CSSS/STAT 564: Problem Set 1

1. First problem

1.1. Background

The Bernoulli distribution is the foundation of many models of data used throughout the social sciences because it models the relative frequency of binary outcomes (or indicator variables):

$$[1.1] \quad Y_i = y \in \mathcal{Y} = \{0, 1\}$$

where Y is a placeholder for the name of an indicator variable; i is a placeholder for the index of an observed or hypothetical case in a sample; y is a placeholder for the value of Y observed for the ith individual; $\mathcal{Y}$ is the domain of variable Y; and $\{0, 1\}$ is the set of all possible values that constitute $\mathcal{Y}$. By definition, any indicator must follow the Bernoulli distribution,

$$[1.2] \quad Y_i \sim \text{Bernoulli}\,(\theta_i)$$

where θ_i is a Bernoulli or probability-of-success parameter for individual i, i.e. $\theta_i = \Pr(Y_i = 1)$, whose domain is

$$[1.3] \quad \theta \in \Theta = [0, 1]$$

The Bernoulli distribution is a key building block of two widely used models in quantitative research: the binomial model, and the logistic regression model.

1.1.1. The binomial model

The binomial distribution is a convolution of Bernoulli distributions under the condition that the Bernoulli parameter θ is constant between observations:

$$[1.4] \quad \{Y_1, Y_2, \ldots, Y_n\} \overset{iid}{\sim} \text{Bernoulli}\,(\theta) \Rightarrow \left[K = \sum_{i=1}^{n} Y_i\right] \sim \text{Binomial}\,(\theta, n)$$

where K is a "count-of-successes" variable, the Bernoulli parameter θ serves as a shape parameter governing the binomial model's direction and degree of skewness, and n serves as a scale parameter for the binomial distribution.

If we wish to fit the Bernoulli distribution to a sample of *iid* indicator variables with a known size n, our inferential target is estimation of the unknown parameter θ. Consequently we require a prior distribution suitable for the domain Θ. The beta distribution is an ideal candidate because the support of this model matches the domain of θ:

$$[1.5] \quad \theta \sim \text{Beta}\,(a, b)$$

where both a and b double as shape (skewness) and concentration parameters for the model. a pushes the concentration of probability up toward 1, while b pushes it down toward 0. The higher the value of either parameter, the more concentrated the beta distribution becomes overall, while the location of this concentration depends on the balance of a relative to b. The distribution will be symmetrical in the case that $a = b$, left-skewed in the case that $a > b$, or right-skewed in the case that $a < b$.

Two popular parameter valuations for this prior model are a "flat" beta prior—

$$[1.6] \quad \theta \sim \text{Beta}\,(a = 1, b = 1)$$

which is equivalent to

$$[1.7] \quad \theta \sim \text{Uniform}\,(0, 1)$$

—and a "Jeffreys" minimally informative prior—

$$[1.8] \quad \theta \sim \text{Beta}\,(a = 0.5, b = 0.5)$$

(The role of flat and minimally informative priors will be discussed further below).

1.1.2. The logistic regression model

Logistic regression is often used to model the dependence of binary outcomes on predictors of interest. Logistic regression poses the dependence of an indicator variable Y on one or more predictors X by framing the Bernoulli parameter as a logit-linear function, i.e. a logistic transformation of a linear predictor:

$$[1.9] \quad \phi_i = \text{logit}\,(\theta_i) = \ln\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \sum_{j=1}^{J} x_{ij}\beta_j = \mathbf{x_i}\boldsymbol{\beta}$$

where

- α is the model's y-intercept
- j is a placeholder for the index of a particular predictor
- J is a placeholder for the number of predictors included in the model
- x_{ij} is the value of predictor j observed for individual i
- β_j is the regression coefficient for the jth predictor
- $\mathbf{x_i}\boldsymbol{\beta}$ is an abbreviation of the linear predictor $\alpha + \sum_{j=1}^{J} x_{ij}\beta_j$

Equivalently, the odds ratio—the number of expected successes per failure—is a log-linear function of the model predictors,

$$[1.10] \quad OR_i = \frac{\theta_i}{1-\theta_i} = e^{\mathbf{x_i}\boldsymbol{\beta}} = e^{\alpha} \times \prod_{j=1}^{J} \left(e^{\beta_j}\right)^{x_{ij}}$$

where e^{α} is the baseline or y-intercept of the odds ratio, which changes multiplicatively by the factor e^{β_j} for each unit increase in x_j.

