Data 102, Fall 2025
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 True or False [3 Pts]

For each of the following, determine whether the statement is true or false. For this question, no work will be graded and no partial credit will be assigned.

(a) [1 Pt] In a hypothesis test, a p-value of 0 means that the alternative hypothesis must be correct.

◯ True ◯ False

(b) [1 Pt] An ROC curve visualizes the tradeoff between row-wise rates and column-wise rates.

◯ True ◯ False

(c) [1 Pt] True or False: When using Poisson likelihood in a generalized linear model (GLM), the inverse link function must never return negative values.

◯ True ◯ False

2 Autograder Accuracy (8 pts)

A professor is evaluating the effectiveness of an automatic grading system for written responses on assignments. For each question, students' responses can be correct (1) or incorrect (0). The autograder gives each response credit (1) or no credit (0), with no partial credit.

(a) [2 Pts] On a pilot assignment with a single question and a class of 100 students, the professor determines that:

- The prevalence of correct responses is 80%.
- The autograder's specificity is 90%
- The autograder made 10 false negatives.

Using the information above, fill in the confusion matrix below with the number of responses for each entry. If you do not have enough information to fill a particular entry, draw an "X" through it.

	Autograder gives no credit	Autograder gives credit
Response incorrect		
Response correct		

(b) [2 Pts] The professor hires a tutor to review all the autograder's decisions. Each of the four cases from the confusion matrix above requires a different amount of time for tutors to review. The professor defines the tutor's time spent per question as his loss function.

The professor computes that on average, the tutor spends 2 minutes on incorrect responses and 3 minutes on correct responses. Which of the following did the professor just compute? Choose the single best answer **by filling in the circle next to it.**

◯ Loss

◯ Frequentist risk

◯ Bayesian posterior risk

◯ Bayes risk

◯ None of the above

(c) [2 Pts] A TA suggests writing harder questions, assuming that the prevalence of correct answers will decrease. The TA claims: "If the prevalence of correct answers is lower for harder questions than for easy questions, then the false discovery proportion (FDP) will increase." Which of the following are true statements about this claim? Select all answers that apply **by filling in the square next to each correct answer**.

- ☐ If the autograder has the same FPR and FNR on hard questions as easy questions, then the claim **must be correct**.
- ☐ If the autograder performs worse (higher FPR and higher FNR) on hard questions than easy ones, then the claim **could be incorrect**.
- ☐ If the autograder has lower FPR but higher FNR on hard questions than easy ones, then the claim **could be correct or incorrect**.

(d) [2 Pts] The professor tries calculating the false omission proportion (FOP) as a function of TPR, TNR, FPR, FNR, and prevalence $\pi_1 = P(R = 1)$ using Bayes' rule, but he makes a mistake. Circle the first error in the work shown below and write the correct value for the quantity you circled. You should only circle the **first occurrence** of the error.

For example, if the question were "solve for x: $3x + 2 = 8$" and the work shown was "$3x = 7, x = 7/3$," you would circle only the 7 in the $3x = 7$ step and write "6" next to your circle, since that's the first step with the error and 6 is the correct value instead of 7. Circling anything more than the 7 would not be specific enough, and would not receive full credit.

Hint: recall that the false omission proportion is the proportion of the "0" decisions that are incorrect.

$$
\begin{aligned}
FOP &= P(R = 1 | D = 0) \\
&= \frac{P(D = 0 | R = 1) P(R = 1)}{P(D = 0)} \\
&= \frac{P(D = 0 | R = 1) P(R = 1)}{P(D = 0 | R = 1) P(R = 1) + P(D = 0 | R = 0) P(R = 0)} \\
&= \frac{FNR \times \pi_1}{FNR \times \pi_1 + TPR \times (1 - \pi_1)} \\
&= \frac{1}{1 + \frac{TPR}{FNR} \frac{1 - \pi_1}{\pi_1}}
\end{aligned}
$$

3 Drug Safety [12 points]

A vaccine researcher is reviewing safety records from n preliminary drug candidate trials in mice. For each trial i, a study used blood pressure data from m mice, $x_{i1}, \ldots, x_{im}$ to test whether the drug was unsafe (H_1) or safe (H_0). She gathers the following data from each study i:

- A p-value p_i
- A test statistic $\hat{\mu}_i$, the sample average of blood pressure data for mice in that study
  - Larger values of $\hat{\mu}_i$ indicate unsafe drugs, and smaller values indicate safe drugs.
- Null and alternative hypotheses

(a) [2 Pts] Which of the following error rates best answers the question: "How often were unsafe drugs incorrectly classified as safe?" Choose the single best answer **by filling in the circle next to it.**

◯ FDR ◯ FNR ◯ FPR ◯ FWER ◯ None of these

(b) [2 Pts] Each statement below describes a possible decision the researcher will make for each p-value. For each one, choose whether she should control family-wise error rate (FWER) at level $\alpha = 0.01$ or false discovery rate (FDR) at level $\alpha = 0.01$, and briefly justify your answer in one sentence or less.

(i) For each candidate, she will use the p-value to determine whether to move forward with the next stage of trials in humans.

◯ Control FWER ◯ Control FDR

Justification:

(ii) For each candidate, she will use the p-value to determine whether to send the drug candidate for a series of low-cost lab tests to find any contamination or toxicity.

◯ Control FWER ◯ Control FDR

Justification:

For parts (c)-(d) only, assume the following: There are four studies, and the decision for each study is made by thresholding the p-value at 0.05. Each one assumes $x_{ij} \sim \mathcal{N}(\mu_i, 4)$, with the following null/alternative hypotheses:

- Study 1: H_0: $\mu_1 = 120$, H_1: $\mu_1 > 120$
- Study 2: H_0: $\mu_2 = 100$, H_1: $\mu_2 > 100$
- Study 3: H_0: $\mu_3 = 120$, H_1: $\mu_3 = 140$
- Study 4: H_0: $\mu_4 = 100$, H_1: $\mu_4 = 160$

(c) [2 Pts] The researcher wants to calculate both power and significance level (FPR). For which study or studies can she calculate **both power and FPR**? Select all answers that apply **by filling in the square next to each correct answer**.

☐ Study 1 ☐ Study 2 ☐ Study 3 ☐ Study 4

(d) [1 Pt] Which study had the highest power? Choose the single best answer **by filling in the circle next to it.**

◯ Study 1 ◯ Study 2 ◯ Study 3 ◯ Study 4 ◯ Cannot be determined

For parts (e) and (f), assume $n = 5$ and the p-values are $0.6, 0.01, 0.1, 0.7$, and 0.12. For each part, to receive full credit, **you should circle or box your answer (either a value of α or the word "impossible"), and show your work**.

(e) [2 Pts] She decides to control the **family-wise error rate (FWER)** at level α using Bonferroni correction. What is the smallest value of α for which she makes exactly two discoveries? If no such value exists (i.e., it is impossible to control FWER and make exactly two discoveries), explain why.

(f) [3 Pts] She decides to control the **false discovery rate (FDR)** at level α using the Benjamini-Hochberg procedure. What is the smallest value of α for which she makes exactly two discoveries? If no such value exists (i.e., it is impossible to control FDR and make exactly two discoveries), explain why.

4 Brandon's Bayesian Basketball (10 pts)

Brandon is a fan of his local basketball team, the Berkeley Bayesians. He goes to n of their m games (where $n < m$), chosen at random, and records how many points they score in each game, calling these $b_1, \ldots, b_n$. He wants to know the average number of points they score across all their games, which he calls λ. He decides to model the scores b_i as **i.i.d. Poisson random variables with parameter** λ (recall that the parameter of the Poisson is its mean).

Hint: you may find it helpful to use facts about the Poisson and Gamma distributions given on the reference sheet. You may use any of these facts without proving or showing them.

(a) [2 Pts] Describe **one** assumption that Brandon is making with his model above, and then explain why that assumption might not hold. Your answer must be two sentences or less, and must be related to the question setting (i.e., basketball game scores) to receive full credit.

For the remainder of this question, assume that the model is an accurate representation of how the team scores points every game.

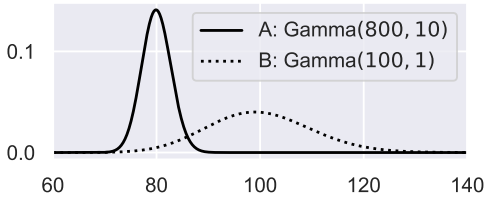
(b) [1 Pt] Consider the maximum likelihood estimate $\hat{\lambda}_{MLE}$ for λ based on the data $b = (b_1, \ldots, b_n)$. If we compare the MLE to a different estimate ρ where $\rho \neq \hat{\lambda}_{MLE}$, which of the following must be true? Select all answers that apply **by filling in the square next to each correct answer**.

