EDUC 452 PS1

###Your name goes here###

2025-04-06 11:51:26.432527

This homework is due by **Friday, April 18th, 8:00am**. Upload a html file to Canvas called `ps1.html`

Tip 1: Code: https://github.com/ben-domingue/educ452/tree/main/problemsets/ps1

Tip 2: Questions in red. If a subquestion has no red, no response is needed.

1 Question 1 Bernoulli Distribution

Simulations with classic distributions [see distributions_bernoulli.R]: A Bernoulli distribution [https://en.wikipedia.org/wiki/Bernoulli_distribution] is useful when you want to simulate a variable that takes one of two states (e.g., true/false, correct/incorrect, etc); we'll generically denote these states as 0 and 1. It has one parameter which is the probability of the distribution generating a 1. If we call that parameter p, the probability of the distribution generating a 0 is 1-p so we've completely characterized the distribution.

1.1 1A

The proportion of 1s in a sample turns out to be a decent guess at p (it's actually the maximum likelihood estimate, but that's a story for another day!). As with all estimates, you can imagine that the quality of our estimate of p depends upon the sample size.

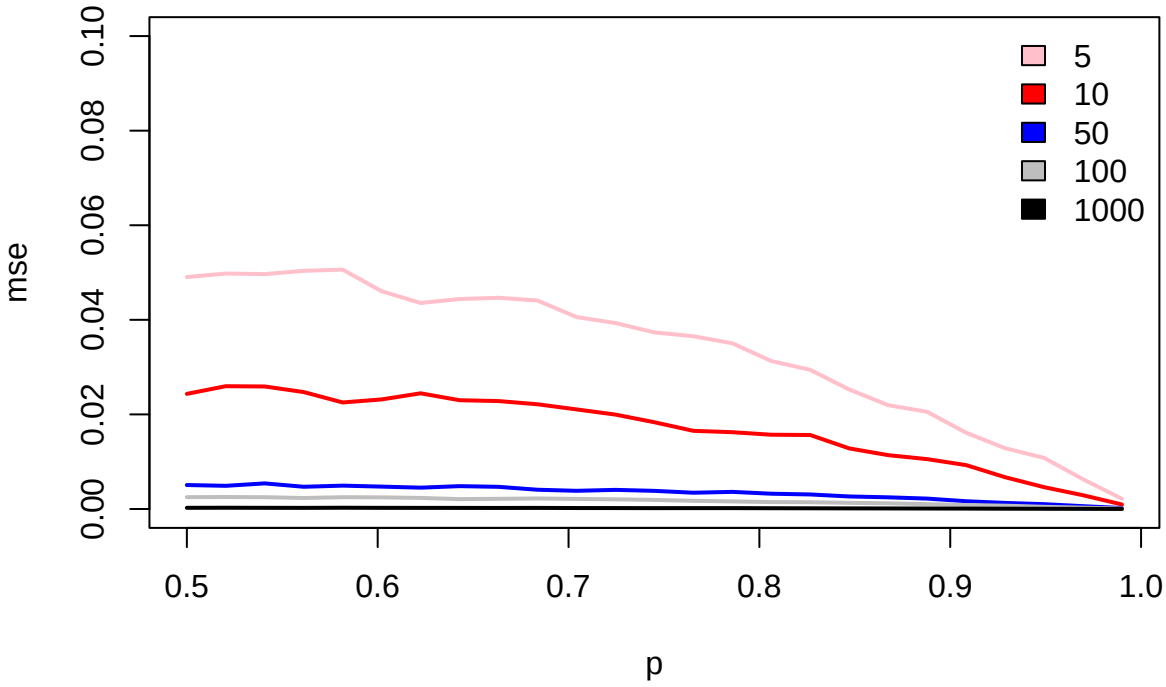
Please describe the interplay between the magnitude of p, the sample size, and the resulting quality of the mean as an estimator for p.

```
##############################################################################################################
##A. bernoulli distribution
##a. how good of an estimate of p is the mean? interplay between the magnitude of p and the sample size

p.est<-list()
for (N in c(5,10,50,100,1000)) {
    for (p in seq(0.5,0.99,length.out=25)) {
        m<-numeric()
        for (i in 1:1000) {
            coins<-rbinom(N,1,p)
            m[i]<-mean(coins)
        }
        p.est[[paste(N,p)]]<-c(N,p,mean((m-p)^2))
    }
}

x<-data.frame(do.call("rbind",p.est))
L<-split(x,x[,1])
cols<-c("pink","red","blue","gray","black")
plot(NULL,ylim=c(0,0.1),xlim=range(x[,2]),xlab='p',ylab='mse')
```

```
for (i in 1:length(L)) lines(L[[i]][,-1],lwd=2,col=cols[i])
legend("topright",bty='n',fill=cols,names(L))
```

