Bayesian statistics problem set 1

Florent Leclercq[a)]
CNRS & Sorbonne Université, UMR 7095, Institut d'Astrophysique de Paris, 98 bis boulevard Arago, F-75014 Paris, France

(Dated: 3 April 2025)

I. BAYESIAN INFERENCE WITH VIROLOGY TESTS

A new virology test is developed to detect a particular viral infection. The test has been evaluated and found to have the following characteristics:

- Sensitivity (true positive rate): 98% (if a patient is infected, the test gives a positive result 98% of the time).
- Specificity (true negative rate): 95% (if a patient is not infected, the test gives a negative result 95% of the time).

Assume that in a given population the prevalence (i.e., the prior probability) of the infection is 2%.

1. What is the probability that a patient is actually infected if they obtain a single positive test result?

We want to compute:

$$p(\text{Infected} \mid \text{Positive}).$$

Bayes' theorem gives:

$$p(\text{Infected} \mid \text{Positive}) = \frac{p(\text{Positive} \mid \text{Infected})\, p(\text{Infected})}{p(\text{Positive})},$$

where

$$p(\text{Positive}) = p(\text{Positive} \mid \text{Infected})\, p(\text{Infected}) + p(\text{Positive} \mid \text{Not Infected})\, p(\text{Not Infected}).$$

Given:

- $p(\text{Positive} \mid \text{Infected}) = 0.98$ (Sensitivity),
- Since Specificity $= 95\%$, the false positive rate is $p(\text{Positive} \mid \text{Not Infected}) = 1 - 0.95 = 0.05$,
- $p(\text{Infected}) = 0.02$,
- $p(\text{Not Infected}) = 1 - 0.02 = 0.98$.

Now, calculate $p(\text{Positive})$:

$$p(\text{Positive}) = (0.98 \times 0.02) + (0.05 \times 0.98) = 0.0196 + 0.049 = 0.0686.$$

Then, the posterior probability is:

$$p(\text{Infected} \mid \text{Positive}) = \frac{0.98 \times 0.02}{0.0686} \approx \frac{0.0196}{0.0686} \approx 0.2856 \quad (\text{or } 28.6\%).$$

2. Suppose the patient receives two independent positive test results. Given that the tests are independent and use the same sensitivity and specificity as above, what is now the probability that the patient is infected?

Step 1: Update the probability after the first positive test. From question 1, after one positive test we have:

$$p(\text{Infected} \mid \text{Positive}_1) \approx 0.2856.$$

This becomes our new prior for the second test.

[a)]Electronic mail: florent.leclercq@iap.fr; https://www.florent-leclercq.eu/; ORCID: 0000-0002-9339-1404

Step 2: Update with the second independent positive test. Using Bayes' rule again with the updated prior:

- New prior $p(\text{Infected}) = 0.2856$,
- $p(\text{Not Infected}) = 1 - 0.2856 = 0.7144$,
- Likelihoods remain: $p(\text{Positive} \mid \text{Infected}) = 0.98$, $p(\text{Positive} \mid \text{Not Infected}) = 0.05$.

Now, compute:

$$p(\text{Infected} \mid \text{Positive}_2, \text{Positive}_1) = \frac{0.98 \times 0.2856}{0.98 \times 0.2856 + 0.05 \times 0.7144} \approx \frac{0.2799}{0.3156} \approx 0.886 \quad (\text{or } 88.6\%).$$

3. Consider now that the patient is tested twice, but the test results are discordant (one positive and one negative, in any order). What is the probability that the patient is infected given one positive and one negative result?

We can assume that the first test is positive and the second test is negative.
Step 1: Update the probability after the first positive test. As before:

$$p(\text{Infected} \mid \text{Positive}) \approx 0.2856.$$

Step 2: Update with the second test (negative result). For a negative test, we have:

- $p(\text{Negative} \mid \text{Infected}) = 1 - \text{Sensitivity} = 1 - 0.98 = 0.02$,
- $p(\text{Negative} \mid \text{Not Infected}) = \text{Specificity} = 0.95$.