- ☐ ρ is less probable than $\hat{\lambda}_{MLE}$ given the observed scores.
- ☐ It could be true that $p(b|\rho) > p(b|\hat{\lambda}_{MLE})$.
- ☐ It could be true that $p(b|\rho) < p(b|\hat{\lambda}_{MLE})$.

For parts (c)-(e), Brandon uses a Bayesian approach with a Gamma prior: $\lambda \sim \text{Gamma}(\alpha, \beta)$.

(c) [1 Pt] Let $t(\alpha, \beta)$ be the most likely value for the average number of points the Berkeley Bayesians score across all their games, according to Brandon's prior. Compute $t(100, 1)$.

(d) [3 Pts] Brandon is debating between two priors, A and B, shown below.



For each statement, pick whether it applies to Prior A, Prior B, both, or neither. Choose the single best answer **by filling in the circle next to it.**

Hint: These can all be answered with minimal calculation.

(i) Brandon should choose this prior if he is very sure the average score will be between 70 and 90 points.

◯ Prior A ◯ Prior B ◯ Both ◯ Neither

(ii) If Brandon chooses this prior and observes 3 games with an average score of 90 points per game, then his MAP estimate for λ will be less than 90.

◯ Prior A ◯ Prior B ◯ Both ◯ Neither

(iii) If Brandon observes data from a very large number of games ($n > 100$) and the observed mean is somewhere between 110 and 120 points, then the LMSE will be closer to the observed mean than to the mean of the prior distribution.

◯ Prior A ◯ Prior B ◯ Both ◯ Neither

(e) [3 Pts] For this part only, Brandon chooses $\alpha = 100$ and $\beta = 1$, and observes data from four games. The first three scores are $b_1 = 60, b_2 = 70$, and $b_3 = 50$. Find the value of b_4 that would make Brandon's LMSE estimate **twice as large** as his MLE estimate, or show why this is impossible.

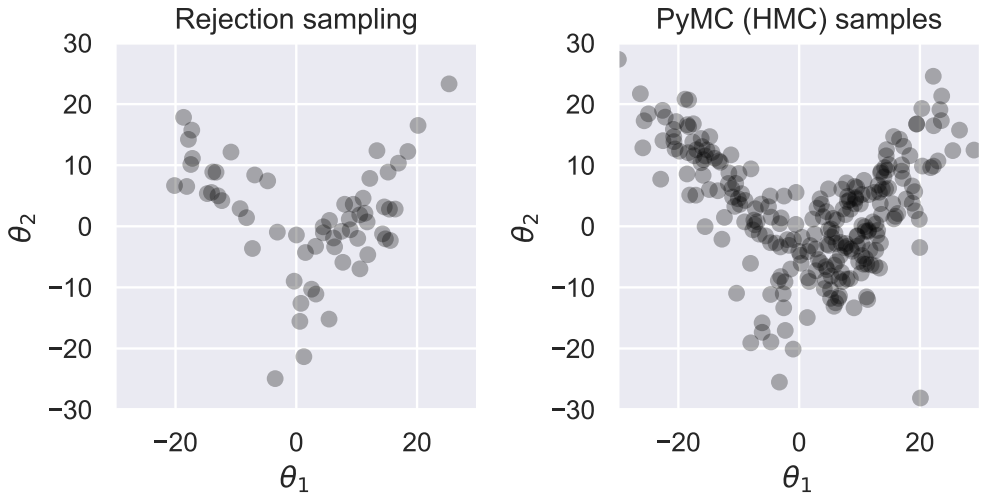
You may use (without proof) the fact that, given i.i.d. observations $x_1, \ldots, x_n$ of a Poisson random variable with parameter λ, the MLE for λ is the sample mean: $\sum_{i=1}^{n} x_i/n$.