The inverse of the logit function is the logistic function (from which logistic regression gets its name). The logistic transformation of the linear predictor ϕ_i yields the relationship between the the model's predictors and regression coefficients on one hand and the Bernoulli parameter θ_i on the other:

$$[1.11] \quad \theta_i = \text{logistic}\,(\phi_i) = \frac{1}{1+e^{-\phi_i}} = \frac{1}{1+e^{-\mathbf{x_i}\boldsymbol{\beta}}}$$

An intercept-only logistic regression model omits predictors, implying a shared odds ratio and a shared probability of success for all observations:

$$[1.12] \quad OR = e^{\alpha}$$

$$[1.13] \quad \theta = \text{logistic}\,(\alpha) = \frac{1}{1+e^{-\alpha}}$$

Consequently the index i has disappeared from the odds ratio and Bernoulli parameter.

In the case of intercept-only logistic regression models, the only prior distribution we require is for α, whose domain is the entire real number line. While many Bayesian textbooks apply a normal model as a default prior for α, Gelman et al. (2008; http://dx.doi.org/10.1214/08-AOAS191) recommend an alternative default: the Cauchy distribution whose location parameter μ (the model's median and mode) is set to 0 and whose scale parameter ψ is set to 10:

$$[1.14] \quad \alpha \sim \text{Cauchy}\,(\mu = 0, \psi = 10)$$

which is equivalent to a three-parameter Student's t distribution whose shape parameter (ν or "degrees of freedom"), location parameter μ, and dispersion parameter ψ are set to 1, 0, and 10, respectively:

$$[1.15] \quad \alpha \sim \text{StudentT}\,(\nu = 1, \mu = 0, \psi = 10)$$

Gelman et al. regard this to be a very weakly informative prior.

Once stripped of predictors, the intercept-only logistic regression model is nothing other than a reparameterized Bernoulli distribution of data, replacing the standard Bernoulli parameter θ with logistic(α). Consequently, the beta and Cauchy priors described above ultimately embody alternative prior specifications for the same model.

As a general rule, default priors are intended to be "weakly informative"—often labeled "vague" or "diffuse"—allowing data to dominate the shape of posterior belief but sometimes introducing just enough prior information to avoid implausible posterior inferences. (So-called "uninformative" priors abstain from this latter role, maximizing the influence of data on posterior inference without imposing safeguards against implausible conclusions.) Default priors are especially useful in contexts of exploratory data analysis (EDA), when little or no background knowledge exists about the data generating process (DGP) under consideration.

While the general spirit of default priors is relatively easy to articulate, identifying models for priors that actually accomplish these goals is less straightforward. As we have just discovered for the Bernoulli distribution, alternative default priors have been recommended depending on whether we are focused on the standard θ parameterization or the reparameterization of implied by an α-only logistic regression model. Two important questions thus arise: (1) How can we compare how weakly informative these defaults are, given the different supports of the parameter spaces $\Theta = [0, 1]$ and $\text{A} = \mathbb{R}$? (2) Which of these defaults is better suited to particular research contexts?

1.2. The questions

Your goal for this exercise is to use Monte Carlo (MC) simulation to compare the forms of the flat beta prior (i.e. uniform prior) for θ, the Jeffreys beta prior for θ, and Gelman et al.'s default Cauchy prior for α, with the goal of of exploring what each implies about prior belief regarding the model's single, unknown parameter. Recall that the distributional characteristics of a sample generated through MC simulation will approach the underlying probability distribution as the size of the simulated sample, $S \to \infty$.