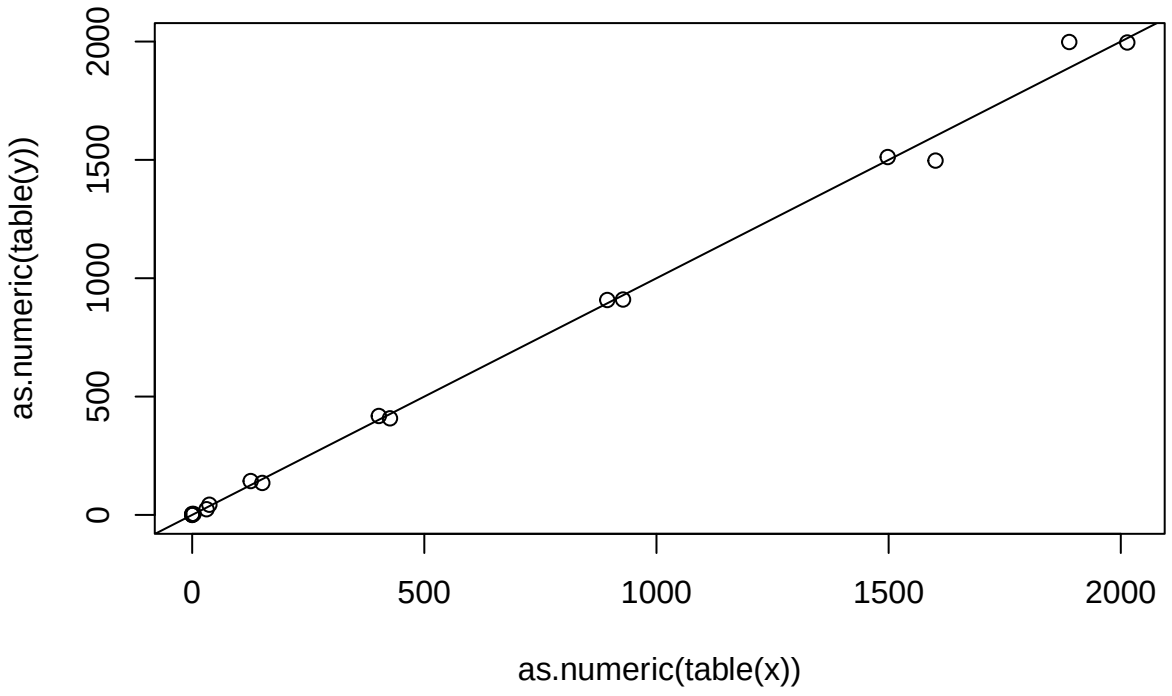


YOUR ANSWER HERE

1.2 1B

Connection to Binomial distribution: The reason that the R command to simulate data from the Bernoulli distribution is "rbinom" is because the "binomial" distribution (hence binom) is a generalized form of the Bernoulli. The binomial distribution asks how many 1s we get from observing n independent draws from a Bernoulli process with parameter p. Let's take a look at how we can translate back-and-forth between these two views.

```
##############################################################################################################
##b. the binomial distribution
ntrial<-15
x<-rbinom(10000,ntrial,.5)
y<-numeric()
for (i in 1:10000) {
    z<-numeric()
    for (j in 1:ntrial) z[j]<-rbinom(1,1,.5)
    y[i]<-sum(z)
}
x<-factor(x,levels=0:ntrial)
y<-factor(y,levels=0:ntrial)
plot(as.numeric(table(x)),as.numeric(table(y))); abline(0,1)
```