Now, applying Bayes' theorem:

$$p(\text{Infected} \mid \text{P then N}) = \frac{p(\text{N} \mid \text{Infected})\, p(\text{Infected} \mid \text{P})}{p(\text{N} \mid \text{Infected})\, p(\text{Infected} \mid \text{P}) + p(\text{N} \mid \text{Not Infected})\, p(\text{Not Infected} \mid \text{P})}.$$

Using the updated prior after the positive result:

- $p(\text{Infected} \mid \text{Positive}) = 0.2856$,
- $p(\text{Not Infected} \mid \text{Positive}) = 1 - 0.2856 = 0.7144$.

Plug in the numbers:

$$p(\text{Infected} \mid \text{P then N}) = \frac{0.02 \times 0.2856}{0.02 \times 0.2856 + 0.95 \times 0.7144} \approx \frac{0.005712}{0.684392} \approx 0.00834 \quad (\text{or } 0.83\%)$$

Interpretation: A discordant result dramatically reduces the probability of infection, even below the original base rate of 2%, because a negative result (with high specificity) is strong evidence against infection when weighted against an uncertain positive.

4. Discuss how the prevalence (prior probability) of the infection affects the interpretation of test results. For example, if the prevalence increases to 20%, recalculate the posterior for a positive test result using the same sensitivity and specificity.

When the prevalence (or base rate) of a disease increases, the same test characteristics (sensitivity and specificity) will yield a higher posterior. Let's redo the calculation for a single positive test when the prevalence changes from 2% to 20%. Now, assume:

- $p(\text{Infected}) = 0.20$,
- $p(\text{Not Infected}) = 1 - 0.20 = 0.80$,
- Sensitivity remains 0.98,
- False positive rate remains 0.05 (since Specificity is 95%).

We compute $p(\text{Positive})$:

$$p(\text{Positive}) = (0.98 \times 0.20) + (0.05 \times 0.80) = 0.196 + 0.04 = 0.236.$$

Now, using Bayes' theorem:

$$p(\text{Infected} \mid \text{Positive}) = \frac{0.98 \times 0.20}{0.236} = \frac{0.196}{0.236} \approx 0.8305 \quad (\text{or } 83.1\%).$$

Interpretation: When the prevalence is higher (20% instead of 2%), a positive test result is much more likely to indicate a true infection. This example highlights one of the key lessons in Bayesian inference: the prior probability (or base rate) matters a great deal when interpreting diagnostic tests.
This exercise demonstrates that Bayesian inference provides a systematic method to update your beliefs (or probabilities) in light of new evidence (test outcomes). It also shows that the usefulness of a diagnostic test depends not only on its sensitivity and specificity but also on the underlying prevalence of the disease in the population.

II. THE MONTY HALL PROBLEM

Solve the "Monty Hall" problem given in the lectures, using Bayes' theorem.

You are a contestant on a game show, and you are presented with three closed doors. Behind one of the doors is a brand new car, while the other two doors have goats behind them. You have no knowledge of which door has the car and which doors have the goats.

You get to choose one of the three doors, but before it is opened, the game show host (Monty Hall) opens one of the other two doors and shows you that it has a goat behind it. Now, you have the option to stick with your original choice or switch to the remaining unopened door. What is the probability that the car is behind the door you originally chose, and what is the probability that the car is behind the other unopened door?

Let the doors be labelled a, b, c, where a is the door you choose initially, and b is the door which is opened. Many, if not all, of the probabilities below should be interpreted as "given that you have chosen a," but for clarity we won't write this explicitly.
Let $p(a)$ = probability that a leads to the car, etc. Let B be the event that door b gets opened and leads to a goat.
What you want is the probability that a leads to the car, given that b is opened and leads to a goat. i.e., the aim is to calculate

$$p(a|B). \tag{1}$$

We can use Bayes' theorem for this:

$$p(a|B) = \frac{p(a,B)}{p(B)} = \frac{p(B|a)p(a)}{p(B)}. \tag{2}$$

Now, clearly $p(a) = p(b) = p(c) = \frac{1}{3}$ (all doors are equally likely, before any experiment is done).
$p(B|a)$ = probability that door b is opened, given that a leads to the car. Evidently

$$p(B|a) = \frac{1}{2}. \tag{3}$$

Monty Hall could have opened either door b or c, since they both lead to a goat.
What about $p(B)$? It is the sum of all the joint probabilities:

$$p(B) = p(B,a) + p(B,b) + p(B,c) = p(B|a)p(a) + p(B|b)p(b) + p(B|c)p(c), \tag{4}$$

