MTH 7241 Fall 2022: Prof. C. King

Notes 7: Bayesian Inference

0.1 Bayesian vs frequentist

The distinction arises from the interpretation of the probability of an event. We write $\mathbb{P}(A)$ in both cases, but the meaning is a little different.

0.1.1 Frequentist 'classical' meaning

$\mathbb{P}(A)$ is the long-run fraction of occurrences of the event A under repeated independent sampling. This leads to, for example, the weak law of large numbers: make independent measurements of a random variable $X_1, \ldots, X_n, \ldots$, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mathbb{E}[X]$$

$\mathbb{P}(A|B)$ is the long-run fraction of occurences of event A under repeated independent sampling when restricted to outcomes where B occurs.

0.1.2 Bayesian meaning

$\mathbb{P}(A)$ is your subjective belief of how likely it is that event A will occur.

$\mathbb{P}(A|B)$ is your updated belief of how likely it is that event A will occur given that event B has occurred.

Recall Bayes rule:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)} \tag{1}$$

0.1.3 Frequentist meaning

A and B are two random events, and it makes sense to consider both $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ (so long as both $\mathbb{P}(A)$ and $\mathbb{P}(B)$ are not zero). Bayes rule is the formula that gives the relation between these conditional probabilities. The order of events A, B in the conditional probabilities has no particular significance.

0.1.4 Bayesian meaning

$\mathbb{P}(A)$ is your belief in the likelihood of event A, and $\mathbb{P}(A|B)$ is your updated belief based on the fact that event B has occurred. Bayes rule tells you how to update your belief based on the measured data. The order of events A, B in the conditional probabilities is significant.

Bayesian application of Bayes rule:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)} \tag{2}$$

In most applications, the event B is measurement of some data, for example some independent measurements of a random variable $X_1, \ldots, X_n$. So we use $\mathcal{D}$ to indicate the data measured. The event A is our guess about the value of some set of parameters θ which determine the distribution of X. So $\mathbb{P}(A) = \mathbb{P}_0(\theta)$ is called the *prior* distribution and represents our belief about θ before the measurements are taken. Then $\mathbb{P}(B|A) = \mathbb{P}(\mathcal{D}|\theta)$ is called the *likelihood* and $\mathbb{P}(A|B) = \mathbb{P}_1(\theta|\mathcal{D})$ is the *posterior*. The normalizing constant $\mathbb{P}(B) = \mathbb{P}(\mathcal{D})$ is called the *evidence*. The formula is used in cases where distributions are discrete, continuous, or a mixture of the two. So the formula is

$$\mathbb{P}_1(\theta|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\theta)\,\mathbb{P}_0(\theta)}{\mathbb{P}(\mathcal{D})}$$

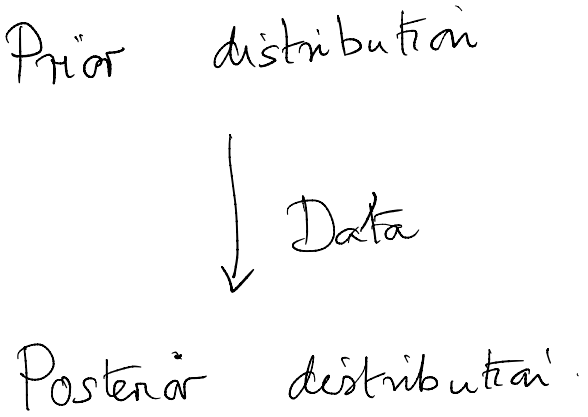
The normalizing constant can be computed by summing over possible values for θ:

$$\mathbb{P}(\mathcal{D}) = \sum_j \mathbb{P}(\mathcal{D}|\theta_j)\,\mathbb{P}_0(\theta_j)$$

Example: Let's look at an example where we want to decide between two competing hypotheses. Suppose your friend claims that she can predict the outcome of a coin toss with 100% certainty.
H_0: your friend is lying, and she has 50% probability of guessing the outcome.
H_1: your friend is telling the truth.
We want to decide between H_0 and H_1. Our prior distribution is

$$\mathbb{P}(H_0) = 0.99 = 1 - \mathbb{P}(H_1)$$

We measure data: a coin is tossed 5 times, and your friend predicts the outcome every time. What is your updated belief in H_0?