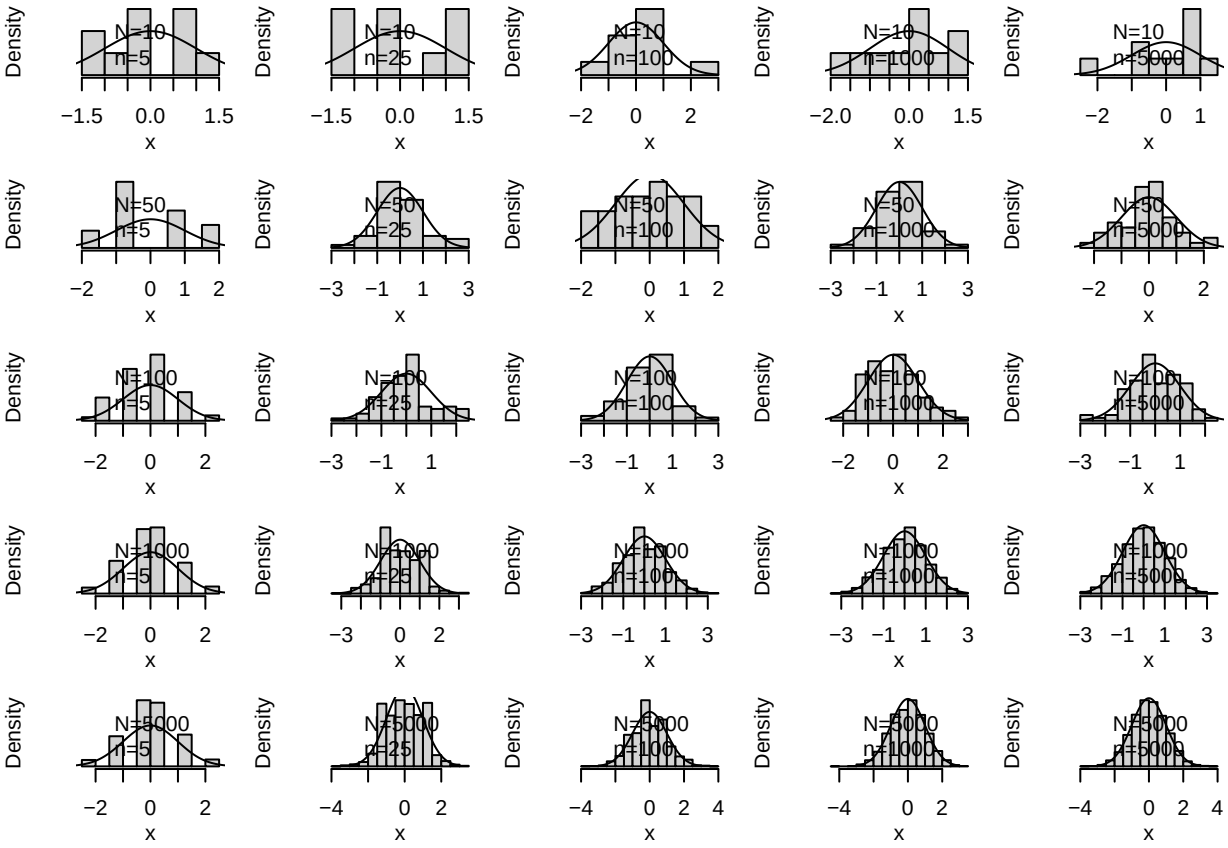


1.3 1C

Connection to CLT: The CLT is a wonderfully useful theorem that tells us under what conditions sums of things converge to normal distributions.

Take a look at the code; which parameter is relevant for applying the CLT? [Note: This is also meant to ensure you're getting increasingly familiar with a range of distributions!]

```
##########################################################################################################################
##c. the CLT, https://en.wikipedia.org/wiki/Central_limit_theorem
#Which is relevant, N or n, in terms of having the CLT apply?
fun<-function(N,n,p=0.5) { #N is the number of people, n is the number of tosses per person
    x<-rbinom(N,n,p)
    x<-(x-mean(x))/sd(x)
    hist(x,main='',freq=FALSE,yaxt='n')
    xv<-seq(-3,3,length.out=1000)
    lines(xv,dnorm(xv))
    legend("topleft",bty='n',c(paste0("N=",N),paste0("n=",n)))
}
par(mfrow=c(5,5),mar=c(3,3,1,1),mgp=c(2,1,0),oma=rep(0,4))
for (N in c(10,50,100,1000,5000)) for (n in c(5,25,100,1000,5000)) fun(N,n)
```



YOUR ANSWER HERE

2 Question 2 Normal Distribution

More than any other distribution, we're going to all become BFF with the normal distribution this quarter. To get started, let's experiment with a few things.

