Math 343 / 643 Spring 2025
Midterm Examination One

Professor Adam Kapelner

February 27, 2025

Full Name ____

Code of Academic Integrity

Since the college is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the college community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a student's performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using an unauthorized cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

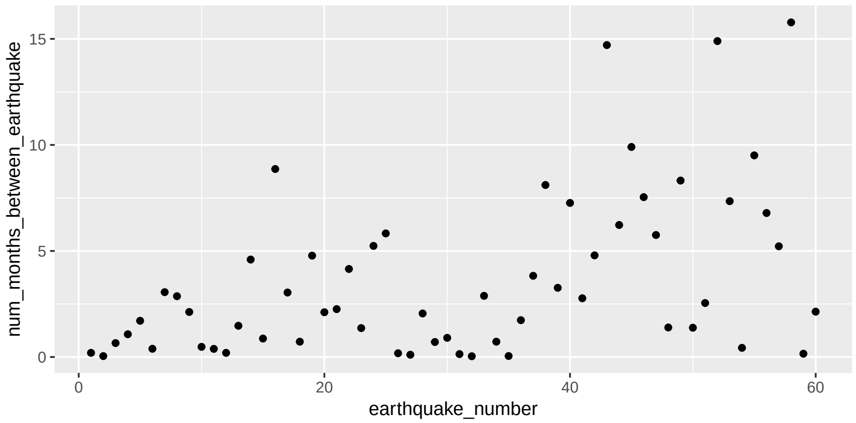
I acknowledge and agree to uphold this Code of Academic Integrity.

____ ________________
signature date

Instructions

This exam is 75 minutes (variable time per question) and closed-book. You are allowed **one** page (front and back) of a "cheat sheet", blank scrap paper (provided by the proctor) and a graphing calculator (which is not your smartphone). Please read the questions carefully. Within each problem, I recommend considering the questions that are easy first and then circling back to evaluate the harder ones. No food is allowed, only drinks.

Problem 1 We measure the time (in months) between $n = 60$ earthquakes. Here is a scatterplot of the raw data:



The time between earthquakes is known to be exponentially distributed. We suspect the distribution's mean changed once (at some point). So the DGP for the above data will be:

$$X_1, X_2, \ldots, X_{\theta_3} \overset{iid}{\sim} \theta_1 e^{-\theta_1 x} \;\text{ independent of }\; X_{\theta_3+1}, X_{\theta_3+2}, \ldots, X_n \overset{iid}{\sim} \theta_2 e^{-\theta_2 x}$$

For this problem, you may need to know some of these facts about the gamma distribution:

$$Y \sim \text{Gamma}(\alpha, \beta) := \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} \mathbb{1}_{y>0}, \quad F(y) = P(\alpha, \beta y), \quad S(y) = Q(\alpha, \beta y)$$

where P, Q are the lower and upper regularized gamma functions.

(a) [2 pt / 2 pts] How many parameters can you draw inference for in this DGP?

3

(b) [5 pt / 7 pts] Find the likelihood of the data, $f(\boldsymbol{x}; \theta_1, \theta_2, \theta_3)$. Denote the sample size as n i.e., do not use its known value of 60 here. You can assume all $x_i > 0$.

$$\begin{aligned} f(\boldsymbol{x}; \theta_1, \theta_2, \theta_3) &= \prod_{i=1}^{\theta_3} f(x_i; \theta_1) \prod_{\theta_3+1}^{n} f(x_i; \theta_2) \\ &= \prod_{i=1}^{\theta_3} \theta_1 e^{-\theta_1 x_i} \prod_{\theta_3+1}^{n} \theta_2 e^{-\theta_2 x_i} \\ &= \theta_1^{\theta_3} e^{-\theta_1 \sum_{i=1}^{\theta_3} x_i} \theta_2^{n-\theta_3} e^{-\theta_2 \sum_{i=\theta_3+1}^{n} x_i} \end{aligned}$$

(c) [3 pt / 10 pts] Assume Laplace's prior of indifference. State the prior for all parameters below. Use correct unambiguous notation.