Bayesian inference provides a systematic way to answer this question. Let $\mathcal{D}$ denote the outcomes of the 5 coin tosses, then

$$\mathbb{P}(\mathcal{D}|H_0) = \left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

$$\mathbb{P}(\mathcal{D}|H_1) = 1$$

Using Bayes rule we get

$$\mathbb{P}(H_0|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|H_0)\,\mathbb{P}(H_0)}{\mathbb{P}(\mathcal{D})}$$

To compute $P(\mathcal{D}$ we use the total probability formula:

$$P(\mathcal{D}) = \mathbb{P}(\mathcal{D}|H_0)\,\mathbb{P}(H_0) + \mathbb{P}(\mathcal{D}|H_1)\,\mathbb{P}(H_1)$$

Putting it together we find

$$\mathbb{P}(H_0|\mathcal{D}) = \frac{0.99/32}{0.99/32 + 0.01} = 0.756$$

So our confidence that your friend is lying has been reduced to 75.6% based on her successes.

Prior: $P_0(H_0) = 0.99$

Posterior: $P_1(H_0|D) = 0.756$

Bayesian inference works just as well when we investigate continuous parameters. Here is the typical setting: the prior is a pdf for the parameter θ, say $f_0(\theta)$. The posterior f_1 is computed using some data $\mathcal{D}$ whose distribution depends on θ:

$$\boxed{f_1(\theta|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\theta)\, f_0(\theta)}{\mathbb{P}(\mathcal{D})}}$$

Here the evidence is computed using an integral with the prior pdf:

$$P(\mathcal{D}) = \int P(\mathcal{D}|\theta)\, f_0(\theta)\, d\theta$$

Parameter θ : prior pdf $f_0(\theta)$.

$\downarrow$ data

posterior pdf $f_1(\theta|\mathcal{D})$.

Example: You are given a coin. It is biased with $\mathbb{P}(H) = q$ but you know nothing about the value of q. You toss the coin n times and it comes up Heads k times, and Tails $n - k$ times. Find the posterior distribution of q. What is the probability of getting Heads on the next toss?

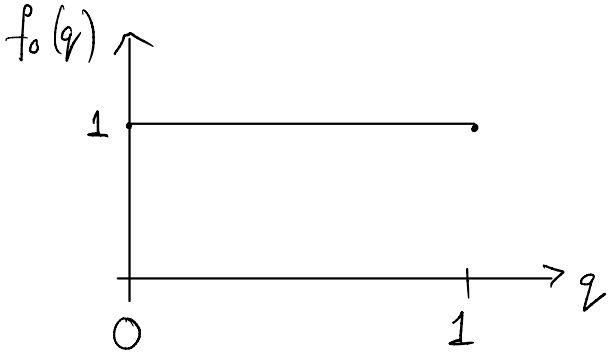
Bias: $\mathbb{P}(H) = q$ (?)

$\mathbb{P}(T) = 1 - q$

Prior for q: you know nothing except $0 \le q \le 1$.

Prior pdf $f_0(q) = \begin{cases} 1 & 0 \le q \le 1 \\ 0 & \text{else} \end{cases}$

uniform distribution on $[0, 1]$.



So we view q as a random variable.

Data: toss coin n times, get k Heads and $n-k$ Tails.

Prior pdf ? ✓

Likelihood ? $\mathbb{P}(\mathcal{D}|q)$ $\mathcal{D}$ = data

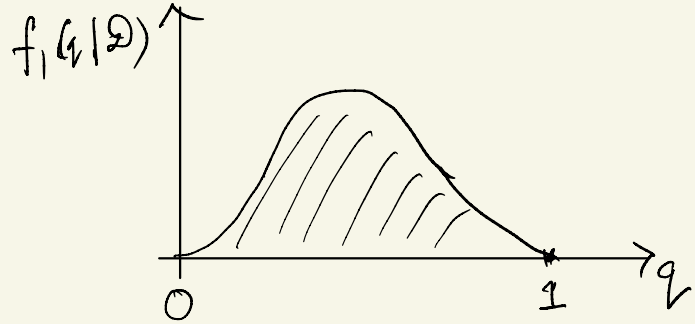
$$= \binom{n}{k} q^k (1-q)^{n-k}.$$

Evidence ? $$\mathbb{P}(\mathcal{D}) = \int_0^1 \mathbb{P}(\mathcal{D}|q)\, f_0(q)\, dq.$$