Question 1.1.

- Set your seed using the `set.seed()` function and the seed 587.
- Simulate a large number of "guesses" ($S = 1,000,000$) at the value of θ from
  - the flat beta prior distribution;
  - the Jeffreys beta prior distribution.
- Generate the same number of "guesses" at the value of α from the Gelman default Cauchy prior.
- Once you have generated MC approximations of all three prior distributions, plot all three on the same plot on the domain of θ.
- Also plot all three sample distributions on the same plot on the domain of α.

In order to draw these plots, you will need to map your two distributions of θ onto the domain of α with the logit function, as well as map your distribution of α onto the domain of θ with the logistic function.

$$[1.16] \quad \alpha^{(s)} = \text{logit}\left(\theta^{(s)}\right) \iff \theta^{(s)} = \text{logistic}\left(\alpha^{(s)}\right)$$

For your convenience, some R packages include `logit()` and `logistic()` functions, relieving you of the need to calculate them by hand. The package I like to use for these functions is `psych`, but there are probably others that include the same functions.

A few helpful hints:

- To plot all three distributions on the same plot once mapped onto a common parameter space, you might either use the nested functions from R's base graphics and stats packages `plot(density(x = ...))` (for the first distribution) and `lines(density(x = ...))` (to superimpose the other two distributions); or you might use the equivalent function `geom_density()` from the `ggplot2` package. (There are probably other packages with functions for plotting density curves as well, so feel free to use what you like.)
- When you are plotting these density estimates, keep in mind the boundaries of the parameters in question: for the plot on the domain of θ, you should limit your density curves to the boundaries [0,1] (e.g. using the `from` and `to` arguments of the `density()` function). Similarly, when plotting on the domain of α, you should *strongly* consider limiting the plotting range to an interval such as [-20,20]; otherwise, using the `density()` or `geom_density()` function to plot sample distributions with extreme outliers (e.g. samples following the Cauchy distribution) will yield some bizarre results.

Question 1.2.

Remember that each of these distributions is supposed to encode some kind of generically weak prior information about the Bernoulli parameter $\theta = \text{logistic}(\alpha)$. Even so, each model encodes different kinds of prior information. To compare/contrast them, answer the following questions:

- While some research questions are concerned with testing hypotheses of fairness in binary outcomes—"fair coin" hypotheses—others focus on events that are believed to be either rare (e.g., epidemiological research focusing on rare diseases) or very common (e.g., questions about ownership of widely used appliances such as the number of households with at least one television). Which of the three default priors introduced above do you think would provide a good default for reliable estimates of Bernoulli parameters for "fair-coin" hypotheses? Which would provide a good default for "extremely rare/common event" hypotheses? Why? In answering this question, think about how each of these distributions concentrates prior belief, or similarly where each distribution's mode(s) fall.
- All of the default priors explored above are symmetrical around $\theta = 0.5$, or equivalently around $\alpha = 0$. What effect do these symmetrical priors have on the posterior estimation of θ? In answering this, recall that the posterior is a compromise between the prior and the likelihood function. Given your response, in what research contexts might an asymmetrical prior over Θ be favored instead, and why?