2.1 2A Univariate normal distribution

i. Let's first remind ourselves where the magical value of 1.96 comes from.

```
##Let's attempt to recover some of the most useful facts about the normal distribution from simulation

##a.i. What proportion of the distribution is more than 1.96 from the mean(=0)?
x<-rnorm(50000,mean=0,sd=1)
sum(abs(x)>1.96)/length(x) #what is this telling us?
```

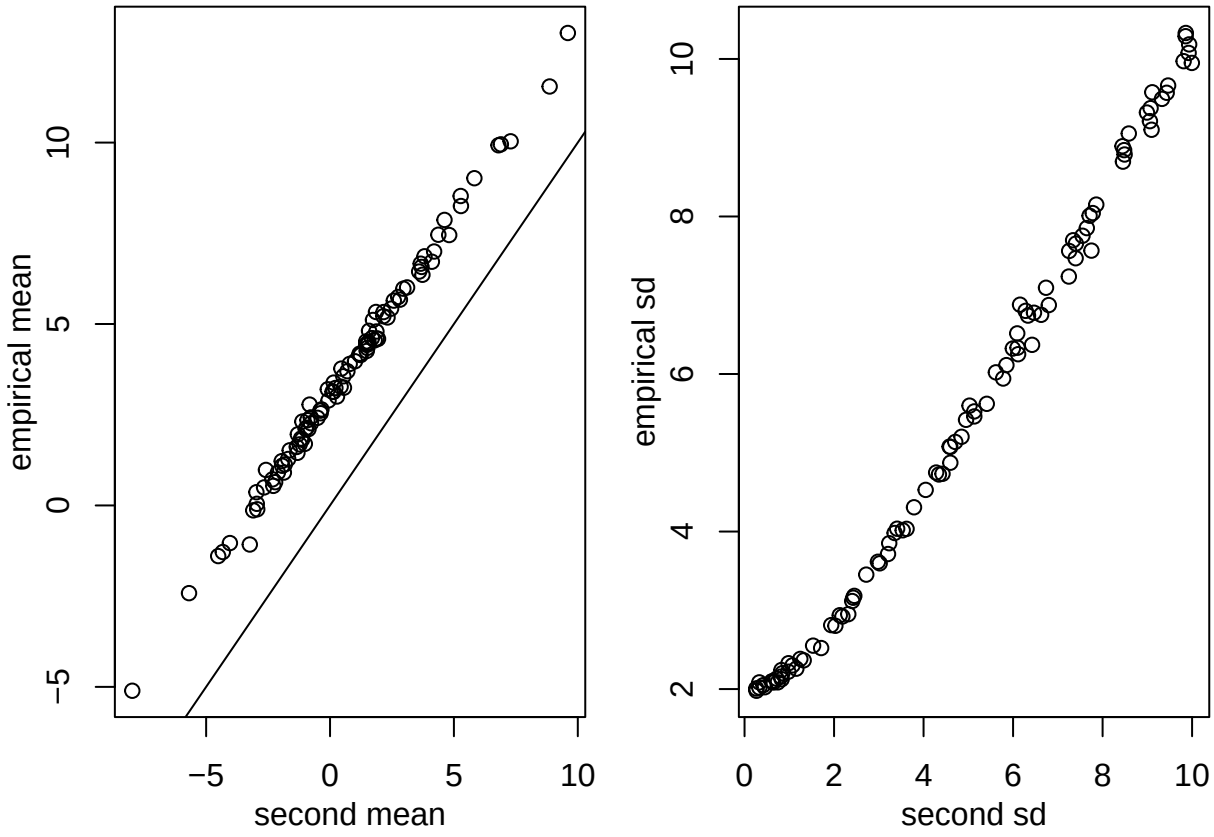
```
## [1] 0.051
```

ii. Suppose x1 has a normal distribution and x2 has a normal distribution (with different mean and variance). What is the distribution of x1+x2? We can get an analytical solution to this question, but let's see how we can also examine this via simulation.

```
##a.ii What does the sum of two normal variables look like?
N<-1000
m1<-3
sd1<-2
m2<-rnorm(100,mean=0,sd=3)
```

```r
sd2<-runif(100,0.1,10)
out<-list()
for (i in 1:length(m2)) {
    x1<-rnorm(N,mean=m1,sd=sd1)
    ##
    x2<-rnorm(N,mean=m2[i],sd=sd2[i])
    ##
    y<-x1+x2
    m<-mean(y)
    s<-sd(y)
    out[[i]]<-c(m,s)
}
z<-do.call("rbind",out)

par(mfrow=c(1,2),mgp=c(2,1,0),mar=c(3,3,1,1))
plot(m2,z[,1],xlab="second mean",ylab="empirical mean"); abline(0,1)
plot(sd2,z[,2],xlab="second sd",ylab="empirical sd")
```



YOUR ANSWER HERE

Note: Analytic derivations of these facts would be preferable for any number of reasons. The goal here is to get our hands dirty with some real simulations but simulations are not always the right tool for the job!

