:** Oh, there's no single best estimator for this problem. But at least I can prove that mine is "admissible." That is, some other estimator might beat my estimator for some p, but not for all p.
- **Basia**: Okay ... then how about the distributions p that you expect to arise in practice? Is your estimator especially well-suited to those?
- **Freddy:** Who knows what p will arise in practice?

- **Basia**: You do, apparently. You already assumed that p will come from a particular family. If it doesn't, your estimator has no guarantees.
- **Freddy:** Ok, you got me. But my assumption is pretty mild. People often have good reasons [e.g., the Central Limit Theorem] to believe that their data-generating distribution *does* come more or less from my family. I want my estimator to work well as long as p is in that family.
- **Basia**: But you'll get your actual dataset from scientists. Won't they have a more specific scientific hunch about what p is likely to be? Then you could get better results by taking that into account.
- **Freddy:** That sounds suspect. The scientists want objective conclusions, not conclusions that reflect their existing assumptions.
- **Basia**: Objective?? Anthropologists and journalists say there's no such thing as an objective viewpoint: as in physics, you always bring your own frame of reference. The most that data can do for you is to update your existing beliefs. Didn't you pay attention in philosophy class? The skeptics tell us there's no way to know anything for sure. The existentialists tell us that you can't give up your free will, as much as you'd like to. You can try to foist off that responsibility by adopting some principle like law or loyalty or religion—or frequentism!—but that's just an indirect way of making your choices.
- **Freddy:** Blah blah blah. Statistics is math, not philosophy.
- **Basia:** Statistics is applied epistemology. You and I are continuing an old philosophical conversation: how can we properly conclude anything from data? Yes, our modern mathematical tools let us rigorously derive precise conclusions, but only once we've made assumptions. Even mathematicians need to assume some axioms. When we deal with data, we also have to make assumptions about where the data came from. No assumptions, no conclusions. Life sucks that way.

- **Freddy:** But I already made a mild assumption and invented a good estimator! All your defeatist talk isn't giving me a better one.
- **Basia**: No, I'm not going to give you one estimator. I'll give you a way to *automatically derive* a better estimator by making more assumptions. Yours is guaranteed to work pretty well for *all* p in the family, but mine will work better for *typical* p.
- **Freddy:** Typical?? I'm doing *worst-case* analysis. You want to do *average-case*analysis, so what's the average case? Am I supposed to average equally over all p?
- **Basia**: Just write down your prior distribution on p, which reflects your true beliefs—prior to the experiment—about where p might fall. If you really aren't sure, then your prior should be "flat" and treat all p more or less equally.
- **Freddy:** And once I've written down the prior distribution ...?
- **Basia**: Then the Bayesian estimator will just fall out! There is nothing left to design. Given a dataset, the Bayesian estimator just reweights your prior belief in each hypothesis p according to that hypothesis's probability of generating the dataset. If you have a loss function, then optimal decisions fall out of the new beliefs, again with no further design, thanks to Bayesian decision theory.
- **Freddy:** "Optimal" decisions ... if you believe the prior.
- **Basia**: Hey, you admitted that *you* don't have a principled way to choose among admissible estimators. Different estimators make different predictions, too.
- I have to specify my prior. I don't have a principled way to choose among different priors; I'm just supposed to *have* a prior belief. But at least I'm being explicit about it! So whoever reads my paper can see exactly what led to my conclusions. I am not slopping around with different estimators. My estimator insists on using all of the data. By Bayesian principles, the data and the prior lead inexorably to the conclusions.
- Of course, my readers are free to argue with me about whether my prior represents the current state of scientific knowledge. We can test how different priors would affect the conclusions.

