

# Data 102, Fall 2024

## Midterm 1

- You have **110 minutes** to complete this exam. There are **6 questions**, totaling **54 points**.
- You may use **one**  $8.5 \times 11$  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 in the **middle** 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: \_\_\_\_\_

# 1 Minion Mayhem (13 pts)

Dr. Nefario, the chief mad scientist in Gru's lab, wants to prevent minions from eating too many bananas, so he builds a gadget called the Banana Analyzer Ray (BAR).

The BAR uses spectrometry to estimate the potassium concentration of minions (bananas are high in potassium). The observed potassium concentration for the  $i^{\text{th}}$  minion,  $x_i$ , is a random variable that depends on the number of bananas that minion eats. The more bananas, the higher the concentration  $x_i$  usually is.

For each minion, Dr. Nefario must decide based on the observed concentration  $x_i$  whether a minion has been eating too many bananas. To do so, he poses a hypothesis test. His hypotheses are:

- $H_0$  : Minion  $i$  has a healthy banana diet ( $\leq 3$  bananas a day). From baseline studies, Nefario further hypothesizes that, in this case,  $x_i \sim \text{Gamma}(\alpha, \beta)$  where  $\alpha = 4$  and  $\beta = 1000$ .
- $H_1$  : Minion  $i$  is overeating ( $> 3$  bananas a day).

By tracking Gru's banana vault, Dr. Nefario knows that  $1/3$  of all minions are eating too many bananas. From initial testing, Dr. Nefario knows that the sensitivity of his test is  $4/5$ , and its specificity is  $9/10$ .

- (a) [3 Pts] Label the rows and columns of the confusion matrix below, then fill in the entries with the *expected* number of FP, TP, FN, and TN events if the BAR is used on 30 minions.


- (b) [2 Pts] What is the probability that Dr. Nefario is wrong when he decides a minion has a healthy banana diet ( $\leq 3$  a day)?

- (c) [2 Pts] If fewer minions were overeating (i.e., less than  $1/10$ ), would your answer to part (b) increase, decrease, or stay the same? Justify your answer in two sentences or less.
- (d) [4 Pts] Pick which of the following statements are true about Dr. Nefario's null hypothesis statistical test. Select all answers that apply **by filling in the square next to each correct answer**.
- His null hypothesis  $H_0$  is a simple hypothesis.
  - If Dr. Nefario uses a  $p$ -value threshold of 0.1, then Dr. Nefario's true positive rate is 0.9.
  - The distribution of the  $p$ -values under the null is a Gamma distribution.
  - The alternative hypothesis is a composite hypothesis.
  - Dr. Nefario could choose a  $p$ -value threshold by studying the trade-off between the power and significance of his test (say by minimizing the Bayes risk)
- (e) [2 Pts] Suppose that Dr. Nefario runs his test on a minion and decides to reject the null hypothesis. Select the statement below that best explains his reasoning. Choose the single best answer **by filling in the circle next to it**.
- If the null hypothesis was true, then the observed concentration  $x_i$  would be very unlikely.
  - If the null hypothesis was true, then the observed concentration  $x_i$  would be unusually large.
  - Given the observed concentration  $x_i$ , it is very unlikely that the null hypothesis is true.
  - Given the observed concentration  $x_i$ , the alternative model would be more likely than the null.

## 2 Plantastic Trials (6 pts)

An agricultural researcher evaluates 100 different plant treatments for improving crop growth, and compares the results for each one to a control group using hypothesis tests. She obtains a  $p$ -value for each test: any treatment that significantly improves crop growth, according to her calculations, will be recommended in follow-up tests.

- (a) [3 Pts] The researcher defines the success rate as the expected fraction of recommended treatments that actually improve crop growth. Shade in the circle next to each correct answer in parentheses.
- Controlling the success rate is equivalent to controlling the ( *FWER*,  *FDR* ).
  - To control the success rate, use ( *Bonferroni*,  *Benjamini-Hochberg* ).
  - A high success rate requires both high sensitivity and high specificity, since lowering the ( *sensitivity*,  *specificity* ) increases the expected number of false positives, and lowering ( *sensitivity*,  *specificity* ) reduces the expected number of true positives.

