Data 102, Spring 2024
Midterm 1

- You have **110 minutes** to complete this exam. There are **7 questions**, totaling **50 points**.
- You may use **one** 8.5×11 sheet of handwritten notes (front and back), and the provided reference sheet. No other notes or resources are allowed.
- You should write your solutions inside this exam sheet.
- You should write your Student ID on every sheet (in the provided blanks).
- Make sure to write clearly. We can't give you credit if we can't read your solutions.
- Even if you are unsure about your answer, it is better to write down something so we can give you partial credit.
- We have provided a blank page of scratch paper at the **beginning** of the exam. No work on this page will be graded.
- You may, without proof, use theorems and facts given in the discussions or lectures, **but please cite them**.
- We don't answer questions individually. If you believe something is unclear, bring your question to us and if we find your question valid we will make a note to the whole class.
- Unless otherwise stated, no work or explanations will be graded for multiple-choice questions.
- Unless otherwise stated, you must show your work for free-response questions in order to receive credit.

Last name	
First name	
Student ID (SID) number	
Berkeley email	
Name of person to your left	
Name of person to your right	

Honor Code [1 pt]:

As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I am the person whose name is on the exam, and I completed this exam in accordance with the Honor Code.

Signature: ____

This page has been intentionally left blank. No work on this page will be graded.

1 Probability Paradigms (6 pts)

Frank the Frequentist, Barbara the Bayesian, and Sam the Statistician meet at the Bear's Lair to debate probability. Each explains their philosophy in turn. Complete each phrase below by circling one option from each set of parentheses. (Options are underlined and in italics).

(a) [1 Pt] Sam the Statistician (circle the correct answer):

- "The **empirical distribution** associated with an observed dataset is the distribution generated by drawing samples uniformly from (*~~the generating process~~,* ***the observed data***)."

(b) [3 Pts] Frank the Frequentist (circle the correct answer under a frequentist approach):

- "The probability an event happens on a single trial is the long-run (*~~average~~,* ***proportion**,* *~~number~~*) of trials in which the event happens if repeated and if the repetitions are (*~~independent~~,* ***independent and identically distributed***)."
- "We should treat data as (*~~fixed~~,* ***random***) and the generating process, underlying model, and/or reality as (***fixed**,* *~~random~~*)."
- "Probability (***can**,* *~~cannot~~*) be used to evaluate hypothetical statements regarding the chance a particular event happens given a simple hypothesis. Probability (*~~can~~,* ***cannot***) be used to evaluate the chance a particular hypothesis is true given data."

(c) [2 Pts] Barbara the Bayesian (circle the correct answer under a Bayesian approach):

- "Probability is a mathematical model for describing uncertainty about an unknown. Thus, we can use a **prior distribution** to model our uncertainty regarding an unknown (***before**,* *~~after~~*) observing data."
- "If the unknown fixes the process generating the data, then the **likelihood** is the probability of the (*~~hypothesis~~,* *~~unknown~~,* ***data***) given the (***hypothesis**,* ***unknown**,* *~~data~~*)."

2 True or False (8 Pts)

For each of the following, determine whether the statement is true or false.

If requested, you should explain your answer in 1-2 sentences. If not requested, no work or explanations will be graded.

(a) [1 Pt] A scientist plans to look for 10 different trends in a dataset. They will report any trend that is significant relative to its null. They do not report negative results. **Claim:** The scientist should adopt a 0.5% significance threshold to ensure that, if all the nulls were true, then the probability they falsely report any trend is less than 5%.

● **True** ○ False

(b) [2 Pts] A materials science lab is testing candidate materials for carbon capture. They plan to use two rounds. In the first round, they will rapidly iterate over many candidates to identify promising materials for more thorough testing. In the second, they will intensively test promising candidates. **Claim:** In the first round of testing the lab should prioritize FWER over FDR. *Explain your answer in two sentences or less.*

○ True ● **False**

Explanation:

> **Solution:** Since they're going to remove any potential false positives in a subsequent round, it's better to optimize for FDR rather than the more conservative FWER, since this makes it more likely they won't accidentally throw out any promising candidates in the first round due to a too-strict threshold.

(c) [1 Pt] The ROC curve associated with a particular test illustrates the trade-off between FPR and TPR over all possible significance thresholds.