each of which we can calculate. $p(a) = p(b) = p(c) = \frac{1}{3}$, as before, and $p(B|a) = \frac{1}{2}$. Now:

$$p(B|b) = 0 : \tag{5}$$

Monty Hall will not open b since it leads to the car in this case.
$p(B|c)$ is the most interesting. Given that you have chosen a (remember this is implicit throughout), then if c leads to the car, then Monty Hall must open door b, i.e.,

$$p(B|c) = 1. \tag{6}$$

So the probability that your original choice a leads to the car is

$$p(a|B) = \frac{p(B|a)p(a)}{p(B|a)p(a) + p(B|b)p(b) + p(B|c)p(c)} \tag{7}$$

$$= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + (0 \times \frac{1}{3}) + (1 \times \frac{1}{3})} = \frac{1}{3}.$$

So you would double your chances (from $\frac{1}{3}$ to $\frac{2}{3}$) if you switch to the other door.

III. THE DOMINANCE OF THE LIKELIHOOD IN BAYESIAN INFERENCE

Suppose you have a coin with an unknown probability of heads, θ. You decide to model the coin toss outcomes as independent Bernoulli trials and use Bayesian inference to learn θ. Assume a Beta prior for θ:

$$\theta \sim \text{Beta}(a, b). \tag{8}$$

Because the Beta distribution is a conjugate prior for the Bernoulli likelihood, the posterior will also be a Beta distribution. We will demonstrate that for small n the choice of prior matters, but as n grows large, the data "overwhelms" the prior.

1. Derivation of the posterior distribution
   (a) Write down the likelihood function for k heads in n tosses.
   (b) Given the Beta(a, b) prior for θ, derive the expression for the posterior distribution $p(\theta|\text{data})$.

1. (a) The likelihood for k heads and $n-k$ tails is

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

(b) The Beta prior is given by

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

Multiplying the prior by the likelihood, we obtain:

$$p(\theta|\text{data}) \propto \theta^{a-1}(1-\theta)^{b-1} \cdot \theta^k(1-\theta)^{n-k} = \theta^{a+k-1}(1-\theta)^{b+n-k-1}.$$

Thus, the posterior is a Beta distribution:

$$\theta|\text{data} \sim \text{Beta}(a+k,\, b+n-k).$$

2. Small sample simulation
   Assume that the true coin bias is $\theta_{\text{true}} = 0.7$ and that you simulate $n = 10$ coin tosses. Consider three different priors:

   - **Prior A**: Weakly informative prior Beta$(2, 2)$
   - **Prior B**: Beta$(5, 5)$
   - **Prior C**: Informative prior Beta$(20, 20)$

   Perform the following tasks:

   (a) Simulate 10 coin tosses (use a random seed for reproducibility).
   (b) For each prior, compute the posterior parameters.
   (c) Plot the posterior distributions on the same graph.
   (d) Discuss the influence of the prior when the sample size is small.

Below is one complete Python code snippet to perform these tasks:

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import beta

# --- Simulation Setup ---
np.random.seed(42)  # For reproducibility

theta_true = 0.7  # True probability of heads
n = 10            # Number of coin tosses

# Simulate coin tosses: 1 indicates heads, 0 indicates tails.
data = np.random.binomial(1, theta_true, size=n)
k = np.sum(data)
print(f"Small Sample: {k} heads out of {n} tosses")

```

```python
# Define the priors as (a, b) tuples
priors = {
    'Beta(2,2)': (2, 2),
    'Beta(5,5)': (5, 5),
    'Beta(20,20)': (20, 20)
}

# Define a grid of theta values for plotting
theta_vals = np.linspace(0, 1, 1000)

# --- Plotting the Posterior Distributions ---
plt.figure(figsize=(10, 6))

for index, (label, (a, b)) in enumerate(priors.items()):
    # Posterior parameters: a_post = a + k, b_post = b + n - k
    a_post = a + k
    b_post = b + n - k
    prior_pdf = beta.pdf(theta_vals, a, b)
    posterior_pdf = beta.pdf(theta_vals, a_post, b_post)
    plt.plot(theta_vals, prior_pdf, ls="--", label=f"Prior {label}", color=f"C{index}")
    plt.plot(theta_vals, posterior_pdf, label=f"{label} -> Beta({a_post},{b_post})", color=
    f"C{index}")


