Ay 119: Bayesian statistics

The basics of statistical inferencing

Matthew J. Graham
mjg@caltech.edu

Outline

- Belief
- Bayes Theorem
- Choice of priors
- Model selection
- MLE
- MCMC and Metropolis-Hastings

A matter of belief

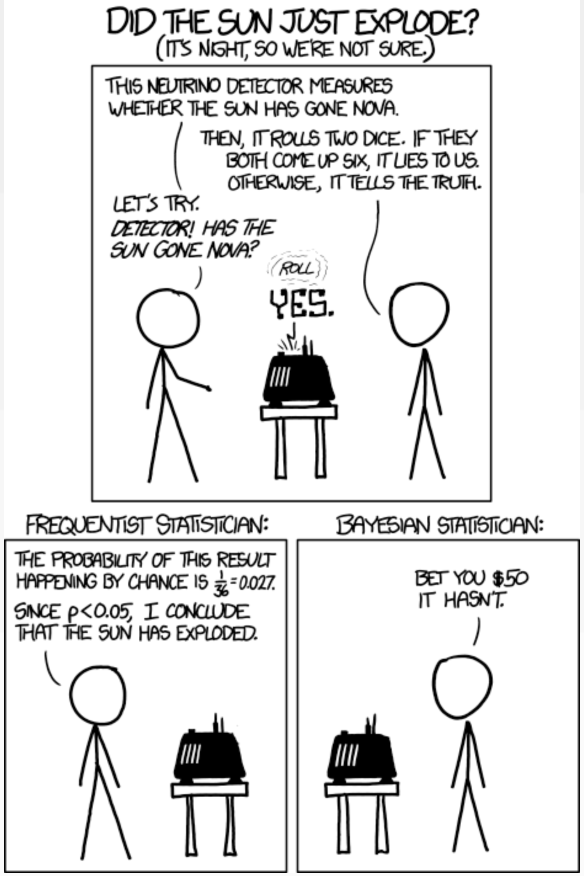
- What is the mean height of people in the room?
- What are the confidence intervals on this?
- What does this mean?

Frequentist:

- The range in which the mean will occur 95% of the time with repeated sampling

Bayesian:

- The interval in which 95% of the population lies


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}=0.027$. SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Bayes Problem

"Let us then imagine a person present at the drawing of a lottery, who knows nothing of its scheme or of the proportion of Blanks to Prizes in it. Let it further be supposed, that he is obliged to infer this from the number of blanks he hears drawn compared with the number of prizes; and that it is enquired what conclusions in these circumstances he may reasonably make.

Let him first hear ten blanks drawn and one prize, and let it be enquired what chance he will have for being right if he guesses that the proportion of blanks to prizes in the lottery lies somewhere between the proportions of 9 to 1 and 11 to 1"

An Essay towards solving a Problem in the Doctrine of Chances

Rev. T. Bayes (1763)

Bayes Problem

- Probability of winning, *p*, is a random variable in [0, 1]
- The result of each draw, X_i, is conditional on p:
  $p(X = 1) = p$ for a prize and $p(X = 0) = 1 - p$ for a blank
- After *n+m* draws, there will be *m* prizes and *n* blanks:

$$f(n,m) = \frac{(n+m)!}{n!\,m!} p^m (1-p)^n = \binom{n+m}{m} p^m (1-p)^n$$

- The chance that *p* lies between two values *a* and *b*:

$$P(a < p < b|m,n) = \frac{\int_a^b \binom{n+m}{m} p^m (1-p)^n dp}{\int_0^1 \binom{n+m}{m} p^m (1-p)^n dp}$$

- For *a* = 1/11, *b* = 1/9, *m* = 1 and *n* = 10, p ~ 0.077

"there would therefore be an odds of about 923 to 76, or nearly 12 to 1 ***against*** *his being right"*