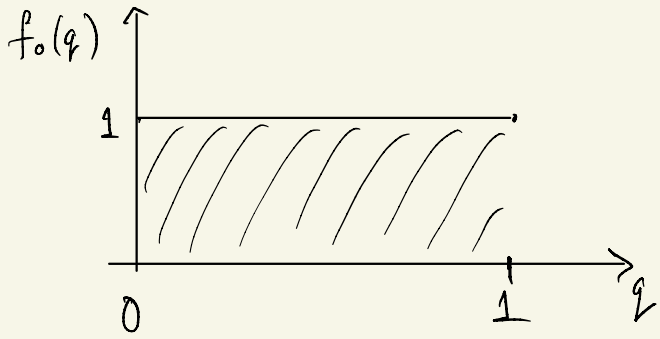
Posterior: $$f_1(q|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|q)\, f_0(q)}{\mathbb{P}(\mathcal{D})}$$

pdf for q → $$= \frac{\binom{n}{k} q^k (1-q)^{n-k} \cdot 1}{\mathbb{P}(\mathcal{D})} \qquad 0 \le q \le 1.$$

$f_1(0|\mathcal{D}) = 0, \quad f_1(1|\mathcal{D}) = 1$



Posterior



Prior

Evidence: is the normalization for the pdf.

$$\mathbb{P}(\mathcal{D}) = \int_0^1 \binom{n}{k} q^k (1-q)^{n-k} \cdot 1 \, dq$$

$$= \binom{n}{k} \frac{k!\,(n-k)!}{(n+1)!} \quad \text{(standard integral)}$$

$\Rightarrow$ substitute into formula

$$f_1(q|\mathcal{D}) = \frac{\binom{n}{k} q^k (1-q)^{n-k}}{\binom{n}{k} \frac{k!\,(n-k)!}{(n+1)!}}$$

$$= \frac{(n+1)!}{k!\,(n-k)!} q^k (1-q)^{n-k}$$

Question: toss the coin again, what is the probability to get Heads?

Initially: prior pdf $f_0(q) = 1 \quad (0 \le q \le 1)$

Toss coin once: what is prob. of Heads?

$$\mathbb{P}(H) = \int_0^1 \mathbb{P}(H|q)\, f_0(q)\, dq$$

↖ condition on the coin's bias

$$= \int_0^1 q \cdot 1 \cdot dq$$

$$= \frac{q^2}{2}\Big|_0^1 = \frac{1}{2}.$$

Question: after we get the data about the n coin tosses, we get the postair pdf.

$$\mathbb{P}(H|\mathcal{D}) = \int_0^1 \mathbb{P}(H|\mathcal{D}, q)\, f_1(q|\mathcal{D})\, dq$$

$$= \int_0^1 q \cdot \frac{(n+1)!}{k!\,(n-k)!}\, q^k (1-q)^{n-k}\, dq$$

$$= \frac{(n+1)!}{k!\,(n-k)!} \int_0^1 q^{k+1} (1-q)^{n-k}\, dq$$

$$= \frac{(n+1)!}{k!\,(n-k)!} \cdot \frac{(k+1)!\,(n-k)!}{(n+2)!}$$

$$\boxed{\mathbb{P}(H|\mathcal{D}) = \frac{k+1}{n+2}}$$

Laplace rule.

Summary: this example shows how to apply Bayesian inference formula for a parameter with continuous pdf.

We will assume that the 'prior' distribution for q is uniform on $[0,1]$, indicating our complete lack of knowledge. We will indicate this by $U(q)$. Let $\mathcal{D}$ denote the observed sequence of k Heads in n tosses, then the Bayesian update for q is

$$f_1(q|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|q)U(q)}{\int_0^1 dq\mathbb{P}(\mathcal{D}|q)U(q)} \tag{3}$$

Now

$$P(\mathcal{D}|q) = \binom{n}{k} q^k(1-q)^{n-k} \tag{4}$$

and $U(q) = 1$. So the normalization on the right side (aka the evidence) is

$$\int_0^1 dq\mathbb{P}(\mathcal{D}|q)U(q) = \binom{n}{k} \int_0^1 q^k(1-q)^{n-k}dq = \binom{n}{k} \frac{k!(n-k)!}{(n+1)!} \tag{5}$$

So the posterior for q is the Beta distribution

$$f_1(q|\mathcal{D}) = \frac{\binom{n}{k} q^k(1-q)^{n-k}}{\int_0^1 dq\mathbb{P}(\mathcal{D}|q)U(q)} = \frac{(n+1)!}{k!(n-k)!} q^k(1-q)^{n-k}$$