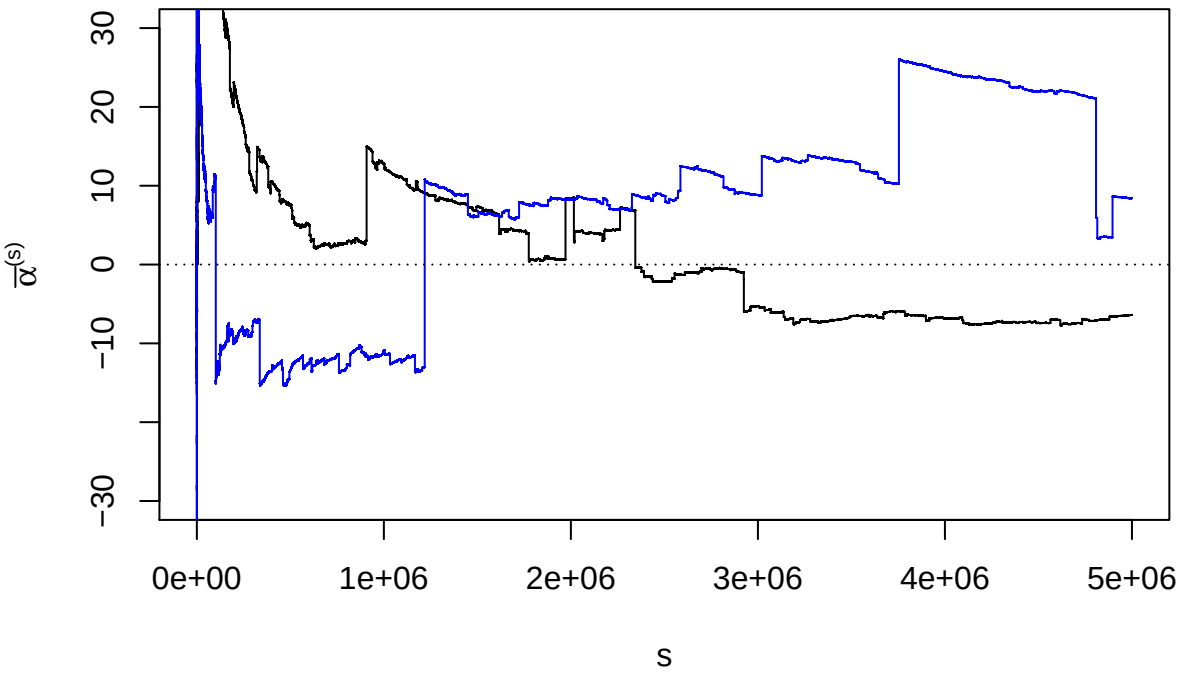
Question 1.3.

Which of these priors is the most diffuse (i.e. dispersed) and which is the most concentrated? To answer this question, calculate two quantile-based measures of scale on the domain Θ: the interquartile range (IQR) and the difference of the 97.5th and the 2.5th percentiles. (Consider using the `IQR()` and `quantile()` functions from R to make these calculations easier.)

An interesting side note about why we are favoring quantile-based rather than moment-based measures of distributional properties here: The mean and variance for the Cauchy distribution are undefined, so MC or "plug-in" approximations of these moments will be unstable, defying the Law of Large Numbers. This is

demonstrated in the simulation performed by the following code, which illustrates the behavior of the MC sample mean of a Cauchy distribution. Note that this code takes several seconds to run because rendering the plot for the cumulative mean of 5 million draws is slow. You do not need to include this illustration as a part of your answer to Question 1.3.

```
set.seed(1978);  cauchyMC1 = rcauchy(n = 5000000, location = 0, scale = 10)
set.seed(2021); cauchyMC2 = rcauchy(n = 5000000, location = 0, scale = 10)
meanPlugin1 = cummean(cauchyMC1)
meanPlugin2 = cummean(cauchyMC2)
plot(
  x = 1:5000000, y = meanPlugin1,
  type = "l",
  xlab = "s",
  ylab = expression(paste(bar(alpha)^"(s)")),
  ylim = c(-30, 30)
  )
lines(
  x = 1:5000000, y = meanPlugin2,
  col = "blue"
  )
abline(h = 0, lty = 3)
```



2. Second problem

2.1. Background

For clinical diagnosticians, arriving at reliable medical diagnoses is far less straightforward than we might initially believe. Making matters even worse, not all health care providers are skilled at effectively communicating about the results and interpretation of medical tests and diagnoses with their patients. Bayesian inference has important implications for how both doctors and their patients should navigate these tricky and important matters.