Bayes Theorem

- The distribution of the data, *n* and *m*, for a given value of the unobserved variable, *p*, is the **likelihood function**: *p(D|M)*
- The initial assumption of the distribution of *p* is the **prior**, *p(M)*
- The denominator is the **marginal likelihood** or **evidence**, *p(D)*
- The final result is the **posterior probability,** *p(M|D)*:

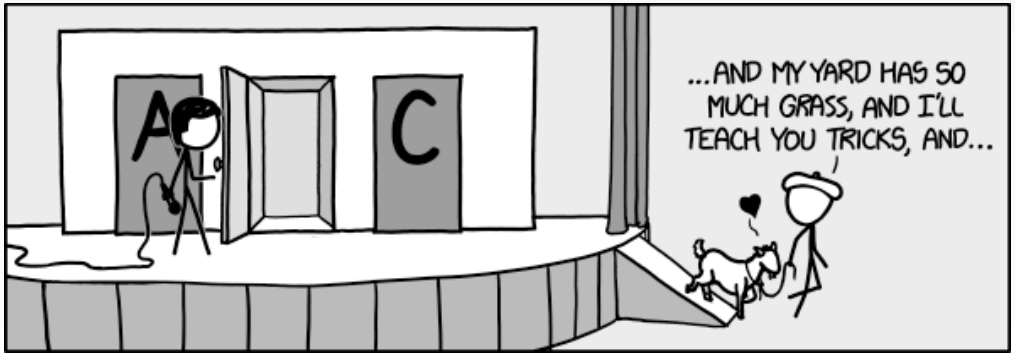
$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

The Monte Hall Problem

- You’re on a game show and are presented with three closed doors
- Behind one is a fabulous prize but behind the other two are goats
- The host asks you to pick one door and shows you that behind one of the others is a goat (it’s always a goat)
- Do you want to stick with your first choice?

$$p(prize\ behind\ A|goat\ behind\ B) =$$

$$\frac{p(goat\ behind\ B|prize\ behind\ A)\cdot p(prize\ behind\ A)}{p(goat\ behind\ B)} = \frac{\frac{1}{2}x\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$


A
C
...AND MY YARD HAS SO MUCH GRASS, AND I'LL TEACH YOU TRICKS, AND...

Choice of prior

- In Bayes' problem, we assumed a uniform distribution for *p*:

$$p(\theta) = const., a < \theta < b$$

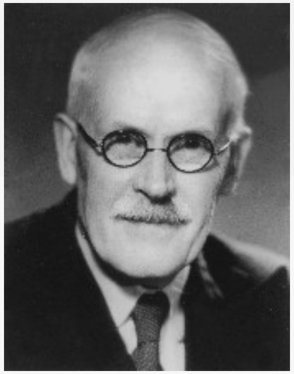
- This is called a **flat** prior or an **uninformative** prior
- Describes a state of knowledge in which we have observed at least one success and one failure, and have prior knowledge that both states are physically possible
- If parameter is limited to positive real values then the prior should be uniform in the logarithmic range:

$$p(\theta) \propto \theta^{-1} \Rightarrow p(\ln\theta) = const.$$

Jeffrey's rule

- The general recommendation for an uninformative prior is the square root of the determinant of the Fisher information for the model: $p(\theta) \propto \sqrt{\det \mathcal{I}(\theta)}$ where $\mathcal{I}(\theta)$ is the second moment of the partial derivative with respect to **θ** of the natural logarithm of the likelihood function:

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\theta)\right)^2\right] = -\mathrm{E}\left[\left(\frac{\partial^2}{\partial\theta^2}\log f(\theta)\right)\right]$$