● **True** ○ False

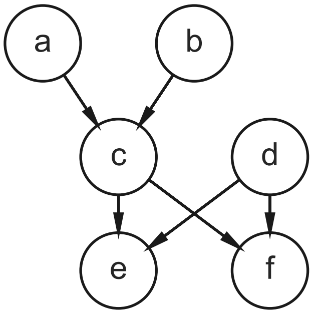
(d) [1 Pt] The posterior risk is the expected loss averaged over all possible unknowns that determine the data-generating process, given the observed data.

● **True** ○ False

(e) [1 Pt] In a generalized linear model with a sigmoid inverse link function, the features must always be between 0 and 1.

○ True ● **False**

Parts (f) and (g) refer to the following graphical model:



(f) [1 Pt] a and d are independent.

● **True** ○ False

(g) [1 Pt] c and d are conditionally independent given f.

○ True ● **False**

3 A Better Medical Test (7 points)

A research group is designing a cheap medical test (which costs $1 per test) to use as the initial screening for a rare but life-threatening condition,“Zyphronitis.” The prevalence of Zyphronitis in the population is 5%.

(a) [1 Pt] When designing the screening test, should the researchers prioritize decreasing FP or decreasing FN? Explain your reasoning in at most two sentences.

Explanation:

Solution: FN. The test will be followed by a supplementary test which may catch FP’s. On the other hand, a FN will not recieve the supplementary test, but is at sever risk as the disease is life-threatening.

(b) [2 Pts] The screening test has a TPR of 95% and a TNR of 80%. Calculate the probability that an individual has Zyphronitis if they test positive and then complete the sentence below by circling an option in the parentheses. You do not need to simply any arithmetic expressions.

Solution: Proceed by Baye’s rule:

$$\text{Pr(true P} \mid \text{test P)} = \frac{0.95 \times 0.05}{0.95 \times 0.05 + (1 - 0.80) \times (1 - 0.05)} = \frac{0.95 \times 0.05}{0.95 \times 0.05 + 0.20 \times 0.95}$$

$$= \frac{0.05}{0.05 + 0.20} = \frac{0.05}{0.25} = \frac{1}{5} = 0.2.$$

The probability an individual who tests positive has Zyphronitis is $(<)$ the sensitivity, and

(~~*decreases,*~~ ~~*stays the same,*~~ ***increases***) if the prevalence of Zyphronitis increases.

(c) [2 Pts] A supplementary test has a TPR of 80% and TNR of 90%. The supplementary test is only administered if the screening test is positive. We define the **combined test** as positive if both tests (screening and supplementary) are positive, and negative otherwise. When applied, the outcome of the supplementary test depends only on the disease status of the patient.

Calculate the TPR and TNR of the combined test. You do not need to simplify your answers.

Solution: For a positive individual to test positive on the combined test they must test positive on both tests. Therefore, the TPR of the combined test is $0.95 \times 0.80 = 0.76$.

For a negative individual to test negative on the combined test they must either test negative on the first, or test negative on the second having tested positive on the first. It is easier to find the probability of a false positive, FPR = $(1 - 0.80) \times (1 - 0.90) = 0.20 \times 0.10 = 0.02$. Then, TPR = $1 - 0.02 = 0.98$.

Equivalently, $0.80 + 0.20 \times 0.90 = 0.80 + 0.18 = 0.98$.

(d) [2 Pts] The researchers want to prepare a report recommending the test with the lowest false discovery proportion (FDP). Select the option below that best summarizes the result of the supplementary test on the FDP. Choose the single best answer by filling in the circle next to it.

Hint: this question can be answered conceptually by checking each statement carefully. You may also compute the answer using the results of parts (b) and (c).

○ We recommend against the combined test. The supplementary test has a lower specificity than the screening test, so the combined test has a higher FDP than the screening test alone.

● **We recommend the combined test. The combined test has greater specificity than the screening test, and a lower FDP.**

○ We recommend the combined test. The combined test has greater specificity and sensitivity than the screening test, thus a lower FDP.

4 Xesla (9 points)

A car company, Xesla, collects data on the lifespan, T, of their electric car batteries. Xesla's engineers want to evaluate whether their batteries outlast the industry standard of 5 years. The engineers formulate two hypotheses:

- Null Hypothesis (H_0): $T \sim \text{Exponential}(1/5)$;
- Alternative Hypothesis (H_1): $T \sim \text{Exponential}(1/8)$.