5 Sampling [6 points]

Liesl defines a Bayesian model with two unknown variables, θ_1 and θ_2. She observes data x and computes the unnormalized posterior $q(\theta_1, \theta_2) \propto p(\theta_1, \theta_2|x)$, where $q(\theta_1, \theta_2) = 0$ for values outside the axis boundaries of the graphs below. She then takes two approaches for inference:

1. She correctly implements rejection sampling to approximate the posterior distribution with **250 proposals**.
2. She correctly implements her model in PyMC and obtains **250 samples** that way.

Here are the samples. Assume that each graph shows all samples produced by that method (i.e., there are no extra samples outside what is shown):

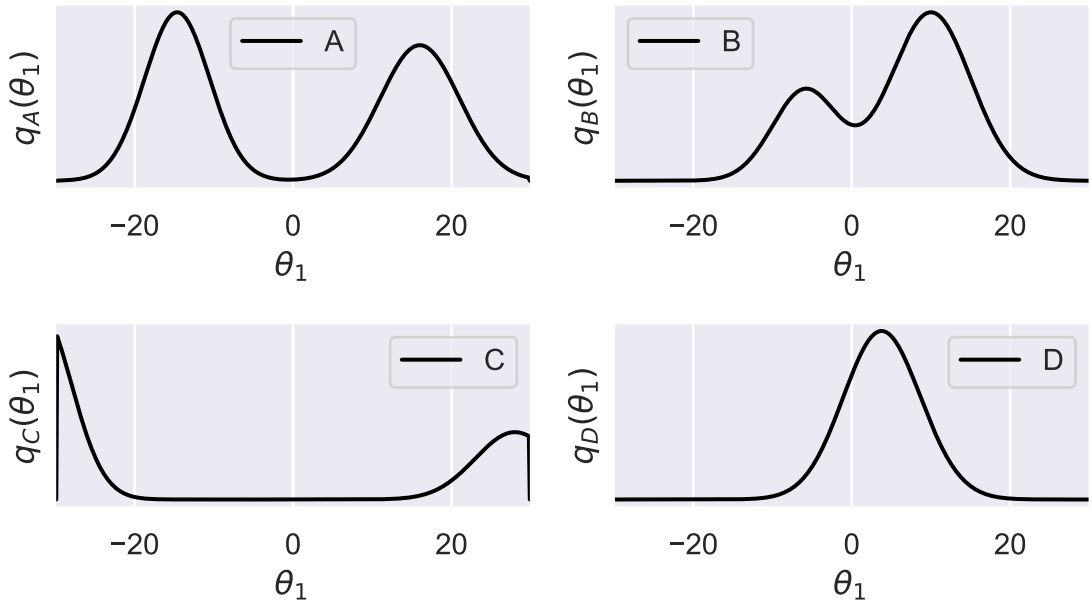


(a) [2 Pts] Given the results above, which of the following are possible valid explanations for why rejection sampling produces so few accepted samples? Select all answers that apply **by filling in the square next to each correct answer**.

Hint: recall that Liesl's implementation of rejection sampling is correct, i.e., there are no errors, mistakes, or bugs, but she hasn't necessarily made the most efficient choices.

- ☐ She chose a scaling factor for her target distribution q that was **too small**.
- ☐ She chose a scaling factor for her target distribution q that was **too large**.
- ☐ She generated proposals using independent Uniform$(-20, 20)$ distributions for θ_1 and θ_2.

(b) [2 Pts] Liesl tries Gibbs sampling. She initializes θ_1 to some value between -20 and 20 that she doesn't tell us, and $\theta_2 = 10$, and decides to start by sampling a new value for θ_1. Which of the following distributions should she sample from to obtain her next value for θ_1? Choose the single best answer **by filling in the circle next to it.**



◯ A

◯ B

◯ C

◯ D

◯ Cannot be determined without knowing the initial value of θ_1

(c) [2 Pts] Liesl stores her PyMC results in a 250×2 numpy array called `smps` (rows represent samples and the columns represent θ_1 and θ_2 respectively). Write 1-2 lines of Python code that uses this array to compute $P(\theta_1 > \theta_2 | x = x)$, or explain in one sentence or less why this is impossible. Answers longer than two lines of code or one sentence of explanation will not be graded.

6 Hierarchical Boba Model (4 pts)

Elaine owns K Boba shops in the Bay Area (assume $K > 1000$). Each shop sells two different kinds of tea: milk tea and fruit tea, and sells nothing else. **For this question, assume that every tea sold is either a milk tea or a fruit tea** (i.e., there are no teas that are both, and no other kinds of tea).