likelihood_pdf = theta_vals**k * (1 - theta_vals)**(n - k)
likelihood_pdf /= np.trapz(likelihood_pdf, theta_vals)
plt.plot(theta_vals, likelihood_pdf, ls=":", label="Likelihood", color="black")

plt.title("Posterior Distributions with n = 10")
plt.xlabel(r"$\theta$")
plt.ylabel("Density")
plt.legend()
plt.show()
```

Discussion: With only 10 coin tosses, the observed data are relatively scarce. You should notice that the posteriors differ noticeably among the three priors. For example, the informative prior Beta$(20, 20)$ (centred around 0.5) tends to "pull" the posterior toward 0.5, even if the data suggest a higher value. In contrast, the uniform prior (Beta$(2, 2)$) is more flexible, resulting in a posterior that reflects the data more directly. This shows that with small n, the choice of prior has a significant impact.

3. Large sample simulation
   Now, simulate $n = 1000$ coin tosses using the same true coin bias $\theta_{\text{true}} = 0.7$ and repeat the analysis using the same three priors:

   - **Prior A**: Weakly informative prior Beta$(2, 2)$
   - **Prior B**: Beta$(5, 5)$
   - **Prior C**: Informative prior Beta$(20, 20)$

   Perform the following:

   (a) Simulate 1000 coin tosses.
   (b) Compute the posterior parameters for each prior.
   (c) Plot the posterior distributions on the same graph.
   (d) Discuss how and why the influence of the prior changes with the larger sample size.

Below is the corresponding Python code:

```python
# --- Large Sample Simulation ---
n_large = 1000  # Large number of tosses
data_large = np.random.binomial(1, theta_true, size=n_large)
k_large = np.sum(data_large)
print(f"Large Sample: {k_large} heads out of {n_large} tosses")

# --- Plotting the Posterior Distributions ---
plt.figure(figsize=(10, 6))

for index, (label, (a, b)) in enumerate(priors.items()):
    # Posterior parameters: a_post = a + k, b_post = b + n - k
    a_post = a + k_large
    b_post = b + n_large - k_large
    prior_pdf = beta.pdf(theta_vals, a, b)
    posterior_pdf = beta.pdf(theta_vals, a_post, b_post)
```

```
    plt.plot(theta_vals, prior_pdf, ls="--", label=f"Prior {label}", color=f"C{index}")
    plt.plot(theta_vals, posterior_pdf, label=f"{label} -> Beta({a_post},{b_post})", color=
    f"C{index}")

likelihood_pdf = theta_vals**k_large * (1 - theta_vals)**(n_large - k_large)
likelihood_pdf /= np.trapz(likelihood_pdf, theta_vals)
plt.plot(theta_vals, likelihood_pdf, ls=":", label="Likelihood", color="black")

plt.title("Posterior Distributions with n = 1000")
plt.xlabel(r"$\theta$")
plt.ylabel("Density")
plt.legend()
plt.show()
```

Discussion: With 1000 tosses, the likelihood (i.e., the data) becomes much more informative. Despite the different starting priors, the resulting posteriors will be very similar, all concentrating sharply around the true value ($\theta \approx 0.7$). This demonstrates that as the sample size increases, the likelihood "dominates" the inference and the influence of the prior becomes negl

4. Comparing posterior means and variances
   For both the small sample ($n = 10$) and large sample ($n = 1000$) cases, compute the posterior mean and variance for each prior. Recall that for a Beta(α, β) distribution:

   - The mean is: $\mu = \dfrac{\alpha}{\alpha + \beta}$
   - The variance is: $\sigma^2 = \dfrac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$

   Write Python code to compute these quantities and observe how they converge with more data.

Here is the Python code that computes the posterior mean and variance for each prior:

```
def posterior_stats(a, b, k, n):
    # Compute the posterior parameters
    a_post = a + k
    b_post = b + n - k
    # Compute the posterior mean and variance
    mean = a_post / (a_post + b_post)
    variance = (a_post * b_post) / ((a_post + b_post)**2 * (a_post + b_post + 1))
    return mean, variance

print("Posterior Statistics for Small Sample (n = 10):")
for label, (a, b) in priors.items():
    mean, var = posterior_stats(a, b, k, n)
    print(f"{label}: Mean = {mean:.4f}, Variance = {var:.6f}")

print("\nPosterior Statistics for Large Sample (n = 1000):")
for label, (a, b) in priors.items():
    mean, var = posterior_stats(a, b, k_large, n_large)
    print(f"{label}: Mean = {mean:.4f}, Variance = {var:.8f}")
```

Expected results and discussion:

- Small sample ($n = 10$): The posterior means and variances will differ noticeably across the three priors. For example, the informative prior Beta$(20, 20)$ might yield a mean that is pulled toward 0.5 compared to the uniform prior.
- Large sample ($n = 1000$): The posterior means will be very close to the true value (approximately 0.7) and the variances will be very small, with only minor differences among the different priors. This numerical evidence reinforces the conclusion that as more data are accumulated, the influence of the prior fades away.

This exercise shows that while the prior can have a strong influence when data are scarce (as seen with $n = 10$), its effect diminishes as the amount of data increases (as seen with $n = 1000$). In the large-sample limit, the posterior is dominated by the likelihood, ensuring that the inference about θ is driven primarily by the observed data rather than the prior beliefs. This is a key idea in Bayesian statistics and is sometimes referred to as "Bayesian consistency."

Feel free to experiment further by modifying the priors or the true value of θ to see how robust this phenomenon is under different settings.

IV. IGNORANCE PRIORS FOR AN URN PROBLEM

Suppose we have an urn containing red and white balls. The probability of drawing a red ball is an unknown parameter θ (with $\theta \in [0,1]$). In the absence of any prior information about θ, we might wish to assign a prior probability density $\pi(\theta)$ that represents "complete ignorance." However, as Jaynes famously argued, the idea of "ignorance" depends on how one parameterises the problem. In what follows we will explore different invariance requirements for an "uninformative" prior and see that they lead to different answers.

1. Invariance under label exchange
   Because the labels "red" and "white" are arbitrary, one might argue that our state of ignorance should be invariant under exchanging the two colors.

   (a) Explain why this invariance condition seems natural for representing ignorance, and write the functional equation that $\pi(\theta)$ must satisfy.

   (b) Verify that the uniform prior

   $$\pi(\theta) = 1, \quad 0 < \theta < 1, \tag{9}$$

   satisfies this invariance.

   (c) Show that the uniform prior on $[0,1]$ is also the maximum entropy prior.

(a) Since we have no reason to favour red over white (or vice versa), our prior should not change if we "flip" the interpretation of θ (red probability) to $1-\theta$ (white probability). Hence, any candidate prior $\pi(\theta)$ should satisfy

$$\pi(\theta) = \pi(1-\theta) \quad \text{for all } \theta \in [0,1].$$