Using observed battery lifetimes, the engineers compute a valid test statistic and report its p-value, p, for a one-sided test.

(a) [3 Pts] Which of the following is/are true about the NHST scheme in this example? Select all answers that apply. No work will be graded for this question.

□ The NHST tests whether the Xesla battery's lifetimes are more likely to have been drawn from the alternative than the null given the value of the test statistic.

■ **To make a decision, the engineers should compare the value of the test statistic to its sampling distribution under the null.**

□ If the engineers reject the null, then they can adopt the alternative since the alternative is a simple hypothesis.

□ If the engineers fail to reject the null, then they should conclude that the battery lifetimes have a true average of 5 years.

(b) [2 Pts] Which of the following is/are true about the p-value in this example? Select all answers that apply. No work will be graded for this question.

□ The p-value is a deterministic value that gives the chance that the testing statistic is greater than a given threshold under the null.

■ **If H_0 is true, then the observed p-value was drawn uniformly between 0 and 1.**

■ **The p-value threshold controls the false positive rate (FPR).**

(c) [1 Pt] Complete the sentence below by circling one option from each set of parentheses.

According to the Neyman-Pearson Lemma, any test statistic that is a one-to-one, monotonic function of the (*~~likelihood under the null~~*, *~~likelihood under the alternative~~*, ***likelihood ratio***) will be the uniformly most powerful test for distinguishing between two (***simple***, *~~composite~~*) hypotheses.

Hint: if a test is the "uniformly most powerful test," that means it will maximize power for any chosen significance threshold.

(d) [3 Pts] The engineers observe battery lifetimes $T_1, T_2, \ldots, T_n$. Suppose that the lifetimes were exponentially distributed with some unknown parameter λ: $T_i \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$. Find the Maximum Likelihood Estimator for λ in terms of $T_1, \ldots, T_n$.

Solution: To find the MLE estimator, we maximize the likelihood. The likelihood is:

$$P(\{T_i\}_{i=1}^n \mid \lambda) = \prod_{i=1}^n \lambda e^{-\lambda T_i} = \lambda^n e^{-\lambda \sum_{i=1}^n T_i} = \exp\left(n \log(\lambda) - \lambda \sum_{i=1}^n T_i\right).$$

It is equivalent (and easier) to maximize the log-likelihood:

$$\log(P(\{T_i\}_{i=1}^n \mid \lambda)) = n \log(\lambda) - \lambda \sum_{i=1}^n T_i.$$

Checking first-order optimality:

$$\frac{d}{d\lambda} \log(P(\{T_i\}_{i=1}^n \mid \lambda)) = \frac{n}{\lambda} - \sum_{i=1}^n T_i$$

The derivative is zero when n/λ equals the sum of observed lifetimes. That is:

$$\frac{1}{\lambda_{MLE}} = \frac{1}{n} \sum_{i=1}^n T_i = \bar{T}$$

where $\bar{T}$ is the average of the observed lifetimes. Therefore, the MLE estimate for λ is the reciprocal of the observed average lifetime.

5 Bayesian Bakeries (8 points)

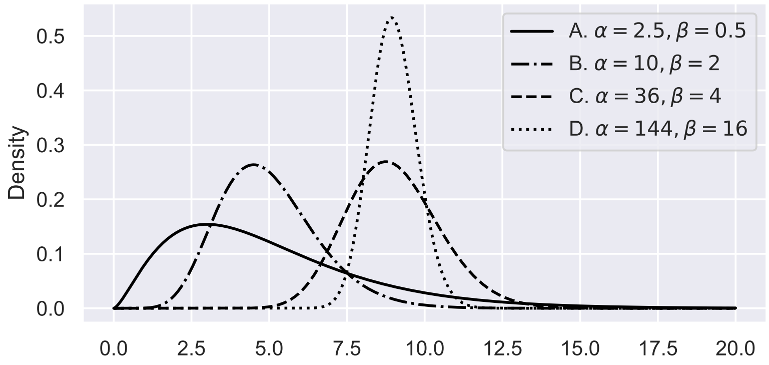
Xavier and Yolanda each own a bakery, and they notice that they have a lot of bread leftover at the end of each day. Let $x_1, \dots, x_n$ be the number of loaves left over each day from Xavier's bakery, and $y_1, \dots, y_m$ be the same for Yolanda's bakery.