Let H be the event to get Heads on the next toss. Then

$$\mathbb{P}(H|\mathcal{D}) = \int_0^1 \mathbb{P}(H|\mathcal{D}, q)\, f_1(q|\mathcal{D})dq = \int_0^1 \mathbb{P}(H|q)\, f_1(q|\mathcal{D})dq \tag{6}$$

Since $\mathbb{P}(H|q) = q$ we get

$$\begin{aligned} \mathbb{P}(H|\mathcal{D}) &= \frac{(n+1)!}{k!(n-k)!} \int_0^1 q\, q^k (1-q)^{n-k} dq & (7)\\ &= \frac{(n+1)!}{k!(n-k)!} \frac{(k+1)!(n-k)!}{(n+2)!} & (8)\\ &= \frac{k+1}{n+2} & (9) \end{aligned}$$

This is known as Laplace's Rule.

Point estimate: given the posterior distribution $f_1(\theta|\mathcal{D})$, what is the best estimate for parameter θ? Of course there is no single answer to this, but a popular choice is the posterior mean, namely

$$\theta_{mean} = \int \theta \, f_1(\theta|\mathcal{D}) \, d\theta$$

Another popular choice is the maximum likelihood estimator θ_{MLE}: this is the value of θ which maximizes the posterior distribution $f_1(\theta|\mathcal{D})$.

θ_{mean} : expected value of θ using the posterior pdf.

θ_{MLE} : maximum likelihood estimator

: value of θ which maximizes the posterior pdf.

Example: Return to the previous coin tossing example. The posterior is

$$f_1(q|\mathcal{D}) = \frac{(n+1)!}{k!(n-k)!}\, q^k (1-q)^{n-k}$$

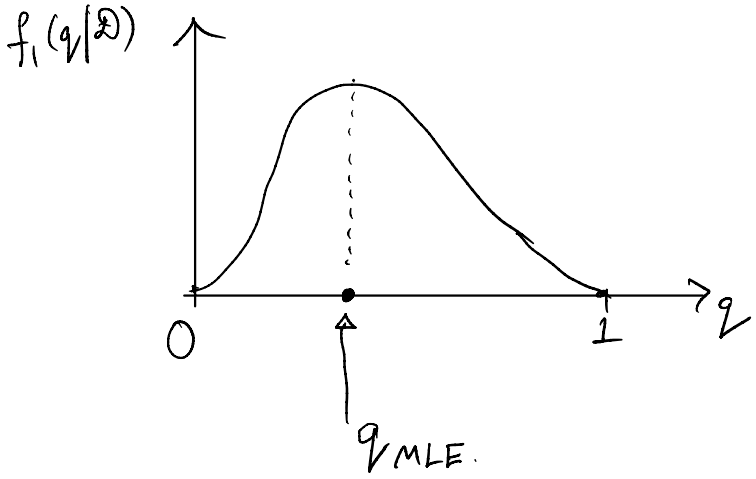
We already computed the mean:

$$q_{mean} = \int_0^1 q\, f_1(q|\mathcal{D})\, dq = \frac{k+1}{n+2}$$

The MLE is found by maximizing $f_1(q|\mathcal{D})$ over q:

$$\frac{d}{dq} f_1(q|\mathcal{D}) = 0 \Leftrightarrow q_{MLE} = \frac{k}{n}$$

Note: when the prior distribution is uniform the estimator θ_{MLE} is the same as the usual MLE computed using the likelihood function; but if the prior is not uniform then it is different in general.



Find MLE : $\frac{d}{dq} f_1(q|\mathcal{D}) = 0$.

$$\frac{d}{dq}\left[q^k (1-q)^{n-k}\right] = k q^{k-1}(1-q)^{n-k} - (n-k) q^k (1-q)^{n-k-1} = 0.$$

$$\Rightarrow \quad k q^{k-1}(1-q)^{n-k} = (n-k) q^k (1-q)^{n-k-1}$$

$$\Rightarrow \quad k(1-q) = (n-k) q$$

$$\Rightarrow \quad \boxed{q_{MLE} = \frac{k}{n}}$$

Mean $q_{mean} = \int_0^1 q \cdot f_1(q|\mathcal{D})\, dq$

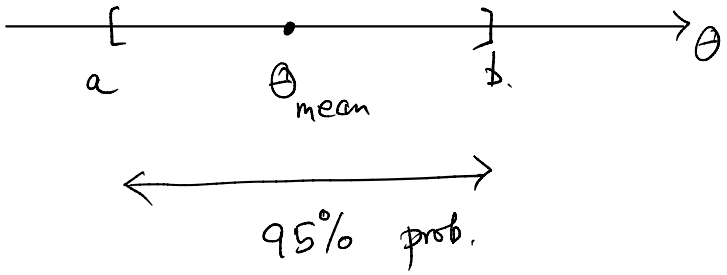
$= \frac{k+1}{n+2}$.