Imagine that a patient has had an annual checkup with their doctor. Unfortunately, this visit did not go so well: a routine "99% accurate" test has yielded a "positive result" for a medically important condition. Pretty frightening, right? Unfortunately for most non-expert patients, the meaning of the expression "99% accurate" is far more ambiguous than we might think, potentially referring to one of several non-identical concepts:

- the **sensitivity** (aka **probability of a true positive**) $P\left(T = 1 \mid C = 1\right)$;
- The **specificity** (aka **probability of a true negative**) $P\left(T = 0 \mid C = 0\right)$;
- The **positive predictive value** (abbreviated **PPV**) $P\left(C = 1 \mid T = 1\right)$;
- The **negative predictive value** (abbreviated **NPV**) $P\left(C = 0 \mid T = 0\right)$.

Both T and C are indicator variables, with $T = 1$ representing a positive test (0 if negative) and $C = 1$ representing the presence of the medical condition of interest (0 if absent). Note that there is no deterministic relationship between T and C such that if one variable equals 1, the other must as well; false positive and false negative tests are very real. Also note that the sensitivity of a test and its PPV are inverse probabilities.

After a sleepless night, our poor patient finally realizes that they are confused about the meaning of their test results, so they follow up with their doctor. The doctor clarifies that "99% accurate" in this case alludes to both the sensitivity and specificity of the test.[1] The doctor further states that the disease is relatively uncommon: its overall **prevalence** in the population, denoted $P\left(C = 1\right)$, is very low, affecting 1 in 10,000 people. Despite all of this, the doctor still fails to explicitly communicate to their patient the one thing that they really want to know: the PPV. Even so, does the doctor's clarification allow the patient to arrive at a solution to this unanswered question? Yes!

2.2. The questions

Question 2.1

Treat C as an estimand—an unknown quantity that you would really like to know—and treat T as observed evidence. Taken together, the epidemiologic term **prevalence** and its complement $P(C = 0) = 1 - P(C = 1)$ correspond with one of the probability functions in Bayes' Rule. Likewise, the **PPV** and **NPV** correspond with another of the probability functions in Bayes' Rule. Finally, the **sensitivity**, **specificity**, the probability of a false positive, and the probability of a false negative correspond with one of the probability functions in Bayes' Rule. Which components of Bayes' Rule correspond with each of these sets of clinical/epidemiological concepts?

Question 2.2

With a positive test and the known values of sensitivity, specificity, and prevalence given above, calculate the PPV.

[1] In reality, it is rare that a test's sensitivity and specificity are equal.

Question 2.3

Compare the PPV to the sensitivity of the test. How different is the patient's understanding of their medical status when they replace sensitivity with the PPV? Is this an example of "updated belief," in the Bayesian sense of that term?

Question 2.4

Compare the PPV to the prevalence of the condition. How different is the patient's understanding of their medical status when they replace prevalence with the PPV? Is this an example of "updated belief," in the Bayesian sense of that term?

Question 2.5

Calculate the NPV assuming two positive tests. How probable is it that a patient does not have the condition assuming two positive tests?

3. Third problem

3.1. Background

The basic linear regression model poses the variable Y as normally distributed around a central tendency defined as a linear predictor $\mathbf{X_i}\boldsymbol{\beta}$:

$$[3.1] \quad Y_i \sim \text{Normal}\left(\mu = \mathbf{X_i}\boldsymbol{\beta}, \sigma^2\right)$$

However, sometimes the sample distribution of Y around the linear predictor is **overdispersed**, showing extreme and/or many outliers. If even a small number of outliers are present, this may bias estimates of regression coefficients, including overestimating the effect of some model predictors that are in fact weakly or unassociated with the outcome. If a large number of outliers are evident, the tails of the distribution of residuals $\hat{\epsilon}$ will be thicker than expected according to the normal distribution even if $\mathbf{X_i}\boldsymbol{\beta}$ is correctly specified.