- (b) [3 Pts] **Benjamini-Hochberg Procedure:**

The researcher decides to use the Benjamini-Hochberg (B-H) procedure to control the FDR at level  $\alpha$ . For this part only, she only looks at 5 treatments, and obtains the following 5  $p$ -values:

$$p_1 = 0.06, p_2 = 0.0001, p_3 = 0.0003, p_4 = 0.01, p_5 = 0.003,$$

For what values of  $\alpha$  will she recommend exactly 3 treatments for follow-up tests? For example, you may answer “If  $\alpha = 0$  or  $\alpha \in [0.05, 0.1]$ ”.

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

### 3 Neyman-Pearson (8 pts)

- (a) [1 Pt] The Neyman-Pearson Lemma suggests a choice of test statistic when distinguishing two hypotheses. To use Neyman-Pearson, how many of the hypotheses must be simple? Choose the single best answer **by filling in the circle next to it.**

- Both must be simple.
- At least one must be simple (one may be composite).
- Neither need be simple (both may be composite).

- (b) [2 Pts] Complete the sentence below by shading in the circle next to each correct answer in parentheses.

The Neyman-Pearson Lemma suggests a choice of test statistic that maximizes the ( *TPR*,  *TNR*) of a test given any chosen ( *FPR*,  *FNR*).

- (c) [2 Pts] The Neyman-Pearson Lemma suggests that the user use the *likelihood-ratio*, LR, as a test statistic, where  $X$  is the data, and where:

$$\text{LR}(X) = \frac{p(X|\text{alternative})}{p(X|\text{null})}.$$

**Note that this uses the convention where the alternative likelihood is in the numerator.**

In the space below, complete the specified decision rule by filling the blanks with  $<$  or  $>$ . Your decision rule should return correct decisions, and does not need to address the case where  $\text{LR} = 2$ .

$$\delta(X) = \begin{cases} \text{alternative} & \text{if } \text{LR}(X) \text{ \_\_\_\_ } 2 \\ \text{null} & \text{if } \text{LR}(X) \text{ \_\_\_\_ } 2 \end{cases}$$

Under this standard the user: “accepts the null if the \_\_\_\_\_ would be at least twice as likely under the \_\_\_\_\_ than the \_\_\_\_\_.”

- (d) [3 Pts] Suppose that we observe a sequence of i.i.d. random variables  $X = \{X_i\}_{i=1}^n$ , and aim to distinguish whether they were drawn from the null ( $H_0$ ) or the alternative ( $H_1$ ):

$$H_0 : X_i \sim \text{Gamma}(4, 1)$$

$$H_1 : X_i \sim \text{Exponential}(1).$$

In this case, which of the following is an optimal test statistic in the Neyman-Pearson sense (returns the same ROC as using the likelihood ratio): the **arithmetic mean**  $a(X)$ , or the **geometric mean**  $g(X)$ , of the observed samples? You must justify your answer to receive credit.

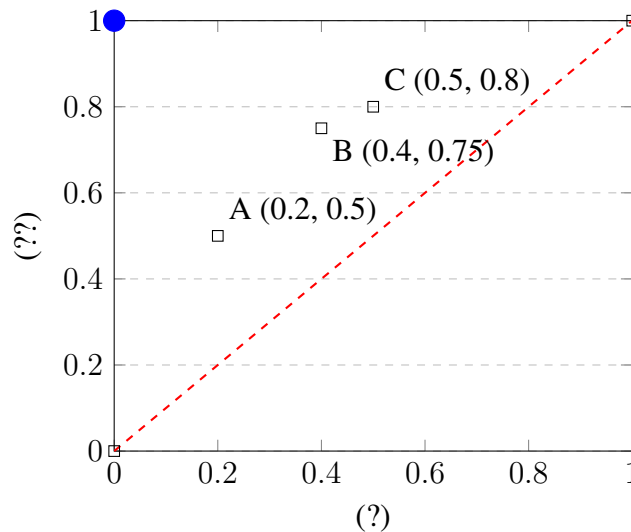
$$a(X) = \frac{1}{n} \sum_{i=1}^n X_i, \quad g(X) = \left( \prod_{i=1}^n X_i \right)^{1/n}$$

- Arithmetic mean  $a(X)$   
 Geometric mean  $g(X)$

**Justification:**

## 4 Decision Theory (9 pts)

A data scientist at a biomedical company is developing a new test for a viral disease. The scientists run three clinical trials to establish the power of the test for three different significance levels. They plot those three points on an ROC curve (A, B, and C below). They aim to use these points to determine an optimal operating point that minimizes the Bayes risk.