They decide to use a Bayesian model to infer the average number of loaves left over, assuming a Poisson model:

$$\begin{aligned} x_i \mid \mu &\overset{\text{iid}}{\sim} \text{Poisson}(\mu), \quad i = 1, \dots, n \\ y_j \mid \eta &\overset{\text{iid}}{\sim} \text{Poisson}(\eta), \quad j = 1, \dots, m \\ \mu &\sim \text{Gamma}(\alpha, \beta) \\ \eta &\sim \text{Gamma}(\alpha, \beta) \end{aligned}$$

Recall from discussion that the Gamma is the conjugate prior for the Poisson likelihood: in other words, if $z_i \mid w \overset{\text{iid}}{\sim} \text{Poisson}(w)$ for $i = 1, \dots, n$ and $w \sim \text{Gamma}(a, b)$, then $w \mid z_1, \dots, z_n \sim \text{Gamma}(a + \sum_{i=1}^n z_i, b + n)$.

For part (b), consider the following four choices of α and β, with the corresponding Gamma density shown:



Hint: Throughout this question, you may find it helpful to use facts about the Gamma distribution provided on the reference sheet.

(a) [2 Pts] Draw a graphical model for the variables in this problem. Your model must include $x_1, \dots, x_n, y_1, \dots, y_m$, μ, η, α, and β.

Hint: don't forget to shade in variables that are observed!

(b) [2 Pts] Suppose Yolanda and Xavier are very sure that the average number of loaves left over is close to 9 per day for both of their bakeries. Which of the four choices of α and β above should they choose? Choose the single best answer by filling in the circle next to it. Explain your answer in two sentences or less.

○ $\alpha = 2.5, \beta = 0.5$

○ $\alpha = 10, \beta = 2$

○ $\alpha = 36, \beta = 4$

● $\alpha = 144, \beta = 16$

Explanation:

Solution: Neither of the first two has a mean of 9, so they are not good choices. Between the last two choices, the fourth one has more certainty and lower variance (as shown in the diagram), so it is the better choice. We can also compute the variance to show this.

For the remainder of the question, they decide to use $\alpha = 25$ and $\beta = 5$.

(c) [1 Pt] Xavier observes data for five days, and correctly computes that $\mu \mid x_1, \ldots, x_5 \sim \text{Gamma}(80, 10)$. Which of the following distributions does this correspond to? Choose the single best answer by filling in the circle next to it. No work or explanations will be graded for this question.

○ Prior ○ Likelihood ● **Posterior** ○ Posterior predictive

(d) [3 Pts] Suppose Xavier has been collecting data for 5 days and Yolanda has been collecting data for 195 days (in other words, $n = 5$ and $m = 195$). Neither of them ever observes more than 25 loaves left over. They both compute the posterior distribution for their respective averages. Which one of them will find a posterior distribution with lower variance? Provide a mathematical justification for your answer.

Hint: if you aren't able to justify it mathematically, you may provide an intuitive explanation in two sentences or less for partial credit.

○ Xavier (μ, with $n = 5$)

○ Yolanda (η, with $m = 195$)

Justification:

Solution: Intuitively, for most prior distributions and assuming that the data are similarly distributed, we should expect that whoever observes more data will have a prior with lower variance, so the answer should be Yolanda.

$$\begin{aligned}\text{Var}(\mu|x_{1:5}) &= \frac{\alpha + \sum_i x_i}{(\beta+5)^2}\\ \text{Var}(\eta|y_{1:195}) &= \frac{\alpha + \sum_j y_j}{(\beta+195)^2}\end{aligned}$$

Using our intuition above, we want to show that the second term above is smaller than the first:

$$\begin{aligned}\text{Var}(\mu|x_{1:5}) &\overset{?}{>} \text{Var}(\eta|y_{1:195})\\ \frac{\alpha + \sum_i x_i}{(\beta+5)^2} &\overset{?}{>} \frac{\alpha + \sum_j y_j}{(\beta+195)^2}\end{aligned}$$

The LHS is smallest when $x_i = 0 \forall i$, and the RHS is largest when $y_j = 25 \forall j$ (since 25 is the maximum value observed). Plugging in:

$$\begin{aligned}\frac{\alpha}{(\beta+5)^2} &\overset{?}{>} \frac{\alpha + 195 \times 25}{(\beta+195)^2}\\ \frac{25}{100} &\overset{?}{>} \frac{196 \times 25}{200^2}\\ \frac{25}{100} &> \frac{196}{2 \times 200}\frac{25}{100}\end{aligned}$$

So, $\text{Var}(\mu \mid x_{1:5}) > \text{Var}(\eta \ \mid y_{1:195})$.