2.2 2B

We're also going to work with the bivariate normal distribution. You'll see this in, for example, question 3. But let's start with a simpler question.

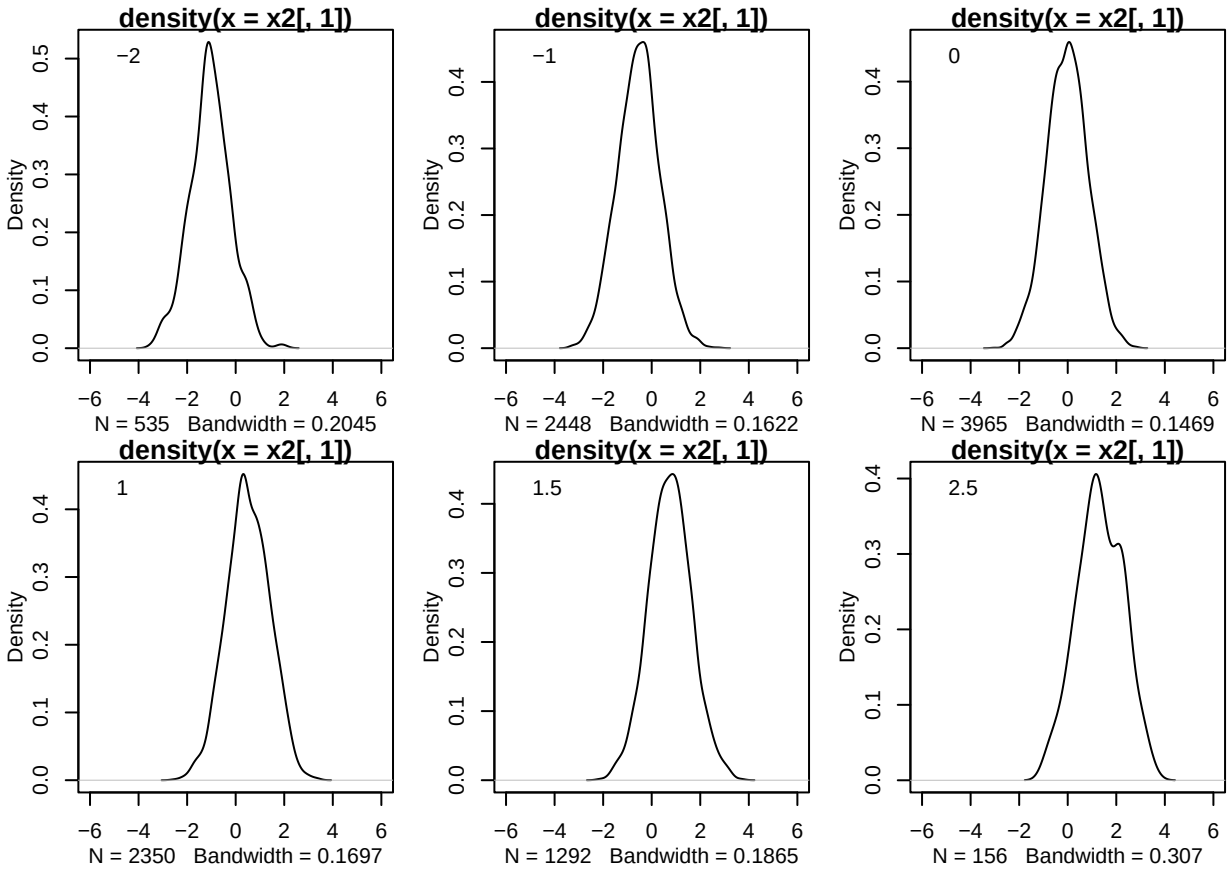
If you have a bivariate normal distribution f(x1,x2), what is f(x1|x2)?

```
##b. Now let's look at a bivariate normal distribution
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
x<-mvrnorm(500000,mu=rep(0,2),Sigma=matrix(c(1,.5,.5,1),2,2))
##What is the distribution of a conditional look like? So if we look at the probability density when th
par(mfrow=c(2,3),mgp=c(2,1,0),mar=c(3,3,1,1))
for (val in c(-2,-1,0,1,1.5,2.5)) {
    x2<-x[abs(x[,2]-val)<.01,]
    plot(density(x2[,1]),xlim=c(-6,6))
    legend("topleft",bty='n',legend=val)
}
```




```
##What do you think?
##Note the centers of these distributions on the x-axis. What would you need to change in the above to
```

YOUR ANSWER HERE

3 Question 3 Multinomial Distribution

We'll often use a simplified version of the Multinomial distribution (https://en.wikipedia.org/wiki/Multinomial_distribution). The Multinomial distribution is a generalized version of the Binomial distribution where each of the n independently observed processes is not Bernoulli but one that generates one of k outcomes. We'll use n=1 and then this distribution will generate outcomes for each person.