$$f(\theta_1, \theta_2, \theta_3) \propto 1$$

(d) [3 pt / 13 pts] The posterior is denoted $f(\theta_1, \theta_2, \theta_3 \mid \boldsymbol{x})$. Find its kernel.

$$\begin{aligned} f(\theta_1, \theta_2, \theta_3 \mid \boldsymbol{x}) &\propto f(\boldsymbol{x}; \theta_1, \theta_2, \theta_3) f(\theta_1, \theta_2, \theta_3) \\ &= \left(\theta_1^{\theta_3} e^{-\theta_1 \sum_{i=1}^{\theta_3} x_i} \theta_2^{n-\theta_3} e^{-\theta_2 \sum_{i=\theta_3+1}^{n} x_i}\right)(1) \\ &= \theta_1^{\theta_3} e^{-\theta_1 \sum_{i=1}^{\theta_3} x_i} \theta_2^{n-\theta_3} e^{-\theta_2 \sum_{i=\theta_3+1}^{n} x_i} \end{aligned}$$

(e) [2 pt / 15 pts] Circle one: this kernel is the kernel of a distribution which is ... known / [✓ unknown]

(f) [4 pt / 19 pts] We wish to create a Gibbs sampler now. Find the conditional distribution $f(\theta_1 \mid \boldsymbol{x}, \theta_2, \theta_3)$ and give its brand name. Note that θ_1 is a mean time and thus it's greater than 0.

$$f(\theta_1 \mid \boldsymbol{x}, \theta_2, \theta_3) \propto \theta_1^{\theta_3} e^{-\theta_1 \sum_{i=1}^{\theta_3} x_i} \propto \text{Gamma}\left(\theta_3 + 1, \sum_{i=1}^{\theta_3} x_i\right)$$

(g) [4 pt / 23 pts] Find the conditional distribution $f(\theta_2 \mid \boldsymbol{x}, \theta_1, \theta_3)$ and give its brand name. Note that θ_2 is a mean time and thus it's greater than 0.

$$f(\theta_2 \mid \boldsymbol{x}, \theta_2, \theta_3) \propto \theta_2^{n-\theta_3} e^{-\theta_2 \sum_{i=\theta_3+1}^{n} x_i} \propto \text{Gamma}\left(n - \theta_3 + 1, \sum_{i=\theta_3+1}^{n} x_i\right)$$

(h) [6 pt / 29 pts] Find the conditional distribution $p(\theta_3 \mid \boldsymbol{x}, \theta_1, \theta_2)$. Assume the parameter space of θ_3 is $\{1, 2, \ldots, n-1\}$.

$$p(\theta_3 \mid \boldsymbol{x}, \theta_1, \theta_2) \quad \propto \quad \theta_1^{\theta_3} e^{-\theta_1 \sum_{i=1}^{\theta_3} x_i} \theta_2^{-\theta_3} e^{-\theta_2 \sum_{i=\theta_3+1}^{n} x_i}$$

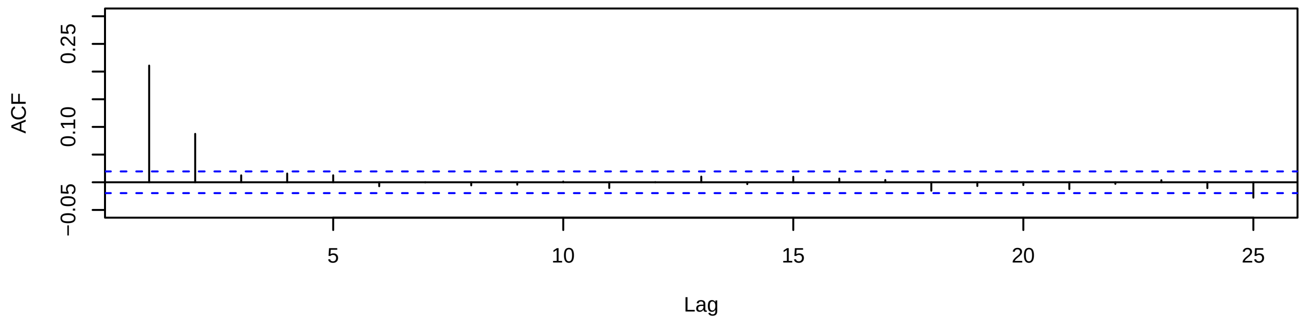
$$\text{Normalizing to sum to one,} \quad p(\theta_3 \mid \boldsymbol{x}, \theta_1, \theta_2) \quad = \quad \frac{\theta_1^{\theta_3} e^{-\theta_1 \sum_{i=1}^{\theta_3} x_i} \theta_2^{-\theta_3} e^{-\theta_2 \sum_{i=\theta_3+1}^{n} x_i}}{\sum_{t=1}^{n-1} \theta_1^{t} e^{-\theta_1 \sum_{i=1}^{t} x_i} \theta_2^{-t} e^{-\theta_2 \sum_{i=t+1}^{n} x_i}}$$