(b) For the uniform prior, $\pi(\theta) = 1$ for all $\theta \in [0,1]$. Then

$$\pi(1-\theta) = 1 = \pi(\theta),$$

so the uniform prior is symmetric under $\theta \to 1-\theta$.

(c) To find the maximum entropy prior, we maximise:

$$H(\pi) = -\int_0^1 \pi(\theta) \log \pi(\theta)\, \mathrm{d}\theta. \tag{10}$$

Since $\pi(\theta)$ is a probability density function, it must satisfy the normalisation condition:

$$\int_0^1 \pi(\theta)\, \mathrm{d}\theta = 1. \tag{11}$$

We introduce a Lagrange multiplier λ to enforce this constraint and define the functional (the Lagrangian):

$$\mathcal{L} \equiv -\int_0^1 \pi(\theta) \log \pi(\theta)\, \mathrm{d}\theta + \lambda \left(\int_0^1 \pi(\theta)\, \mathrm{d}\theta - 1 \right). \tag{12}$$

Taking the functional derivative of $\mathcal{L}$ with respect to $\pi(\theta)$ and setting it to zero gives:

$$\frac{\delta \mathcal{L}}{\delta \pi(\theta)} = -(1 + \log \pi(\theta)) + \lambda = 0. \tag{13}$$

Solving for $\pi(\theta)$ gives $\pi(\theta) = \mathrm{e}^{\lambda - 1} = \text{const.}$ for all θ; and the normalisation constraint fixes $\pi(\theta) = 1$.

2. Scale invariance of the odds
   An alternative idea is to require that our state of ignorance be invariant under reparameterisation. A common reparameterisation is in terms of the *odds*:

   $$\phi = \frac{\theta}{1-\theta}, \quad \text{with inverse } \theta = \frac{\phi}{1+\phi}. \tag{14}$$

   Since $\phi \in [0, \infty)$, one might require that ignorance about ϕ be expressed by a prior that is *scale invariant*. In other words, if we "rescale" the odds by a positive constant, our ignorance should remain the same.

(a) Functional equation for scale invariance
Assume that the prior density for ϕ, called $\pi_\phi(\phi)$, satisfies the following invariance property for any scaling factor $a > 0$:

$$\pi_\phi(a\phi) = \frac{1}{a}\,\pi_\phi(\phi). \tag{15}$$

Show that (up to a multiplicative constant) the unique solution of this functional equation is

$$\pi_\phi(\phi) \propto \frac{1}{\phi}. \tag{16}$$

(b) Transforming back to the θ parameter
The densities $\pi(\theta)$ and $\pi_\phi(\phi)$ are related by the usual change-of-variables formula:

$$\pi(\theta) = \pi_\phi(\phi)\left|\frac{\mathrm{d}\phi}{\mathrm{d}\theta}\right|. \tag{17}$$

i. Compute $\dfrac{\mathrm{d}\phi}{\mathrm{d}\theta}$.
ii. Express $\pi(\theta)$ in terms of θ by using the result from question (a).

(a) Let $\pi_\phi(\phi)$ be a positive function on $[0, \infty)$ satisfying

$$\pi_\phi(a\phi) = \frac{1}{a}\,\pi_\phi(\phi)$$

for every $a > 0$ and $\phi > 0$. A standard method is to set

$$g(\phi) = \phi\,\pi_\phi(\phi).$$

Then, replacing ϕ by $a\phi$ in g we get

$$g(a\phi) = a\phi\,\pi_\phi(a\phi) = a\phi\,\frac{1}{a}\,\pi_\phi(\phi) = \phi\,\pi_\phi(\phi) = g(\phi).$$

Thus, $g(a\phi) = g(\phi)$ for all $a > 0$ and all ϕ; that is, g is constant. Denote this constant by C. Then

$$\phi\,\pi_\phi(\phi) = C \quad \Longrightarrow \quad \pi_\phi(\phi) = \frac{C}{\phi}.$$

So, up to the overall multiplicative constant C, we have

$$\pi_\phi(\phi) \propto \frac{1}{\phi}.$$

(b) i. Compute the derivative:

$$\phi = \frac{\theta}{1-\theta} \quad \Longrightarrow \quad \frac{\mathrm{d}\phi}{\mathrm{d}\theta} = \frac{(1-\theta) - (-1)\theta}{(1-\theta)^2} = \frac{1}{(1-\theta)^2}.$$