- (a) [1 Pt] The axes in the ROC plot are not labeled. In the spaces below, write in the appropriate axis label. Select from TPR, TNR, FPR, FNR, TDP, TOP, FDP, FOP:

Horizontal axis label: \_\_\_\_\_ Vertical axis label: \_\_\_\_\_

- (b) [2 Pts] Consider the filled circle in the upper left-hand corner, and the dashed diagonal line.
- What are the error rates (FPR and FNR) for a test located at the filled circle on the ROC plot?
  - Specify a test procedure that can achieve any point on the dashed diagonal line.

- (c) [2 Pts] Risk and Loss. Shade one option where options are given, otherwise, fill in the blank.

- Frequentist risk averages over ( *reality*,  *data*) given ( *reality*,  *data*).
- Bayesian Posterior risk averages over ( *reality*,  *data*) given ( *reality*,  *data*).
- Bayes risk averages over \_\_\_\_\_.



- (d) [1 Pt] Missing a case can have severe consequences, so false negatives are five times worse than false positives. Fill in the two blanks to define a loss function that corresponds to the description above.

$$\begin{cases} \ell(D = 1 | R = 0) = \underline{\hspace{2cm}} \\ \ell(D = 0 | R = 1) = \underline{\hspace{2cm}} \\ \ell(D = 0 | R = 0) = \ell(D = 1 | R = 1) = 0 \end{cases}$$

- (e) [3 Pts] Minimizing Risk. Suppose that the rate of cases is reported to be roughly 2,000 per 10,000 people. Which point along the ROC curve minimizes the risk?  
*Hint: you may assume that the optimal point is either A, B, or C.*

## 5 Shifting Prior-ities (9 points)

Kevin plans to attend an upcoming baseball game between the Exponential Eagles and the Geometric Giants. He always roots for whichever team he thinks will win, so he wants to figure out which of the two teams is more likely to win.

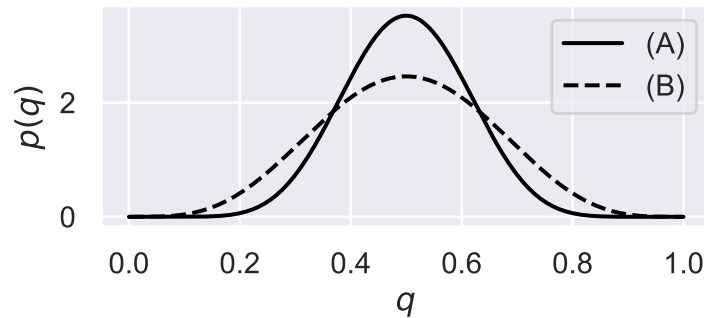
He acquires the match history between the two teams, and resolves to try to use this data to figure out the probability the Eagles will beat the Giants.

He assumes each game outcome (win or loss) is independent of the other games, and the game outcomes are identically distributed over the period of time when he has collected data. Accordingly, he decides to model the number of wins the Eagles have over the Giants as a binomial distribution with probability  $q$  - the probability the Eagles win.

- (a) [2 Pts] In Kevin's data the Eagles won 5 of their last 8 games against the Giants. Calculate the maximum likelihood estimator (MLE) for  $q$ . You do not need to re-derive the formula for the MLE if you know it.
- (b) [2 Pts] Which sentence correctly describes the MLE for this problem? Choose the single best answer **by filling in the circle next to it**.
- The MLE is the win probability  $q$  that is most likely given the data.
  - The MLE is the win probability  $q$ , which, if true, would make the observed data most likely.

For the remainder of this question, Kevin uses Bayesian inference instead.

- (c) [2 Pts] Kevin now decides to use a Beta distribution as his prior. He is deciding between a Beta(5, 5) prior and a Beta(10, 10) prior. The plot below shows these two prior distributions (assume they are plotted correctly). Which of the two curves shows the Beta(10, 10) distribution? Choose the single best answer **by filling in the circle next to it**.



- (A) Solid line     (B) Dashed line

- (d) [1 Pt] Kevin decides to proceed with the Beta(5, 5) prior. Compute the MAP estimator in this case and call it  $\hat{q}_{\text{MAP}}$ .

- (e) [2 Pts] Complete the sentences below by shading the circle next to the correct answer in the parentheses. Beneath each sentence, explain your answer *without computing* the new MAP.

1. Suppose that Kevin had used a Beta(10, 10) prior. Then his MAP estimator would be ( *closer to*,  *farther from*)  $1/2$  than  $\hat{q}_{\text{MAP}}$ .