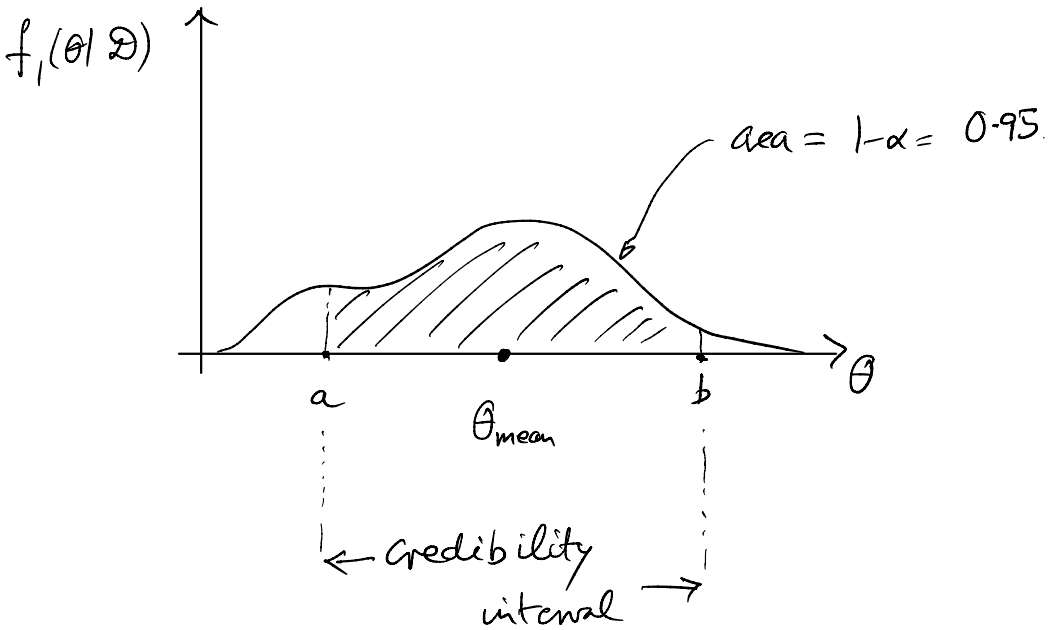
When n is large these are close but not identical.

Credibility interval: going beyond the point estimate, we often want an interval estimate for θ. This is not the same as the confidence interval we encounter in usual hypothesis testing, but they are often quite close. Suppose that θ is a real parameter. We say that the interval (a, b) is a $100(1-\alpha)\%$ credibility interval for θ if

$$\mathbb{P}(a < \theta < b|\mathcal{D}) = 1 - \alpha$$

where the probability on the left side is computed using the posterior pdf $f_1(\theta|\mathcal{D})$. Note that here θ is a random variable, and we are finding numbers a, b so that the probability that θ will lie between a, b is 0.95.





Example: Contrast credibility interval with confidence interval for the previous coin tossing example. Suppose for concreteness that $n = 25$ and $k = 12$. The posterior is

$$f_1(q|\mathcal{D}) = \binom{25}{12}\, 26\, q^{12}\,(1-q)^{13} \quad 0 \le q \le 1$$