She gathers data for one day on how many orders of tea each shop sold, and how many of those orders are milk teas, and she defines the following variables, where $i \in \{1, \ldots, K\}$:

- $\phi_i \in [0, 1]$: proportion of daily orders at shop i that will be milk tea
- $r_i \in \{1, 2, \ldots\}$: total number of orders at shop i in a day
- $m_i \in \{1, \ldots, r_i\}$: number of orders at shop i in a day that are milk tea

For parts (a) - (d), assume r_i is fixed and known for each shop.

(a) [2 Pts] Fill in the blanks to define a Bayesian hierarchical model, assuming that s and t are fixed (nonrandom) parameters of the distribution for ϕ_i. You must choose a conjugate prior to receive full credit.

$$m_i|\phi_i \sim \text{Binomial}\Big(\underline{\qquad\qquad}, \underline{\qquad\qquad}\Big)$$

$$\phi_i \sim \underline{\qquad\qquad\qquad\qquad}(s, t)$$

(b) [2 Pts] Elaine correctly determines that given the model above, she can use exact inference to compute the posterior distribution over her unknown variables.

She considers several modifications to her model/setting. Which of the following modifications would **require** her to use approximate inference (i.e., exact inference would be impossible)? Select all answers that apply **by filling in the square next to each correct answer**.

- ☐ She triples the number of boba shops.
- ☐ She places a prior distribution on s and t, and also wants to infer the posterior over those variables.
- ☐ She also wants to infer the proportion of fruit tea sold at each shop.
- ☐ She wants to treat the total number of orders r_i at each shop as random variables and also infer the posterior over these.

7 Alien Detector (6 pts)

A NASA scientist uses Mars rover motion sensor data to calculate a daily motion measurement $x \in (0, 1)$. Smaller values of x correspond to typical dust patterns on Mars (H_0), but larger values of x are unusual and may indicate the presence of alien life (H_1).

The scientist formulates two hypotheses, each with a likelihood for x:

$$H_0 : p(x|H_0) = 2x \qquad\qquad H_1 : p(x|H_1) = 3x^2$$

(a) [1 Pt] Compute the likelihood ratio $LR(x) = \frac{p(x|H_1)}{p(x|H_0)}$.

For the remainder of the question, we use the decision rule "reject H_0 if the the observed value x is k times more likely under the alternative than under the null," for some $k \in (0, 1.5]$. Let $g(k)$ be the significance level (FPR) of this test, and let $h(k)$ be the power of this test.

(b) [3 Pts] Compute $g(k)$. You should leave your answer as a function of k, but you must simplify all integrals and/or algebraic expressions to receive full credit.

(c) [2 Pts] Which of the following must be true about $g(k)$ and $h(k)$ for all $k \in (0, 1.5]$? Select all answers that apply **by filling in the square next to each correct answer**.

Hint: this question can be answered without knowing the answer to the previous part.

- ☐ If we require a significance level of **at most** $g(k)$, then no other test can achieve a power greater than $h(k)$.
- ☐ If we require a significance level of **at least** $g(k)$, then no other test can achieve a power greater than $h(k)$.
- ☐ If we use the same null likelihood and test statistic threshold but a different alternative likelihood, then we will get **the same significance level** $g(k)$.
- ☐ If we use the same null likelihood and test statistic threshold but a different alternative likelihood, then we will get **the same power** $h(k)$.

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	$np(1-p)$	$\lfloor (n+1)p \rfloor$
$X \sim \text{Beta}(\alpha, \beta)$	$0 \leq x \leq 1$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{\alpha+\beta+1}$	$\frac{\alpha-1}{\alpha+\beta-2}$
$X \sim \text{Gamma}(\alpha, \beta)$	$x \geq 0$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\frac{\alpha-1}{\beta}$
$X \sim \mathcal{N}(\mu, \sigma^2)$	$x \in \mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$	μ	σ^2	μ
$X \sim \text{Exponential}(\lambda)$	$x \geq 0$	$\lambda \exp(-\lambda x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	0
$X \sim \text{InverseGamma}(\alpha, \beta)$	$x \geq 0$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$	$\frac{\beta}{\alpha-1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$	$\frac{\beta}{\alpha+1}$

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

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

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