3.1 3A

So, for example, can you use rmultinom to generate language backgrounds for 1000 students (let's say 40% of the students grew up in primarily English-speaking homes, 30% in primarily Spanish-speaking homes, and 30% in primarily Tagalog-speaking homes)?

```
### YOUR CODE HERE ###
```

3.2 3B

Can you now introduce another person-level covariate x such that the overall proportions (40%/30%/30%) are the same but an individual's probability of being in a home of a given kind depends on x?

```
### YOUR CODE HERE ###
```

4 Question 4 Linear Regression

Linear regression. I simulate a dataset in a straightforward way based on the linear regression model.

```
set.seed(8675309)
library(MASS)
xz<-mvrnorm(10000,mu=rep(0,2),Sigma=diag(1,2))
x<-xz[,1]
z<-xz[,2]
y<-.5*x+.7*z+rnorm(length(x))
df<-data.frame(x=x,z=z,y=y)

m<-lm(y~x+z,df) #note that here we are observing z!! different than what we previously considered here.
```

Building on what is in linear_regressions.R, please put together a small example wherein you show the sensitivity of the parameter estimates and associated quantities in the "m" model to assumptions about the correlation between x & z. - In class, we looked at what happens when you don't observe z. This led to bias in estimates of association between x and the outcome. - Now, we'll observe z. This will lead to a subtler problem. - You'll need to vary the structure shown above to induce a problem. In particular, if you change the way we simulate the xz variable, you can generate interesting behavior in the regression outputs. - Explain your example.

```
### YOUR CODE HERE ###
```

YOUR ANSWER HERE

5 Question 5 Logistic Regression

Logistic regression supposes that a binary outcome y depends on some predictor in the following way: $\Pr(y = 1 \mid x) = \frac{1}{1+\exp(-(b_0+b_1 x))}$

For and b_1=1, let's consider one of the 'sinister problems' from Class1. In particular, suppose that the covariate x is observed with measurement error. Through the use of simulation studies along the lines of what we considered in class, I want you to consider the following question: is measurement error a bigger

problem when b_0=0 or when b_0=2? Notes: - I'm allowing some flexibility in how you conceptualize what 'bigger problem' means. - You will need to simulate outcomes y using rbinom (but where the parameters you pass to rbinom depend upon x).

```
### YOUR CODE HERE ###
```

YOUR ANSWER HERE

6 Session info

Information about this R session including which version of R was used, and what packages were loaded.

```
sessionInfo()
```

```
## R version 4.4.3 (2025-02-28)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 20.04.6 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C               LC_TIME=en_US.UTF-8
##  [4] LC_COLLATE=en_US.UTF-8     LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                  LC_ADDRESS=C
## [10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## time zone: America/Los_Angeles
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] MASS_7.3-63     lubridate_1.9.4 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [6] purrr_1.0.2     readr_2.1.5     tidyr_1.3.1     tibble_3.2.1    ggplot2_3.5.1
## [11] tidyverse_2.0.0 knitr_1.49
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.6      compiler_4.4.3    tinytex_0.54      tidyselect_1.2.1
##  [5] scales_1.3.0      yaml_2.3.10       fastmap_1.2.0     R6_2.5.1
##  [9] generics_0.1.3    munsell_0.5.1     pillar_1.10.1     tzdb_0.4.0
## [13] rlang_1.1.4       stringi_1.8.4     xfun_0.50         timechange_0.3.0
## [17] cli_3.6.3         withr_3.0.2       magrittr_2.0.3    digest_0.6.37
## [21] grid_4.4.3        rstudioapi_0.17.1 hms_1.1.3         lifecycle_1.0.4
## [25] vctrs_0.6.5       evaluate_1.0.1    glue_1.8.0        colorspace_2.1-1
## [29] rmarkdown_2.29    tools_4.4.3       pkgconfig_2.0.3   htmltools_0.5.8.1
```