**Explanation:**

2. Suppose Kevin used a Beta(5, 5) prior, but considered 80 games, of which the Eagles won 50. Then his MAP estimator would be ( *closer to*,  *farther from*) the MLE than  $\hat{q}_{\text{MAP}}$ .

**Explanation:**



## 6 Testing Taquerias (8 pts)

Olegario, a wealthy investor, is looking to build and invest in taquerias in the Bay Area. He collects the following data for every taqueria in fifteen Bay Area cities in a dataframe called `taquerias`. Note that some of the cities are large (e.g., San Francisco, San Jose), some are medium (e.g., Millbrae, Lafayette), and some are small (e.g., Colma, Atherton).

1. `city`, the city it's in
2. `logrevenue`, the log of the revenue from the taqueria from the last year
3. `rating`, the average rating on Yelp (assume this is a continuous value between 1 and 5).

He notices that the small cities only have 1-3 taquerias each, while the larger cities have hundreds. Olegario wants to open new taquerias in the city where they'll generate the highest revenue.

**For parts (a) and (b) only**, his friend suggests using the following line of Python code to find such cities:

```
taquerias.groupby('city')['logrevenue'].mean()
```

(a) [1 Pt] Is this a frequentist or Bayesian approach?

- Frequentist    Bayesian

(b) [2 Pts] Explain in two sentences or less why this approach might not give Olegario the best chance of success with his new taquerias.

*Hint: it might help to think about pooling in your answer, but you don't need to discuss it to earn full credit.*

For the remainder of the question, Olegario decides to use a hierarchical model. He defines the following variables:

- $z_i$  is the average log-revenue of a taqueria in city  $i$ .
- $y_{ij}$  is the log-revenue for taqueria  $j$  in city  $i$ .
- $\alpha$  is the average log-revenue for all taquerias across the entire Bay Area.

He decides on the following probability models.

$$z_i \mid \alpha \sim \mathcal{N}(\alpha, \sigma_z^2) \quad (10)$$

$$y_{ij} \mid z_i \sim \mathcal{N}(z_i, \sigma_y^2) \quad (11)$$

He treats  $\sigma_y$  and  $\sigma_z$  as fixed parameters. He does some background research to determine a reasonable value for  $\sigma_y$ .

- (c) [2 Pts] For this part only, he decides to treat  $\alpha$  as a fixed parameter. Which of the following describes the best empirical Bayes approach to choosing values for  $\alpha$  and  $\sigma_z$ ?
- Set  $\sigma_y := \sigma_z$  and  $\alpha$  to be the average log-revenue across all taquerias. First, compute the average log-revenue for *medium and large* cities only. Then, use maximum likelihood to estimate the mean and variance of that distribution, and use those estimates for  $\alpha$  and  $\sigma_z$  respectively.
  - First, compute the average log-revenue for *small* cities only. Then, use maximum likelihood to estimate the mean and variance of that distribution, and use those estimates for  $\alpha$  and  $\sigma_z$  respectively.

For the remainder of the question, assume he treats  $\alpha$  as a random variable with the following distribution:

$$\alpha \sim \mathcal{N}(r, \sigma_\alpha^2), \quad (12)$$

with  $r$  and  $\sigma_\alpha$  as fixed parameters.