- For a Gaussian with an unknown mean:

$$f(x|\mu) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$$

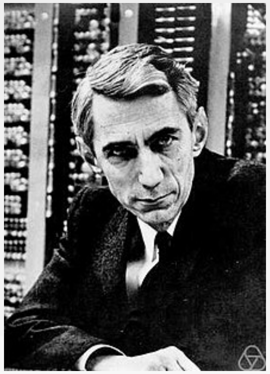
and so the prior is:

$$p(\mu) \propto \sqrt{\mathrm{E}\left[\left(\frac{x-\mu}{\sigma^2}\right)^2\right]} = \sqrt{\int_{-\infty}^{\infty} f(x|\mu)\left(\frac{x-\mu}{\sigma^2}\right)^2 dx} = \frac{1}{\sigma}$$

Maximum entropy

- We want a measure that captures the *information content* of a distribution and then assign a prior that reflects our ignorance of the true parameter value by optimizing this quantity:

$$S(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i \ln p_i \quad \text{(Shannon entropy)}$$



- The **maximum entropy** prior maximizes:

$$S = -\int dx\, p(x) \ln p(x)$$

- So if we know the variance σ^2 is finite for an arbitrary mean, μ, we can use Lagrange multipliers to show that:

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

Amongst all real-valued distributions with a specified variance, the Gaussian has the maximum entropy

Bayes factor

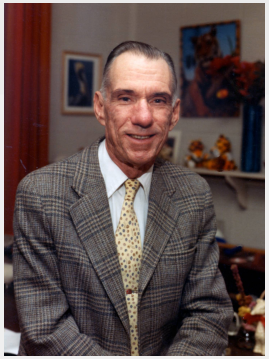
- The ratio of the relative likelihoods of the data under each model:

$$K = \frac{p(D|M_1)}{p(D|M_2)} = \frac{\int p(D|\theta_1, M_1)p(\theta_1|M_1)d\theta_1}{\int p(D|\theta_2, M_2)p(\theta_2|M_2)d\theta_2}$$

- Posterior odds = Bayes factor x prior odds

2 ln K	**K**	**Strength of evidence**
0 - 2	1 - 3	Not worth more than a bare mention
2 - 6	3 - 20	Positive
6 - 10	20 - 150	Strong
> 10	> 150	Very strong

Evidence for ESP?



- A study of whether psychokinesis could influence an electronic random number generator reported 52263471 successes out of 104490000 Bernoulli trials (ratio = 0.50017768)

- In the absence of any effect, ratio = 0.5 so *p*-value = 0.0003 (3.6σ)
- Assume $\theta_1 = 0.5$ for model 1 and an unknown θ_2 for model 2 with Jeffreys' prior for Bernoulli, $f(\theta) = 1/\sqrt{(\theta(1-\theta))}$:

$$K = \frac{p(D|\theta_1 = 0.5)}{\int_0^1 p(D|\theta_2) f(\theta_2) d\theta_2} = \frac{\pi \cdot 0.5^N}{B(S+0.5, N-S+0.5)} = e^{2.93} = 18.7$$

So there is positive evidence against ESP. For a uniform prior, K ~ 15.4.

Maximum likelihood estimation

- The likelihood of a data set given a particular model is the joint probability of each individual data point given the model:

$$L \equiv p(\{x_i\}|M(\theta)) = \prod_{i=1}^{n} p(x_i|M(\theta))$$

- Although it is often useful to deal with the log-likelihood:

$$L = \prod_{i=1}^{n} \ln(p(x_i|M(\theta)))$$

- For the best-fit form of a model to a data set, the likelihood will be optimized in terms of the model parameters:

$$\frac{\partial L}{\partial \theta_i}\left(x_i, \hat{\theta}_i\right) = 0$$

MLE example

- Consider a power law (Pareto) model of the form:

$$f = \frac{a x_m^\alpha}{x^{\alpha+1}} \text{ for } x \geq x_m$$

- This gives a log-likelihood of:

$$\ln L = n \ln \alpha + n\alpha \ln x_m - (\alpha + 1) \sum_{i=1}^{n} \ln x_i$$

So

$$\hat{\alpha} = \frac{1}{\sum_{i=1}^{n} \ln \frac{x_i}{n} - \ln x_m}$$

- As the sample size increases, the distribution of the MLE tends to a Gaussian distribution with mean θ and covariance matrix equal to the inverse of the Fisher information matrix, $\mathcal{I}(\theta)$.
- For the Pareto model: $\mathcal{I}(\alpha) = n/\alpha^2$ so $\hat{\alpha} \sim G(\hat{\alpha}, \frac{\alpha^2}{n})$