(i) [3 pt / 32 pts] Assume all conditional distributions are correct. We run the Gibbs Sampler for 10,000 runs and burn appropriately. Then we look at the autocorrelations for all chains (see next page). At what point should we thin if we wish to retain as many iterations as possible?

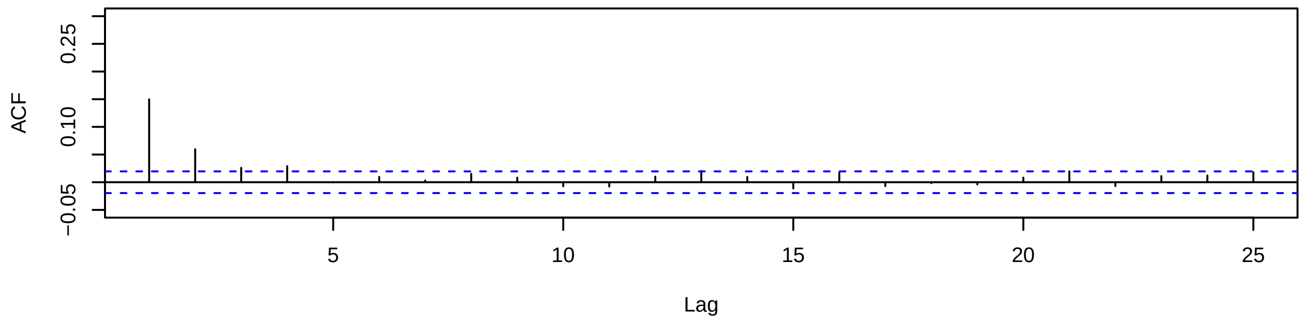
5

(j) [3 pt / 35 pts] After thinning, how many samples would be lift in the Gibbs chains approximately? (Assume we burned a very small number of iterations).