The point estimates are

$$q_{mean} = \frac{13}{27} = 0.481, \qquad q_{MLE} = \frac{12}{25} = 0.48$$

Can check that

$$\int_{0.297}^{0.663} f_1(q|\mathcal{D})\, dq = 0.95$$

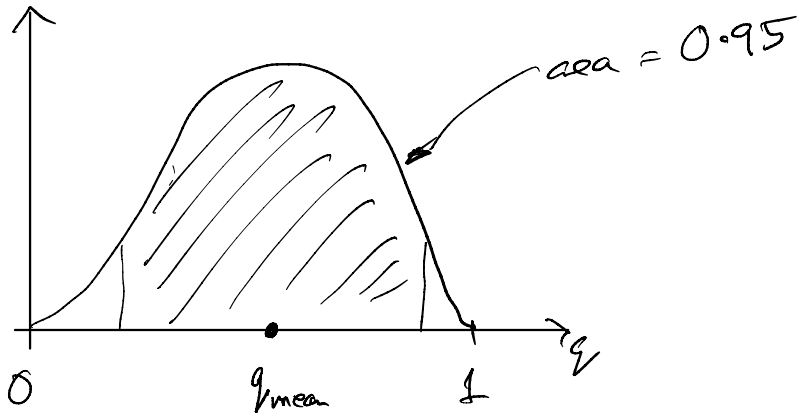
So $(0.297, 0.663)$ is a (symmetric) 95% credibility interval for q given the data $\mathcal{D}$. [This was found by trial and error].

$n=25$: toss coin 25 times

$k=12$: # Heads $= 12$, # Tails $= 13$.

$\Rightarrow$ posterior $f_1(q|\mathcal{D}) = \binom{25}{12}\, 26\, q^{12}(1-q)^{13}$.

$q_{mean} = \frac{k+1}{n+2}$



What would we say about a confidence interval for q given the same data? In this case we would be saying that q is some fixed number q_0 which is unknown to us, and we are trying to estimate q_0 by taking measurements. Our point estimate would be

$$\widehat{q} = \frac{k}{n} = \frac{12}{25} = 0.48$$

To get the 95% confidence interval for q, we would find numbers c, d so that

$$\mathbb{P}(c < \widehat{q} < d) = 0.95$$

The meaning is: if we repeat the measurement many times (ie tossing the coin 25 times and counting how many Heads each time), then in the long-run q_0 will lie in this interval 95% of the time. Now by the usual normal approximation we have

$$\widehat{q} \sim \mathcal{N}(q_0, \frac{\sigma^2}{n}), \quad \sigma^2 = \frac{q_0(1-q_0)}{n}$$

We approximate σ^2 by using $\widehat{q}$ in place of q_0. Then the 95% confidence interval is

$$\widehat{q} \pm 1.96 \left(\frac{\widehat{q}(1-\widehat{q})}{25} \right)^{1/2} = 0.48 \pm 0.196$$

So $(0.284, 0.676)$ is the 95% confidence interval for q_0. Notice that the intervals are different, but close. This is often the case, but there are instances where they can be very different.

Example.

3 strains of COVID in population.

labelled $\{1, 2, 3\}$.

Want to know the proportions in the population:

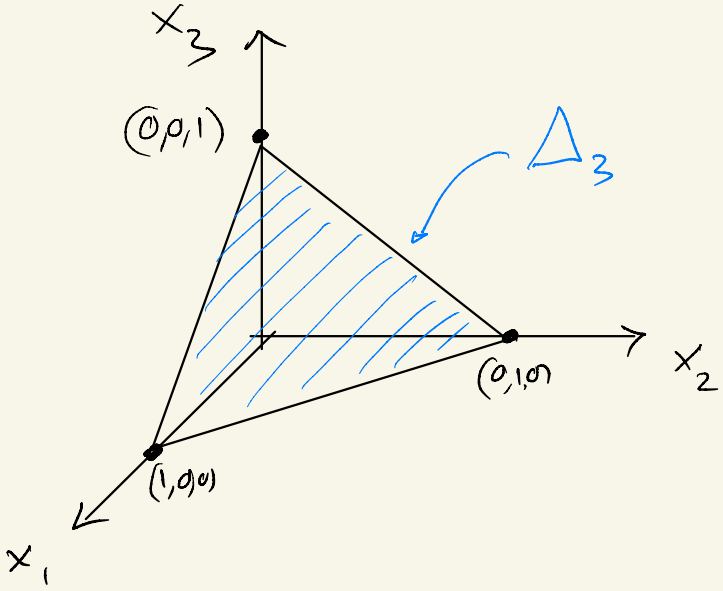
Call the proportions: q_1, q_2, q_3

$$q_1 + q_2 + q_3 = 1.$$

Bayesian inference.

Prior? $f_0(q_1, q_2, q_3) =$ uniform

$$\Delta_3 \subset \mathbb{R}^3 = \{(x_1, x_2, x_3) : 0 \le x_i \le 1, \; x_1 + x_2 + x_3 = 1\}$$



so $q = (q_1, q_2, q_3)$ is a point in Δ_3.