ii. Now, since we found $\pi_\phi(\phi) \propto \frac{1}{\phi}$ and noting that $\phi = \theta/(1-\theta)$, we have

$$\pi(\theta) \propto \frac{1}{\phi} \cdot \frac{1}{(1-\theta)^2}.$$

But

$$\frac{1}{\phi} = \frac{1-\theta}{\theta}.$$

Thus,

$$\pi(\theta) \propto \frac{1-\theta}{\theta} \cdot \frac{1}{(1-\theta)^2} = \frac{1}{\theta\,(1-\theta)}.$$

3. Jeffreys prior
The Jeffreys prior is defined by

$$\pi_J(\theta) \propto \sqrt{|I(\theta)|}, \tag{18}$$

i.e. its density is proportional to the square root of the determinant of the Fisher information matrix $I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log p(r|\theta)\right)^2\right]$.

(a) Compute the Fisher information for the binomial likelihood of our problem,

$$p(r|\theta) = \binom{N}{r}\theta^r\,(1-\theta)^{N-r}. \tag{19}$$

Show that

$$\pi_J(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}}. \tag{20}$$

(b) Compare the three candidate priors for θ (the uniform prior, the odds-invariance prior, and the Jeffreys prior). Discuss what these differences imply about the notion of a "noninformative" or "ignorance" prior.

(a) A practical way to compute the Fisher information $I(\theta)$ is to use the identity

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log p(r|\theta)\right)^2\right] = -E\left[\frac{\partial^2}{\partial\theta^2}\log p(r|\theta)\right]. \tag{21}$$

The log-likelihood function is

$$\log p(r|\theta) = \log\binom{N}{r} + r\log\theta + (N-r)\log(1-\theta).$$

Differentiating the log-likelihood with respect to θ gives:

$$\frac{\partial}{\partial\theta}\log p(r|\theta) = \frac{r}{\theta} - \frac{N-r}{1-\theta}.$$

$$\frac{\partial^2}{\partial\theta^2}\log p(r|\theta) = -\frac{r}{\theta^2} - \frac{N-r}{(1-\theta)^2}.$$

We take the expectation and use $E[r] = N\theta$ (because r is binomially distributed) to find:

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log p(r|\theta)\right] = \frac{N\theta}{\theta^2} + \frac{N(1-\theta)}{(1-\theta)^2} = \frac{N}{\theta} + \frac{N}{1-\theta} = \frac{N}{\theta(1-\theta)}.$$

Therefore, we have

$$\pi_J(\theta) \propto \sqrt{I(\theta)} \propto \frac{1}{\sqrt{\theta(1-\theta)}},$$

(with the factor $\sqrt{N}$ absorbed into the constant of proportionality).

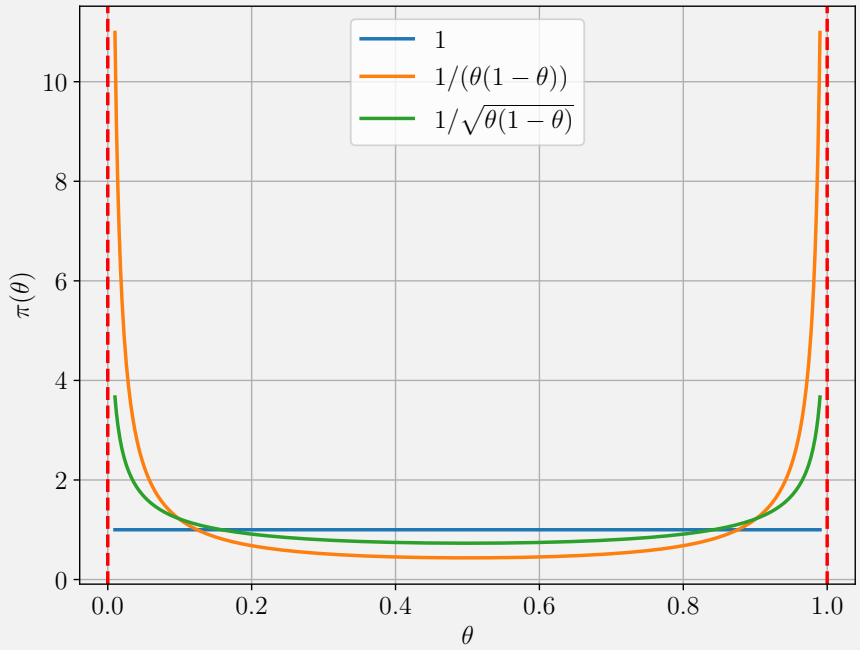
(b) The uniform prior on θ is "flat" on the interval $[0,1]$. In contrast, the odds-invariance argument leads us to a prior for θ that is

$$\pi(\theta) \propto \frac{1}{\theta(1-\theta)}, \tag{22}$$

which puts relatively more weight near $\theta \approx 0$ and $\theta \approx 1$ compared to the uniform prior. The Jeffreys prior is less "extreme," namely,

$$\pi_J(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}}.$$

The three differ substantially—especially near the endpoints $\theta \to 0$ and $\theta \to 1$.



This exercise thus shows that the concept of a "noninformative" or "ignorance" prior is not unique; it depends on the parameterisation: demanding invariance under color exchange suggests symmetry (and the uniform prior satisfies this), while demanding invariance on the odds scale (a form of scale invariance) leads to a very different answer. In short, there is no unique "noninformative" prior—different requirements lead to different choices.

ACKNOWLEDGEMENTS

I thank Benjamin Wandelt, Alan Heavens and colleagues of the SOC of the ICIC Data Analysis workshop 2021 for their own lectures, from which this document is inspired.