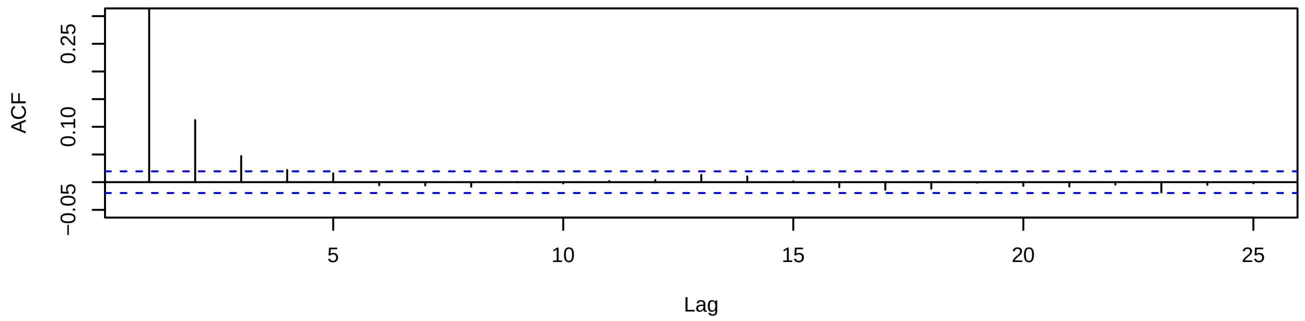
2000



Series theta1s



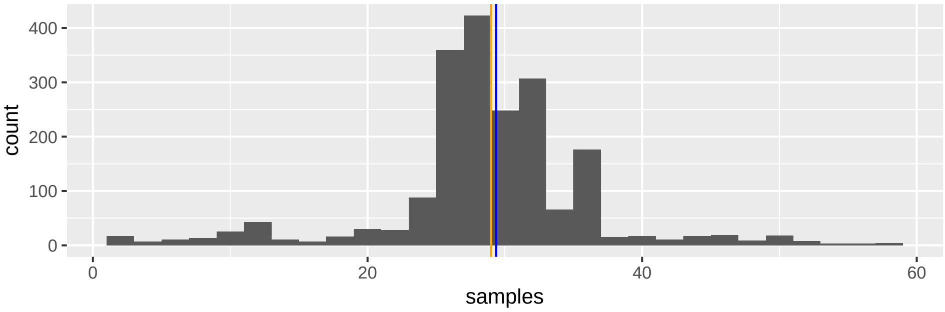
Series theta2s



Series theta3s

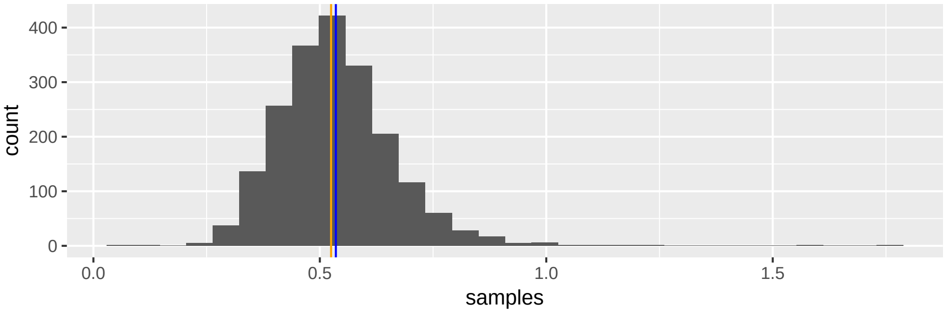
(k) [3 pt / 38 pts] Assume we thinned appropriately. Here is the histogram of the samples of θ_3 given $\boldsymbol{x}$. What is your point estimate of θ_3? Approximate it from the samples. No need to justify.

29



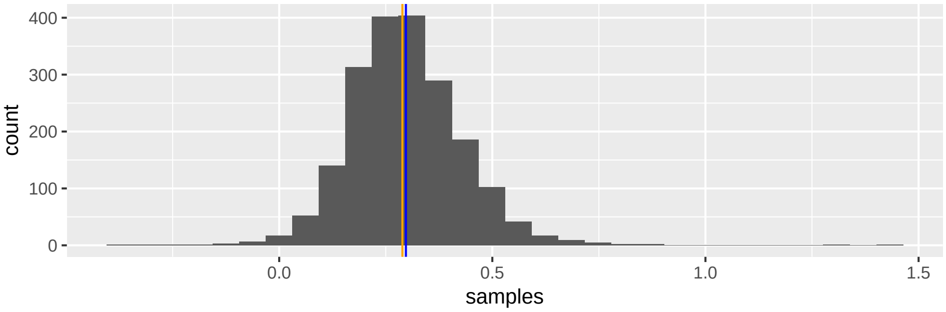
(l) [4 pt / 42 pts] Here is the histogram of the samples of θ_1 given $\boldsymbol{x}$. Provide a 95% CR for θ_1. Approximate it from the samples. No need to justify.

$[0.3, 0.8]$



(m) [4 pt / 46 pts] Here is the histogram of the samples of θ_1 given $\boldsymbol{x}$ minus the samples of θ_2 given $\boldsymbol{x}$. Provide a decision on $H_0 : \theta_1 = \theta_2$. No need to justify.

Reject H_0



Problem 2 We collect data on yearly returns (in percentage) of two financial assets. The returns come from DGP_1 and DGP_2 respectively. The sample sizes are $n_1 = 17$ and $n_2 = 37$. Some statistics are as follows: $\bar{x}_1 = 6.17$, $s_1 = 1.91$, $\bar{x}_2 = 6.28$, $s_2 = 3.09$.