Prior $f_0(q)$ is a functional on Δ_3.

$f_0(q)$ is uniform $\Leftrightarrow$ $f_0(q) =$ constant on Δ_3.

Total probability $= \int_{\Delta_3} f_0(q)\, dq_1\, dq_2\, dq_3$

$= 1.$

$\Rightarrow$ $f_0(q) = 2$ for q in Δ_3.

Collect data. ②

Take 10 measurements and classify the strains of COVID.

$$\mathcal{D} = (1, 3, 3, 2, 3, 1, 2, 1, 3)$$

Likelihood:

$$\mathbb{P}(\mathcal{D}|q) = ?$$

Strains	1	2	3
Prob.	q_1	q_2	q_3
Observed frequency	4	2	4

Measurements $X_1, X_2, \ldots, X_{10}$.

$$P(X_1 = \text{Type } 1) = q_1$$

$$P(X_2 = \text{Type } j) = q_j$$

$$\vdots$$

$$P(\mathcal{D}|q) = q_{X_1} q_{X_2} q_{X_3} \cdots q_{X_{10}}$$

$$= q_1^4 q_2^2 q_3^4 \cdot K.$$

where K is the normalization so that summing over all possible outcomes gives 1.

Evidence: $\mathbb{P}(\mathcal{D}) = \int_{\Delta_3} \mathbb{P}(\mathcal{D}|q)\, f_0(q)\, dq_1 dq_2 dq_3$

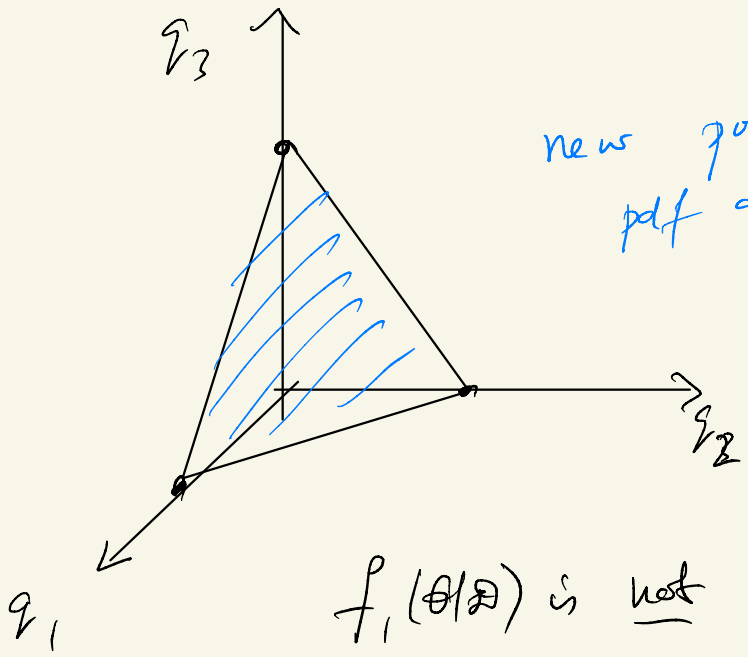
$\Rightarrow$ posterior pdf is

$$f_1(q|\mathcal{D}) = \frac{K q_1^4 q_2^2 q_3^4 \cdot 2}{\mathbb{P}(\mathcal{D})}$$

($f_0(q)$)

$$= 8316\, q_1^4 q_2^2 q_3^4$$

standard integrals.



new posterior pdf on Δ_3

$f_1(\theta|\mathcal{D})$ is not uniform.

Point estimates for $\theta_1, \theta_2, \theta_3$:

$$\theta_{1,\text{mean}} = \frac{5}{13}, \quad \theta_{2,\text{mean}} = \frac{3}{13}, \quad \theta_{3,\text{mean}} = \frac{5}{13}.$$

$$\theta_{1,\text{MLE}} = \frac{4}{10}, \quad \theta_{2,\text{MLE}} = \frac{2}{10}, \quad \theta_{3,\text{MLE}} = \frac{4}{10}.$$

The integrals needed for mean values are these:

$$\int_{\Delta_3} q_1 \, f_1(q_\alpha \mid \mathcal{D}) \, dq = \frac{4+1}{10+3} = \frac{5}{13}$$

$q_1^4 \, q_2^2 \, q_3^4$

Dirichlet distribution

Credibility intervals:

with 95% certainty each of the q_i lie within intervals around these point estimates.