- **Narrator:** Thank you for a stimulating discussion! This is getting very tl;dr. The audience is now free to leave.
- **Freddy:** I see that you really want to squeeze every drop of value out of the data. But why try to define the One True Estimator? Mine is good enough. I can bound the bias and variance of my estimator as a function of the dataset size, so I can prove to you that large errors are not very likely for practical datasets.
- As a practical matter, my estimator is also easy to compute. In fact, that's how I came up with it: I conjectured a simple reasonable procedure and then proved that it had good properties. Your Bayesian estimator was easy enough to write down mathematically, but maybe it's hell on wheels to compute, which also makes it hard to analyze.
- **Basia**: That's fair. In fact, usually I have no practical way to compute it exactly. I have to design a randomized algorithm or variational approximation. So my practical conclusions don't follow inexorably from the data plus the prior. They are also affected by the computational approximation.
- But perhaps drawing exact conclusions from data *should* be computationally intensive. Scientific reasoning is quite involved when humans do it. Scientific processes are intricate, which leads to complex families of models. Scientific experiments produce heterogenous, noisy, incomplete data.
- The Bayesian approach handles all of this complexity seamlessly. Once you've designed your model, Bayesianism consists of a single simple statistical principle, backed in practice by a library of computational tricks.

- **Freddy:** I grant you that in these fancy situations, frequentist estimators would become computationally expensive too. I also admit that it would be hard for me to devise an estimator for such a situation (let alone for many related situations) that had provably good frequentist properties.
- I'd probably fall back on a maximum-likelihood estimator. That's like a pared-down version of your Bayesian estimator, so it's at least as feasible to compute. And it doesn't need a prior.
- **Basia:** I'm not crazy about maximum likelihood. It *ignores* the information in the prior. And it gives only a silly point estimate, instead of representing posterior uncertainty. This will lead you to worse decisions.
- **Freddy:** So maybe I'll add a regularizer. Regardless, the effect of your prior diminishes as the dataset grows, and so does your posterior uncertainty. So at least we'll agree with each other in the infinite-data limit. And at that point we'll also agree with the truth: I'm not crazy about maximum likelihood estimation either, but at least it's consistent.
- **Narrator:** Ok then! Great to see you in agreement.
- **Freddy:** Bye, non-expert audience! Hope you had fun. You can upvote us on your way out.
- But Basia, between the two of us, I still don't share your philosophical stance on what we want from an estimator. Let's drop the infinite-data fantasy. We'll have finite data, so we want the estimator's risk to decrease rapidly as a function of the dataset size. If I were considering an estimator for a complicated model, I would *try* to prove that it did this for *any* distribution in the family. That wouldn't require any prior.
- **Basia**: But what do you mean by "any distribution in the family"? With complicated models, is that even a natural concept? Let me sketch a basic hierarchical Bayesian model:
- draw some hyperparameters from the prior distribution
- draw parameters from distributions controlled by the hyperparameters

- draw data from distributions controlled by the parameters
- What's the family here?
- **Freddy**: Here I'd treat the hyperparameters and the parameters differently. I'm willing to assume that p has your hierarchical form: as you pointed out before, I'll accept hard constraints on p. I only throw away your prior over the hyperparameters, which is a soft constraint on p. Every setting of the hyperparameters is a different distribution p, so I want to design a frequentist method that works well for *any* such setting.
- **Basia**: But you didn't throw away the distributions that generate the parameters.
- **Freddy**: Right. So I have to regard those parameters at step 2 as *unobserved data* that get generated by the model along the way to step 3. They're "nuisance" variables. So when I average over random datasets, I'm doing average-case analysis of the parameters too. But since I try to show that this analysis comes out well for for *any* distribution, I'm doing worst-case analysis of the hyperparameters.
- **Basia:** What's your motivation for treating these two levels so differently??
- **Freddy:** Oh, I always distinguish two levels. There's some *set of distributions*. For each distribution in the set, I want to do well on average.
- **Basia:** You look at this three-level hierarchical model and you see a *set of distributions over distributions*. By using a prior over the hyperparameters, I turn that into a *distribution over distributions over distributions.* Or equivalently, one big distribution. So I'm just analyzing everything in the average case. I don't see why you'd draw a special line between levels 1. and 2. of my model.