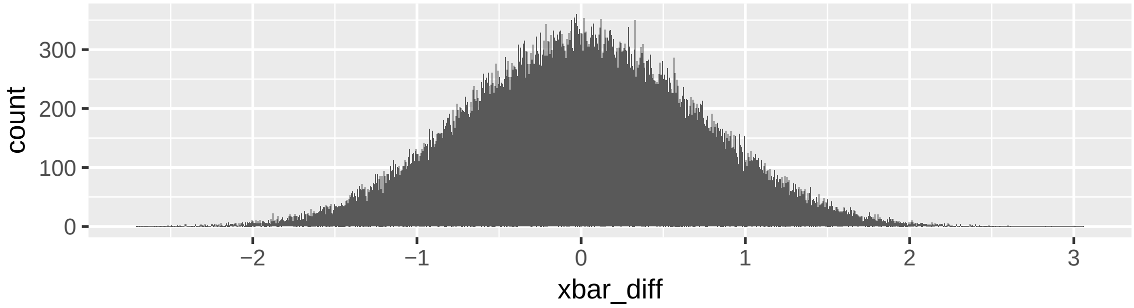
(a) [2 pt / 48 pts] We run a two-sample permutation test. What is the null hypothesis?

$$H_0 : DGP_1 = DGP_2$$

(b) [4 pt / 52 pts] If we were to run all the permutation samples, how many samples would there be? You can answer using any well-known mathematical notation. You do not need to explicitly provide the number itself.

$$\binom{17+37}{17} = \binom{17+37}{37} = \binom{54}{17} = \binom{54}{37}$$

(c) [4 pt / 56 pts] We run a permutation test using $B = 10^5$ with the test statistic defined as $\bar{x}_1 - \bar{x}_2$. Below is a histogram of the samples.



Test the hypothesis in (a) at $\alpha = 5\%$. Justify why you retained or rejected.

Retain H_0. The true test statistic from the raw data is $\bar{x}_1 - \bar{x}_2 = -0.11$ which falls within a 95% RET which estimated above looks to be about $[-1.5, 1.5]$.

(d) [4 pt / 60 pts] We run a permutation test using $B = 10^5$ with the test statistic defined as $s_1 - s_2$. Below are some quantiles for all the B test statistics:

```
> quantile(test_stats, c(0.025, 0.975))
     2.5%     97.5%
-1.241878  1.264630
```