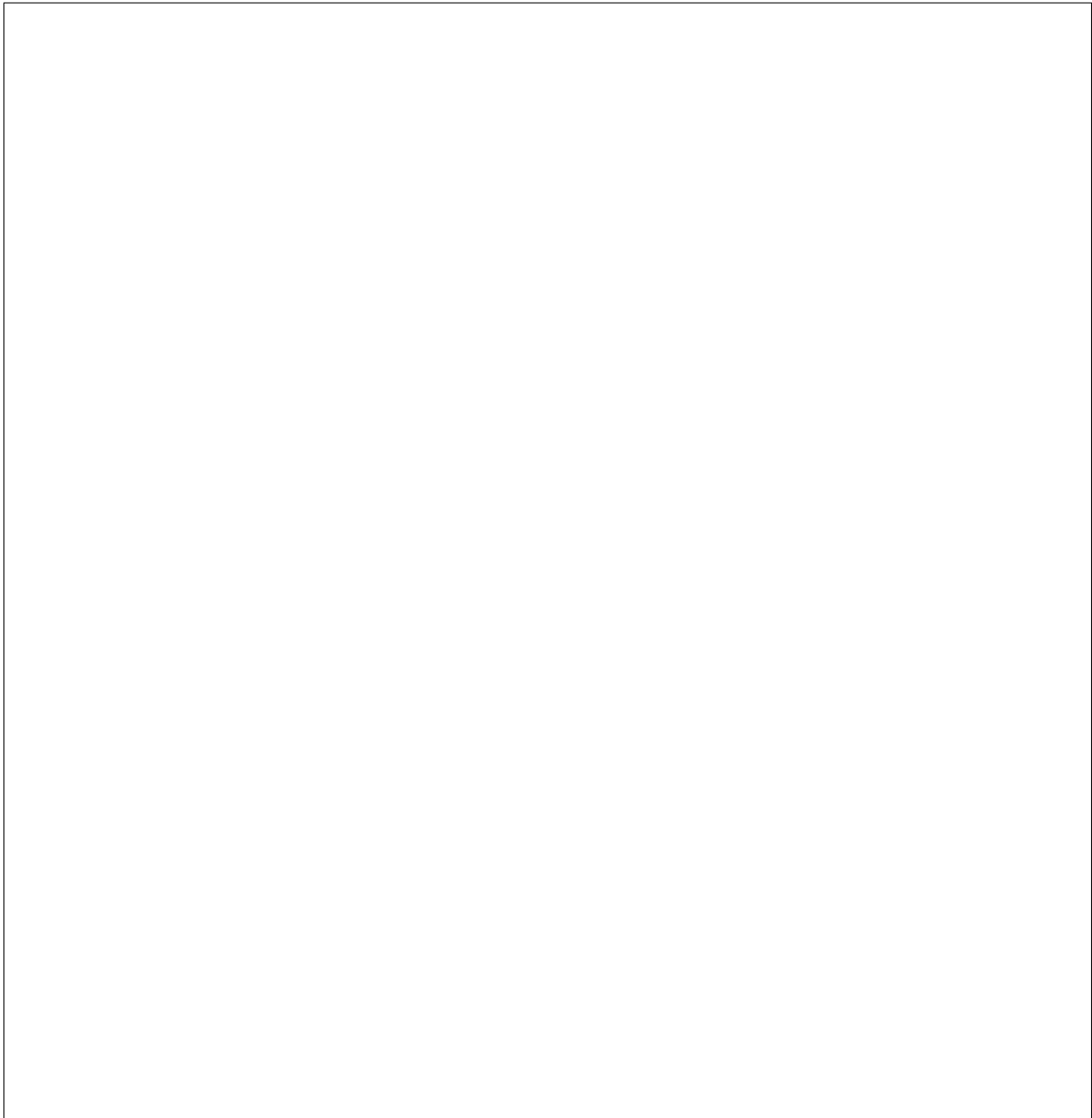
- (d) [3 Pts] Carefully draw a graphical model for this hierarchical model. Clearly label all components of your model. Your model should include all variables in equations (10), (11), and (12). **If you shade in any nodes, please only shade the border of the node so that we can read what's inside while grading.**

## 7 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
$X \sim \text{Poisson}(\lambda)$	$x = 0, 1, 2, \dots$	$\frac{\lambda^x e^{-\lambda}}{x!}$	$\lambda$	$\lambda$	$\lfloor \lambda \rfloor$
$X \sim \text{Binomial}(n, p)$	$x \in \{0, 1, \dots, n\}$	$\binom{n}{x} p^x (1-p)^{n-x}$	$np$	$np(1-p)$	$\lfloor (n+1)p \rfloor$
$X \sim \text{Beta}(\alpha, \beta)$	$0 \leq x \leq 1$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha}{\alpha+\beta} \frac{\beta}{\alpha+\beta} \frac{1}{\alpha+\beta+1}$	$\frac{\alpha-1}{\alpha+\beta-2}$
$X \sim \text{Gamma}(\alpha, \beta)$	$x \geq 0$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\frac{\alpha-1}{\beta}$
$X \sim \mathcal{N}(\mu, \sigma^2)$	$x \in \mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$	$\mu$	$\sigma^2$	$\mu$
$X \sim \text{Exponential}(\lambda)$	$x \geq 0$	$\lambda \exp(-\lambda x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$0$
$X \sim \text{InverseGamma}(\alpha, \beta)$	$x \geq 0$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$	$\frac{\beta}{\alpha-1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$	$\frac{\beta}{\alpha+1}$

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

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

## 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