In either case, a robust alternative to the basic linear regression model may be preferred:

$$[3.2] \quad Y_i \sim \text{StudentT}\left(\nu, \mu = \mathbf{X_i}\boldsymbol{\beta}, \psi\right)$$

where ν is a strictly positive shape parameter commonly known as "degrees of freedom," μ is the mode and median of the distribution (and the mean, conditional on $\nu > 1$), and ψ is a dispersion parameter. The two-parameter normal distribution is a limiting case of the three-parameter Student's t:

$$[3.3] \quad \text{Normal}(\mu, \sigma^2 = \psi^2) = \lim_{\nu \to \infty} \text{StudentT}(\nu, \mu, \psi)$$

Conversely, the tails of the Student's t distribution are increasingly thick relative to those of the normal distribution as ν approaches zero (e.g. the thick-tailed Cauchy distribution, which fixes ν at 1 as noted earlier in Problem 1). Consequently, the Student's t distribution with a small value of ν is better able to accommodate extreme values of Y than the normal distribution, and replacing the normal with the Student's t distribution in regression models therefore tends to yield more robust estimates of regression coefficients. By extension, sparser regression models are often favored during variable selection than when the normal model is used.

The claim has been made that the Student's t model can be interpreted as a latent, continuous mixture model, under certain circumstances. According to this interpretation, the normal distribution of Y_i is conditioned not only on the ith observation's linear predictor $\mathbf{X_i}\boldsymbol{\beta}$ but also on an observation-specific variance V_i—

$$[3.4] \quad Y_i \sim \text{Normal}\left(\mu = \mathbf{X_i}\boldsymbol{\beta}, \sigma^2 = V_i\right)$$

—which in turn follows a "scaled" or two-parameter inverse chi-square distribution:

$$[3.5] \quad \{V_1, \ldots, V_n\} \overset{iid}{\sim} \text{Inv}\chi^2\left(\nu, \psi^2\right)$$

where ν is a shape parameter and ψ is a dispersion parameter. This distribution of variances embodies the latent, continuous mixture part of the model because each observation's unique variance parameter is an unknown quantity. In short, this interpretation claims the following:

$$[3.6] \quad \left\{ \begin{array}{c} Y_i \sim \text{Normal}\left(\mu = \mathbf{X_i}\boldsymbol{\beta}, \sigma^2 = V_i\right) \\ \cap \\ \{V_1, \ldots, V_n\} \overset{iid}{\sim} \text{Inv}\chi^2\left(\nu, \psi^2\right) \end{array} \right\} \Rightarrow Y_i \sim \text{StudentT}(\nu, \mu = \mathbf{X_i}\boldsymbol{\beta}, \psi)$$

In other words, the marginal probability of Y_i $p(y_i|\nu, \mu = \mathbf{X_i}\boldsymbol{\beta}, \psi)$ on the left side of the following equation is the Student's t pdf, providing a convenient analytical solution for an otherwise tedious integral. By extension, we can calculate the likelihood for this model over the joint parameter space of μ, $\boldsymbol{\beta}$, and ψ without estimating each individual "nuisance" parameter V_i.

$$[3.7] \quad p(y_i|\nu, \mu = \mathbf{X_i}\boldsymbol{\beta}, \psi) = \int_0^\infty p(y_i, V|\nu, \mu = \mathbf{X_i}\boldsymbol{\beta}, \psi)\, dV$$

$$= \int_0^\infty p(y_i|\mu = \mathbf{X_i}\boldsymbol{\beta}, \sigma^2 = V)\, p(V|\nu, \psi)\, dV$$

3.2. The questions

Question 3.1

The goal of this problem is to prove this claim through MC simulation. To operationalize this simulation:

- Set $S = 1,000,000$.
- For the sake of simplicity, assume an intercept-only model: $\mu = \alpha$ for all data points.
- Set the Student's t distribution's parameters according to the following:

$$[3.8] \quad \{Y_1, \ldots, Y_n\} \overset{iid}{\sim} \text{StudentT}\left(\nu = 5, \mu = 100, \psi = 10\right)$$

According Eq. 3.6, this distribution would follow from the two assertions:

$$[3.9] \quad Y_i \sim \text{Normal}\left(\mu = 100, \sigma^2 = V_i\right)$$

$$[3.10] \quad \{V_1, \ldots, V_n\} \overset{iid}{\sim} \text{Inv}\chi^2\left(\nu = 5, \psi^2 = 10^2\right)$$

A simulation-based proof that Eq. 3.8 follows from Eqs. 3.9 and 3.10 should involve the following:

- First, simulate S observations of Y from the Student's t distribution given by Eq. 3.8.
- Second, simulate S observations of V from Eq. 3.10, then pass these results to Eq. 3.9 to simulate S observations of Y.
- Third, visually compare the simulated samples of Y from the first and second simulation. Answer the question: Is the difference between these two MC sample distributions large, or negligible? If negligible, your proof is successful.

A few helpful hints:

- To simulate these samples, you will need to use packages that include the three-parameter Student's t and the two-parameter inverse chi-square distributions. The package `LaplacesDemon` contains both.
- Pay careful attention to which scale parameters are required by which distribution models in the packages you use, so that you pass the appropriate values to the appropriate arguments in the random number functions.
- When plotting thick-tailed sample distributions using `density()` or `geom_density()`, recall that you may need to limit the plotting range to a narrower interval than the full sample range. Otherwise, the rare extreme observations often lead to some bizarre plotting outcomes. You might consider the range [25, 175].

4. Fourth problem

4.1. Background

The (negative) exponential distribution is a one-parameter model often used in Event History Analysis to model survival/failure-time processes, appropriate when the age-specific hazard of "failing" remains constant over age. The model's one parameter is a concentration or rate parameter, λ. The entropy of the exponential distribution is

$$[4.1] \quad H(\lambda) = 1 - \ln \lambda$$

Let's consider two competing hypotheses about what λ might be for a particular class of observations: a "Cambridge" hypothesis that asserts a median life or "half-life" of 5730, a mean life of $-5730/\ln(0.5) \approx 8266.64$, and a rate parameter equaling the reciprocal of the mean[2]—

$$[4.2] \quad \lambda = -\ln(0.5)/5730 \approx 1.209681 \times 10^{-4}$$

—and a "Libby" hypothesis that asserts a half-life of 5568, a mean life of $-5568/\ln(0.5) \approx 8032.93$, and a rate parameter equaling the reciprocal of the mean[3]—

$$[4.3] \quad \lambda = -\ln(0.5)/5568 \approx 1.244876 \times 10^{-4}$$

4.2. The questions

Question 4.1

The R code below simulates the lifespan of a sample of 1000 observations, assuming the "Cambridge" rate.

```
CambHalfLife = 5730; LibbyHalfLife = 5568
(CambRate = -log(0.5)/CambHalfLife); 1/CambRate
```

```
## [1] 0.0001209681
```

```
## [1] 8266.643
```

```
(LibbyRate = -log(0.5)/LibbyHalfLife); 1/LibbyRate
```