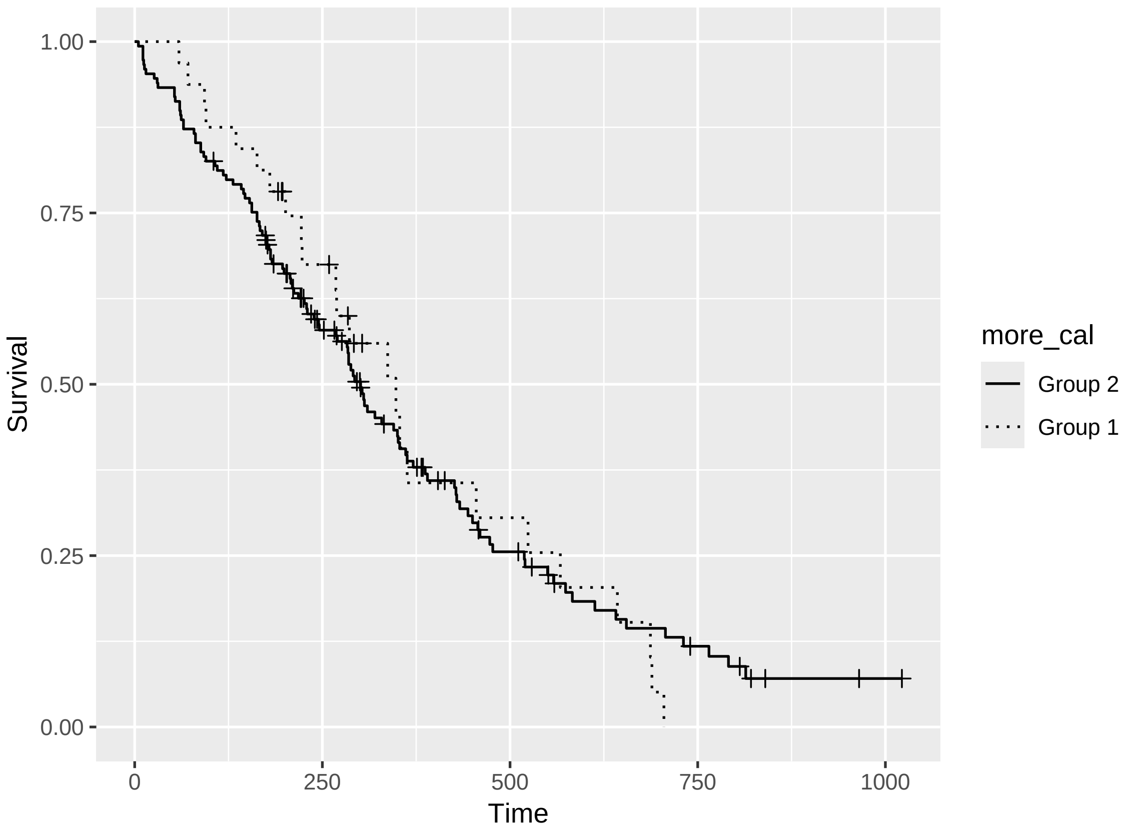
Test the hypothesis in (a) at $\alpha = 5\%$. Justify why you retained or rejected.

Retain H_0. The true test statistic from the raw data is $s_1 - s_2 = -1.18$ which falls inside the 95% RET given above.

(e) [2 pt / 62 pts] Circle one: did you have to make any parametric assumptions to run the test in (d)? yes / [✓ no]

(f) [2 pt / 64 pts] The tests in both (c) and (d) share the same H_0. Circle one: do the tests in (c) and (d) share the same power? yes / [✓ no]

Problem 3 We are interested in the differential survival (measured in days) of lung cancer patients by the amount of daily calories they consume in their diet. There are $n = 181$ total subjects in the study. Group 1 is those that consume more than 1200 per day (and there are $n_1 = 32$) and the number that consume 1000 calories or less per day is $n_2 = 149$. Below are the KM curves. The + signs indicate a censored observation.



(a) [6 pt / 70 pts] Assume for this question only there are *no censored observations* (even though we know it's not true). Provide an approximate 95% CI for survival more than 750 days in Group 2 (those that consume less than or equal to 1200 calories per day) to the nearest three digits.

$$CI_{S(y),1-\alpha} \approx \left[\hat{\hat{S}}(y) \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\hat{S}}(y)(1-\hat{\hat{S}}(y))}{n}}\right]$$

$$CI_{S(750),95\%} \approx \left[0.12 \pm 1.96\sqrt{\frac{0.12(1-0.12)}{149}}\right] = [0.068, 0.172]$$

(b) [2 pt / 72 pts] Circle one: do the censored observations need to have the same observed survival times? yes / [✓ no]

(c) [3 pt / 75 pts] Circle one: the observation with the greatest value for group 2 is censored? [✓ yes] / no