6 Sampling (5 pts)

Dulé sets up a Bayesian model with two unknown variables θ_1 and θ_2 and observed data $x_1, \ldots, x_n$, where θ_1 and θ_2 are each between 0 and 1. After correctly defining a prior distribution and likelihood model, he wants to use sampling to approximate the posterior distribution.

(a) [2 Pts] He uses rejection sampling, and finds that his acceptance probability is very small (i.e., most of his samples are rejected). Which of the following, if true, would correctly explain why? Select all answers that apply.

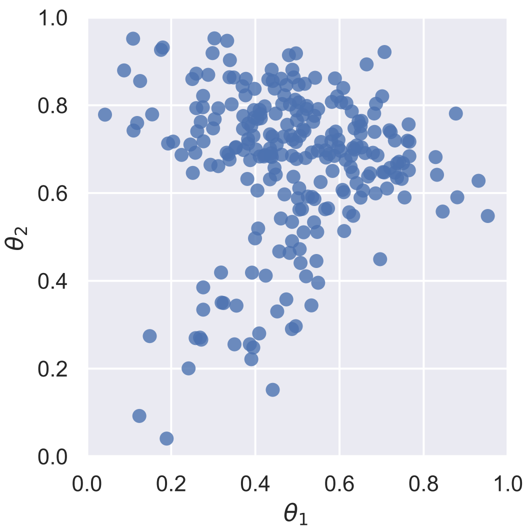
■ **The scaled target distribution is too large.**

□ The number of proposals is too large.

■ **For most values of θ_1 and θ_2, the true posterior density is very close to 0.**

(b) [3 Pts] Next, he correctly defines his model in PyMC, and obtains the 250 samples shown in the scatterplot below.

Assuming the samples accurately reflect the true posterior distribution, which of the following are correct interpretations? Select all answers that apply.



□ $\mathbb{P}(\theta_1 > \theta_2 \mid x_1, \ldots, x_n) > 0.5$

■ **If the prior was uniform over $(0,0)$ to $(1,1)$, then most of the data observed were consistent with values of θ_2 greater than 0.5.**

■ **If the observed data were independent of (θ_1, θ_2), then the samples computed above are just as good an approximation of the prior as they are of the posterior.**

7 School Funding, Again (6 points)

In HW2, we used a Bayesian hierarchical model for the state-level average funding gap, treating the state-level variance as known. In this question, we'll instead examine a hierarchical model for the state-level variances, treating the means μ_i as known and fixed:

$$\begin{aligned} y_{ij} \mid \sigma_i &\overset{\text{iid}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2) \\ \sigma_i^2 &\sim \text{InverseGamma}(\alpha, \beta) \end{aligned}$$

The **inverse Gamma** distribution is the conjugate prior for the variance of a normal distribution when the mean is known. *Hint: information about the parameters of the inverse Gamma and the posterior distribution is provided on the reference sheet, but you should be able to answer this entire question without doing any computation.*

In this question, we'll compare three approaches:

(1) A frequentist model where we treat each σ_i as fixed but unknown, and estimate each one separately from the data for the corresponding state using maximum likelihood estimation (MLE).

(2) A Bayesian model where α and β are chosen using empirical Bayes.

(3) A fully hierarchical Bayesian model where we treat α and β as random variables, with exponential priors for each one. Note that the exponential distribution is not the conjugate prior for either parameter of the inverse gamma distribution.

(a) [3 Pts] **For this part, only consider models (1) and (2)**. Which of the following statements are true? Select all answers that apply. No work or explanations will be graded for this question.

■ **For large states (e.g., California), the MLE frequentist estimates from model (1) and MAP Bayesian estimates from model (2) will be similar.**

□ The MLE frequentist estimates from model (1) for small states will depend on the choice of α and β.

■ **Model (1) is an example of an approach with no pooling.**

(b) [3 Pts] **For this part, only consider model (3)**. Which of the following statements are true? Select all answers that apply. No work or explanations will be graded for this question.