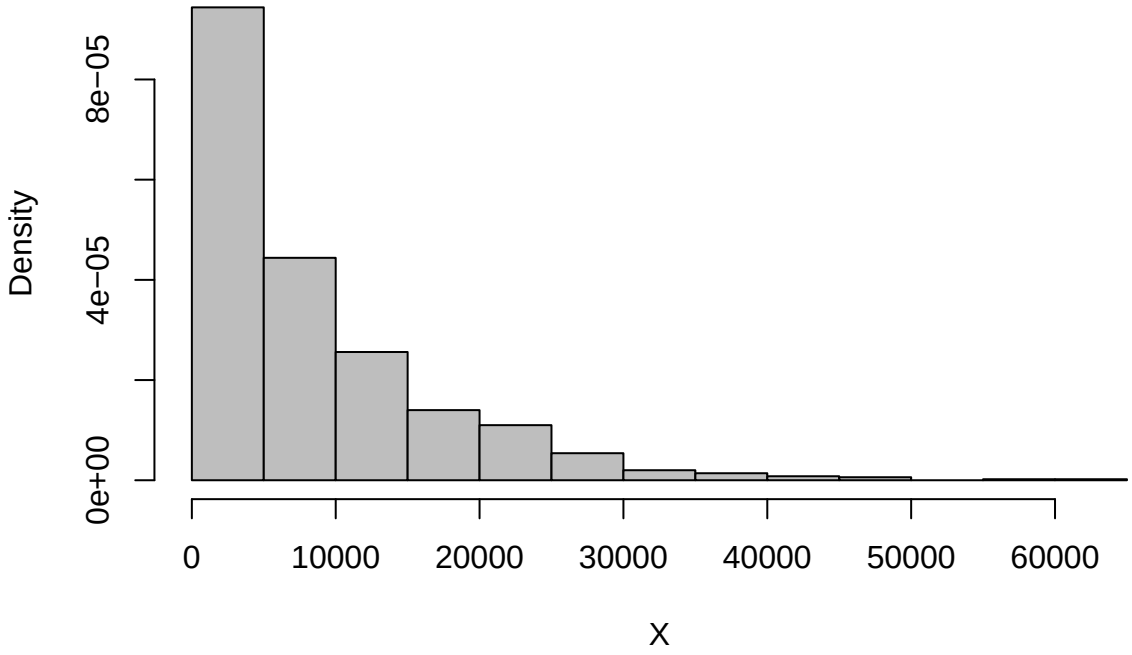
```
## [1] 0.0001244876
```

```
## [1] 8032.926
```

```
set.seed(1949); negExpMC = rexp(n = 1000, rate = CambRate)
hist(x = negExpMC, xlab = "X", main = NA, col = "gray", freq = FALSE)
```

[2]This is the "Cambridge" estimate of the instantaneous decay rate at which unstable radiocarbon atoms "decay" or stabilize into nitrogen-14 atoms. This estimate is based on an estimate of the radiocarbon half-life of 5730, where 'half-life' denotes the median lifespan of radioisotopes.

[3]This is Willard Libby's estimate of a 5568-year radiocarbon half-life. Libby introduced radiocarbon dating to geology, archaeology, and other historical sciences in 1949.



Using R as a calculator, calculate:

- entropy for the negative exponential distribution assuming the "Cambridge" rate λ;
- entropy for the negative exponential distribution assuming the "Libby" rate λ.

Question 4.2

Using R as a calculator, calculate the average sample surprisal for the simulated sample:

- assuming the Cambridge rate;
- assuming the Libby rate.

Given these average sample surprisals, which rate makes the sample most surprising?

Question 4.3

Subtract the Cambridge entropy from the Cambridge average sample surprisal. Likewise, subtract the Libby entropy from the Libby average sample surprisal. How different is each rate's average surprisal from its entropy (and in which direction)? Which sample surprisal comes closest to the corresponding entropy?

Question 4.4

The maximum likelihood estimator for the exponential model's rate parameter is the reciprocal of the sample average,

$$[4.4] \quad \hat{\lambda} = \frac{1}{\bar{x}}$$

Using R as a calculator, calculate

- the maximum likelihood estimate of λ for the simulated sample;
- the entropy implied by this maximum likelihood estimate;
- the average sample surprisal for the simulated sample based on this estimate.

Based on these calculations, answer the following questions:

- Is the sample more or less surprising given $\hat{\lambda}$ than assuming either the Cambridge or the Libby rate?
- Does the average surprisal implied by this maximum likelihood estimate come closer to or fall further from the entropy estimated for it than for the other two rates?