Maximum a posteriori (MAP) estimation

- The MLE is the most probable Bayesian estimator assuming a flat prior.
- For other priors, the maximum a posteriori estimate is better:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} L(\{x_i\}|\theta)p(\theta)$$

- Suppose we want to fit a Gaussian model to a data set and we believe the mean, μ, is drawn from a different Gaussian $G(\sigma_0, \sigma_m^2)$:

$$L(\{x_i\} \mid \mu)p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_m}\exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_m^2}\right)^2\right)\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x_j-\mu}{\sigma^2}\right)^2\right)$$

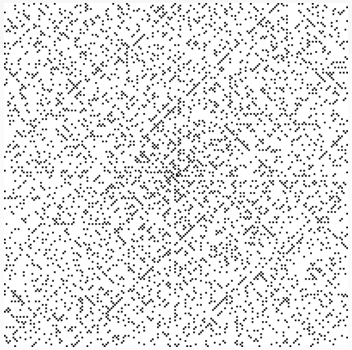
- From this the MAP estimate for μ is:

$$\hat{\mu}_{MAP} = \frac{n\sigma_m^2}{n\sigma_m^2+\sigma^2}\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right) + \frac{\sigma^2}{n\sigma_m^2+\sigma^2}\mu_0$$

which is a linear interpolation between the prior mean and the sample mean weighted by their respective covariances

Random sampling

- In 1946, Stan Ulam was playing solitaire and decided to try to compute the chances that a particular solitaire laid out with 52 cards would come out successfully. After attempting exhaustive combinatorial calculations, he decided to go for the more practical approach of laying out several solitaires at random and then observing and counting the number of successful plays.
- The idea of selecting a statistical sample to approximate a hard combinatorial problem is at the heart of Monte Carlo simulations
- We often want to generate a random sample from a particular distribution to approximate the distribution or compute an integral involving the distribution.



Estimating multidimensional integrals

- The general multidimensional integration problem is of the form:

$$I(\theta) = \int g(\theta)p(\theta)d\theta$$

- This can be computed numerically by generic Monte Carlo where a random set of *M* values uniformly sampled from within the integration value V_θ gives an estimate of the integral:

$$I \simeq \frac{V_\theta}{M}\sum_{j=1}^{M} g(\theta_j)p(\theta_j)$$

- It would be much better if we could guarantee that the random set of values we use is (at least) asymptotically proportional to $p(\theta)$ to give:

$$I(\theta) = \frac{1}{M}\sum_{j-1}^{M} g(\theta_j)$$

Markov Chain Monte Carlo

- Consider a sequence of random variable where the probability of a given state at point *t+1* only depends on the state at point *t*. Let *T* be a matrix of the transition probabilities between states.
- Now we want the sequence to reach a stationary distribution proportional to some $p(\theta)$ and so the probability of arriving at point *t+1* must be proportional to $p(\theta_{t+1})$:

$$p(\theta_{t+1}) = \sum_y T(\theta_{t+1}, y) p(y)$$

$$p(\theta_{t+1}|\theta_t) = p(\theta_t|\theta_{t+1})$$

- The most popular MCMC algorithm is Metropolis-Hastings and this adopts:

$$T(\theta_{t+1}|\theta_t) = p_{acc}(\theta_{t+1}, \theta_t)(\theta_{t+1}|\theta_t)$$

$$p_{acc}(\theta_t, \theta') = \frac{Q(\theta_t, \theta')p(\theta')}{Q(\theta'|\theta_t)p(\theta_t)}$$