■ **Let $Y = y_{11}, \ldots, y_{nm}$ be the collection of all observed funding gaps for all districts in all states. Then the posterior distribution is $p(\alpha, \beta, \sigma_1, \ldots, \sigma_n \mid Y)$.**

□ The posterior distribution can be computed exactly using numerical techniques: no approximation is necessary.

■ **The posterior distribution for α and β represents nation-wide information about the variation in funding gaps.**

□ Model (3) is an example of complete pooling.

8 Congratulations [0 Pts]

Congratulations! You have completed Midterm 1.

- **Make sure that you have written your student ID number on *every other page* of the exam.** You may lose points on pages where you have not done so.
- Also ensure that you have **signed the Honor Code** on the cover page of the exam for 1 point.
- If more than 10 minutes remain in the exam period, you may hand in your paper and leave. If $\leq$ 10 minutes remain, please **sit quietly** until the exam concludes.

[Optional, 0 pts] Draw a picture or cartoon that's related to your favorite thing you've learned in Data 102 so far.

Midterm 1 Reference Sheet

Useful Distributions:

Distribution	Support	PDF/PMF	Mean	Variance	Mode
$X \sim \text{Poisson}(\lambda)$	$x = 0, 1, 2, \ldots$	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ	$\lfloor \lambda \rfloor$
$X \sim \text{Binomial}(n, p)$	$x \in \{0, 1, \ldots, n\}$	$\binom{n}{x} p^x (1-p)^{1-x}$	np	$np(1-p)$	$\lfloor (n+1)p \rfloor$
$X \sim \text{Beta}(\alpha, \beta)$	$0 \le x \le 1$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha}{\alpha+\beta} \frac{\beta}{\alpha+\beta} \frac{1}{\alpha+\beta+1}$	$\frac{\alpha-1}{\alpha+\beta-2}$
$X \sim \text{Gamma}(\alpha, \beta)$	$x \ge 0$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\frac{\alpha-1}{\beta}$
$X \sim \mathcal{N}(\mu, \sigma^2)$	$x \in \mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$	μ	σ^2	μ
$X \sim \text{Exponential}(\lambda)$	$x \ge 0$	$\lambda \exp(-\lambda x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	0
$X \sim \text{InverseGamma}(\alpha, \beta)$	$x \ge 0$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$	$\frac{\beta}{\alpha-1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$	$\frac{\beta}{\alpha+1}$

Conjugate Priors: For observations x_i, $i = 1, \ldots, n$:

Likelihood	Prior	Posterior
$x_i \mid \theta \sim \text{Bernoulli}(\theta)$	$\theta \sim \text{Beta}(\alpha, \beta)$	$\theta \mid x_{1:n} \sim \text{Beta}\left(\alpha + \sum_i x_i, \beta + \sum_i (1 - x_i)\right)$
$x_i \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$	$\mu \sim \mathcal{N}(\mu_0, 1)$	$\mu \mid x_{1:n} \sim \mathcal{N}\left(\frac{\sigma^2}{\sigma^2+n}\left(\mu_0 + \frac{1}{\sigma^2}\sum_i x_i\right), \frac{\sigma^2}{\sigma^2+n}\right)$
$x_i \mid \lambda \sim \text{Exponential}(\lambda)$	$\lambda \sim \text{Gamma}(\alpha, \beta)$	$\lambda \mid x_{1:n} \sim \text{Gamma}\left(\alpha + n, \beta + \sum_i x_i\right)$
$x_i \mid \lambda \sim \text{Poisson}(\lambda)$	$\lambda \sim \text{Gamma}(\alpha, \beta)$	$\lambda \mid x_{1:n} \sim \text{Gamma}\left(\alpha + \sum_i x_i, \beta + n\right)$
$x_i \mid \lambda \sim \mathcal{N}(\mu, \sigma^2)$	$\sigma \sim \text{InverseGamma}(\alpha, \beta)$	$\sigma \mid x_{1:n} \sim \text{InverseGamma}\left(\alpha + n/2, \beta + \left(\sum_{i=1}^n (x_i - \mu)^2\right)/2\right)$

Generalized Linear Models

Regression	Inverse link function	Likelihood
Linear	identity	Gaussian
Logistic	sigmoid	Bernoulli
Poisson	exponential	Poisson
Negative binomial	exponential	Negative binomial

Some powers of e:

x	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$y = e^x$	1.05	1.11	1.22	1.35	1.49	1.65	1.82	2.01	2.23	2.46	2.72