(d) [4 pt / 79 pts] Consider the following R code

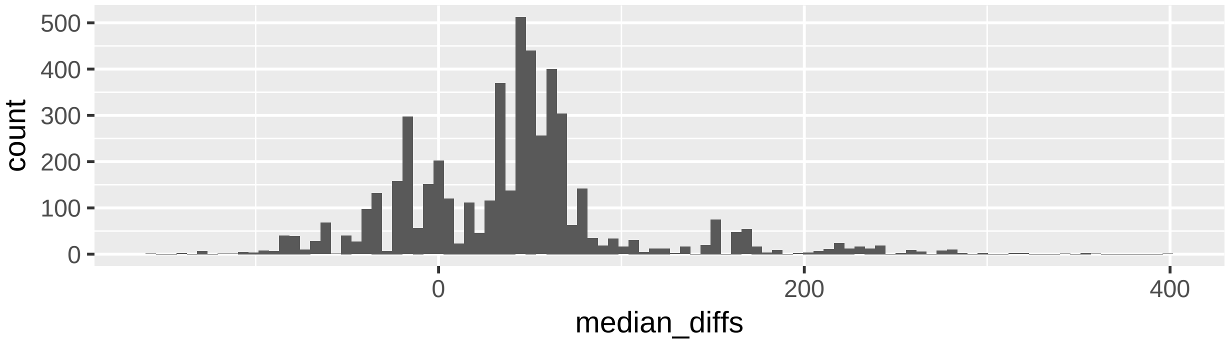
```
#n_1: number of subjects in group 1
#n_2: number of subjects in group 2
#y_1: survival times for group 1
#y_2: survival times for group 2
#c_1: censor vector for group 1
#c_2: censor vector for group 2

B = 5000
median_diffs = array(NA, B)
for (b in 1 : B){
  idx_1_b = sample(1 : n_1, n_1, replace = TRUE)
  idx_2_b = sample(1 : n_2, n_2, replace = TRUE)
  y_1_b = y_1[idx_1_b]
  c_1_b = c_1[idx_1_b]
  y_2_b = y_2[idx_2_b]
  c_2_b = c_2[idx_2_b]
  km_1 = survfit2(Surv(y_1_b, c_1_b) ~ 1)
  km_2 = survfit2(Surv(y_2_b, c_2_b) ~ 1)
  res_1 = summary(km_1)$table
  res_2 = summary(km_2)$table
  median_diffs[b] = res_1["median"] - res_2["median"]
}
```

What is (1) the type of test used here and (2) the null hypothesis?

(1) bootstrap (2) H_0 : the median survival in both groups are equal

(e) [6 pt / 85 pts] Below is a histogram of the B values of `median_diffs` from the previous R code. Run the test from the previous question at $\alpha = 5\%$ and justify the decision.



Retain H_0 since 0 is within the 95% bootstrap CI for median differences.

(f) [5 pt / 90 pts] We now wish to test H_0: the survival DGP for group 1 is the same as the survival DGP for group 2. We use the Log Rank test. Here is some relevant numeric output:

```
                      n Observed Expected (O-E)^2/E
more_cal=Group 2 149      110    109.8  0.000274
more_cal=Group 1  32       24     24.2  0.001243
```

Make a decision on H_0 at $\alpha = 5\%$.

The log rank statistic is the sum of the $(O - E)^2/E$ column which is near zero, i.e. much less than the χ^2_1 cutoff at $\alpha = 5\%$ which is $1.96^2 = 3.84$. Hence we retain H_0.

(g) [8 pt / 98 pts] We now fit an iid Gamma model to group 2's survival data i.e. we assume the positive values $\boldsymbol{y} :=< y_{2,1}, y_{2,2}, \ldots, y_{2,n_2} >$ were drawn from the DGP

$$Y_{2,1}, Y_{2,2}, \ldots, Y_{2,n_2} \overset{iid}{\sim} \text{Gamma}(\theta_1, \theta_2)$$

and we also have the censoring binary data $\boldsymbol{c} :=< c_{2,1}, c_{2,2}, \ldots, c_{2,n_2} >$. Write the likelihood for the parameters. You may need to use the facts about the gamma distribution (see Problem 1). No need to simplify.

$$\begin{aligned} \mathcal{L}(\theta_1, \theta_2; \boldsymbol{y}, \boldsymbol{c}) &= \prod_{\{i: c_i = 1\}} f(y_i; \theta_1, \theta_2) \prod_{\{i: c_i = 0\}} S(y_i) \\ &= \prod_{\{i: c_i = 1\}} \frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} y_i^{\theta_1 - 1} e^{-\theta_2 y_i} \prod_{\{i: c_i = 0\}} Q(\theta_1, \theta_2 y_i) \end{aligned}$$

(h) [2 pt / 100 pts] Circle one: does the likelihood in the previous question have a closed form solution for either parameter? yes / [✓ no]