- **Freddy:** But I don't have to draw it there. I can draw it anywhere I choose. You want to throw out worst-case analysis altogether.
  But I get to mix worst-case and average-case analysis in different ways.
- When I draw the line above level 1., then everything is average-case and my analysis is indistinguishable from a Bayesian's. In that case, the family contains only one distribution p, which generates the hyperparameters, parameters, and data. So my estimator isn't estimating properties of p, which is known. It is imputing values of the nuisance variables, given p and the observed dataset.
- And here my estimator's risk no longer depends on a choice of p. It's an average over everything including the hyperparameters.
- **Basia**: Good! That is what I *always* minimize. My estimator is explicitly *defined* to minimize the Bayes risk—that is, the expected loss of the prediction, according to the posterior given the dataset. Since my estimator minimizes the Bayes risk for any dataset that it is given, then it also minimizes the frequentist risk you're talking about, which additionally averages over all possible datasets.
- **Freddy**: Yes, your estimator looks like an ideal solution if I draw the line above level 1, accepting your prior as part of the model itself. But that's a single, rather weak result. By choosing to draw the line in other places, I also get to formulate additional theorems about estimators. Theorems that contain ∀ symbols because they're doing worst-case analysis.
- **Basia**: That "weak result" is all I ever need in practice. Your additional theorems are true enough, but how do they help you?

- **Freddy**: Well, I become more comfortable recommending an estimator to the scientists. I can tell them what known properties it has, including various kinds of worst-case properties.
- **Basia**: But another frequentist might equally well recommend a different estimator, which also has good properties but will make different predictions.
- Your theorems are just talking points; they confuse the issue. I don't need any theorems to make a recommendation. My Bayesian recommendation is to derive the estimator directly from your scientific assumptions and engineering goals. I am *always* going to tell the scientists to use a generalized Bayes rule: if they actually trust their model and prior, then the best prediction from the data is the one that minimizes the Bayes risk.
- **Freddy:** I think you're actually leaning on the complete class theorem. Which you feel solves all of statistics. What do you do all day, then? Must be a cushy job.
- **Basia:** Well, I help the scientists formalize their model, prior, and loss function. That doesn't require new *statistical* theorems—but there's still math to do. I may have to design and analyze new probability distributions. I also design and analyze algorithms to help the scientists compute the best prediction.
- **Freddy:** They deserve to know whether that "best prediction" will be any good. So maybe I should do frequentist analysis of your Bayesian estimator.
- **Basia:** Why bother? I'd just alert them to the Bayes risk of their actual prediction. That number is highly useful information because it conditions on their actual dataset.
- Your frequentist analysis would pay just as much attention to distributions p that are *ruled out* by their actual dataset. Who cares about doing well on those?? Especially when "doing well" means average performance over a lot of fictional datasets. Those are irrelevant.

- **Freddy:** But what if the scientists don't have an "actual dataset" yet? They will be analyzing *many* datasets. They need to make some decisions beforehand. First, should they adopt your statistical software? Second, how much data should they collect?
- These are indeed questions about how well your software—or mine—will do on the average dataset of size n, for a range of distributions p. Any software box should have a "nutritional information" sticker on it with answers to those questions.
- **Basia:** Ok, but that sticker doesn't have to focus on the worst-case p. The scientists have a prior over p. My software consults the prior, and yours doesn't. But in each case, the scientists want to know how well the software will do on distributions p chosen *from their prior*. I could estimate that for them by sampling distributions and datasets from their prior.
- **Freddy:** In principle you could. But in practice, you may want to publish the sticker before you know who will be using the software. Frequentist theorems are nice and portable that way—just like nutrition labels, they are aimed at helping lots of different users, who may have different priors.
- We can formulate a frequentist estimator without knowing the user's prior. And we can publish its worst-case risk without knowing the user's prior. The user knows that the worst-case risk is at least an upper bound on their average-case risk, no matter how they prefer to average.
- **Basia:** I think your objection is coming down to computational inconvenience again! You want to devise general estimators and prove general theorems ... in order to avoid doing specific computations that would give you the best possible answer in your precise situation.
- It's no wonder that statistics has historically focused on general theorems. It wasn't computationally feasible to do more. Maybe I'm a Bayesian because I came of age surrounded by computational power and techniques like MCMC. I respect the generality and elegance of theoretical bounds, in the simple cases where you can get them. But I also appreciate machine learning work that focuses on measuring and maximizing the performance of specific predictive systems, rather than proving broader theorems about weaker systems.

"Bayesian" Milestones:

- Bayes (1763), Laplace (1774), Jeffreys (1939)
  Almost nothing until 1990's, when Gibbs sampling arrived

- Very prominent critics in 20th century:
  Pearson (Egon), Neyman, Fisher

- Today:
  Explosion led by efficient codes
  (BUGS/JAGS, STAN, MCMC samplers) and fast computers,
  Bayes dominates in some fields (astronomy, physics,
  bioinformatics, data science),
  Frequentist more common in medicine, economics and humanities

And then, the geologist Harold Jeffreys made Bayes' Theorem useful for scientists, proposing it as an alternative to Fisher's 'p-values' and 'significance tests', which depended on "imaginary repetitions." In contrast, Bayesianism considered data as fixed evidence. Moreover, the p-value is a statement about data, but Jeffreys wanted to know about his hypothesis *given* the data. He published the monumental *Theory of Probability* in 1939, which remained for many years the only explanation of how to use Bayes to do science.

For decades, Fisher and Jeffreys were the world's two greatest statisticians, though both were practicing scientists instead of theoreticians. They traded blows over probability theory in scientific journals and in public. Fisher was louder and bolder, and frequentism was easier to use than Bayesianism.

In 1950, an economist preparing a report asked statistician David Blackwell (not yet a Bayesian) to estimate the probability of another world war in the next five years. Blackwell answered: "Oh, that question just doesn't make sense. Probability applies to a long sequence of repeatable events, and this is clearly a unique situation. The probability is either 0 or 1, but we won't know for five years." The economist replied, "I was afraid you were going to say that. I've spoken to several other statisticians, and they all told me the same thing."

In 1954, Savage published *Foundations of Statistics*, which built on Frank Ramsey's earlier attempts to use Bayes' Theorem not just for making inferences but for making decisions, too. His response to a classic objection to Bayesianism is worth remembering. He was asked, "If prior opinions can differ from one researcher to the next, what happens to scientific objectivity in data analysis?" Savage explained that as we gain data, subjectivists move into agreement, just as scientists come to consensus as evidence accumulates about, say, cigarettes causing lung cancer. When they have little data, scientists are subjectivists. When they have tons of data, they agree and become objectivists.

Finally, in 1983 the US Air Force sponsored a review of NASA's estimates of the probability of shuttle failure. NASA's estimate was 1 in 100,000. The contractor used Bayes and estimated the odds of rocket booster failure at 1 in 35. In 1986, *Challenger* exploded.

Summary

1) In this course we adopt Bayesian statistics not because it is superior or more correct (although the Freddy-Basia dialog makes a compelling case it is), but because it is easier and usually is as good as the best classical statistics: it has only one equation, and everything follows from it: no need to learn anything but probability (i.e., write down likelihoods). But we will study some non-Bayesian concepts and we will insist on verification

2) Priors are subjective: This can be a good thing. Likelihoods are also subjective in practice: e.g. we typically assume data are uncorrelated and that we know $p(d|\lambda)$. This remains one of the main issues of Bayesian st.

3) In practice for intervals in most cases very little difference between confidence interval (with coverage guarantee) and credible interval (corresponding Bayesian concept)

4) Hypothesis testing: Bayesian versions typically weaker than p-value. This is because alternating hypotheses can also give an "unlikely" data draw, weakening a null hypothesis rejection.

Literature

- D. Mackay, *Information Theory, Inference, and Learning Algorithms* (See course website)

  Chapter 2.1 – 2.3, 3. Exercises very instructive
- M. Kardar, *Statistical Physics of Particles*, Chapter 2