ESS 575: Probability Labs

Team England

16 September, 2022

Contents

Team England:

- Caroline Blommel
- Carolyn Coyle
- Bryn Crosby
- George Woolsey

cblommel@mail.colostate.edu, carolynm@mail.colostate.edu, brcrosby@rams.colostate.edu, george.woolsey@colostate.edu

Probability Lab 1: Chain rule, joints distributions, and DAGs

Questions

The lab assignment can be found at this link.

Answers

Completed work for 2022 ESS-575 Team England was uploaded to Dropbox here.

Probability Lab 2 - Probability Distributions

Question 0

Barn swallows form pair bonds (male/female pairings) in the spring before mating season. Each male/female pair share a nest and care for the offspring of the female. Often a number of the female's offspring were sired by males other than her mate. We are interested in the random variable, the number of offspring sired by the females mate. Suppose previous literature suggests the probability an offspring's father is the female's mate is 0.8. Plot the probability mass function.

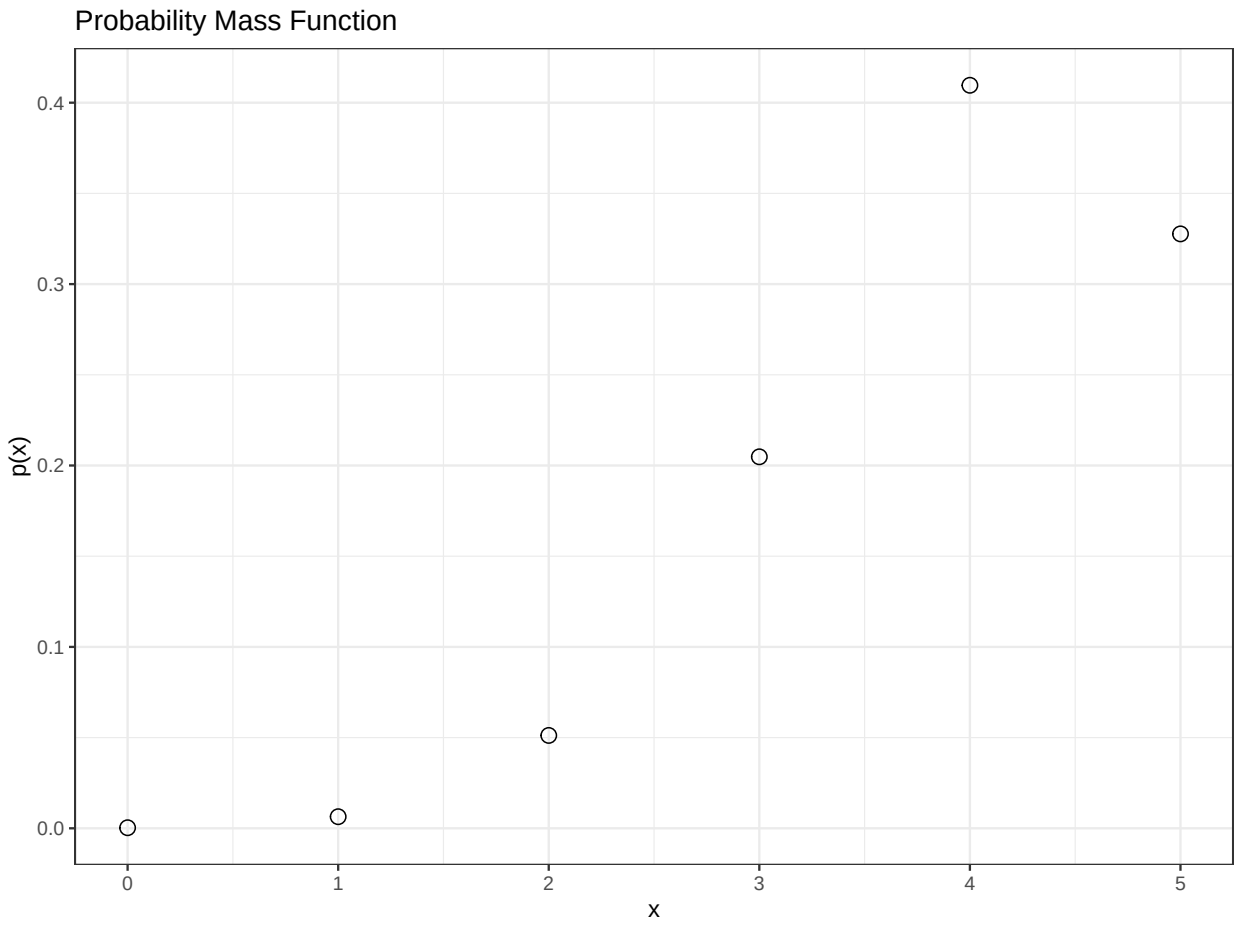
Answer

Write a model describing how the data arise.

$$y_i \sim \mathsf{binomial}(5, 0.8) \tag{}$$

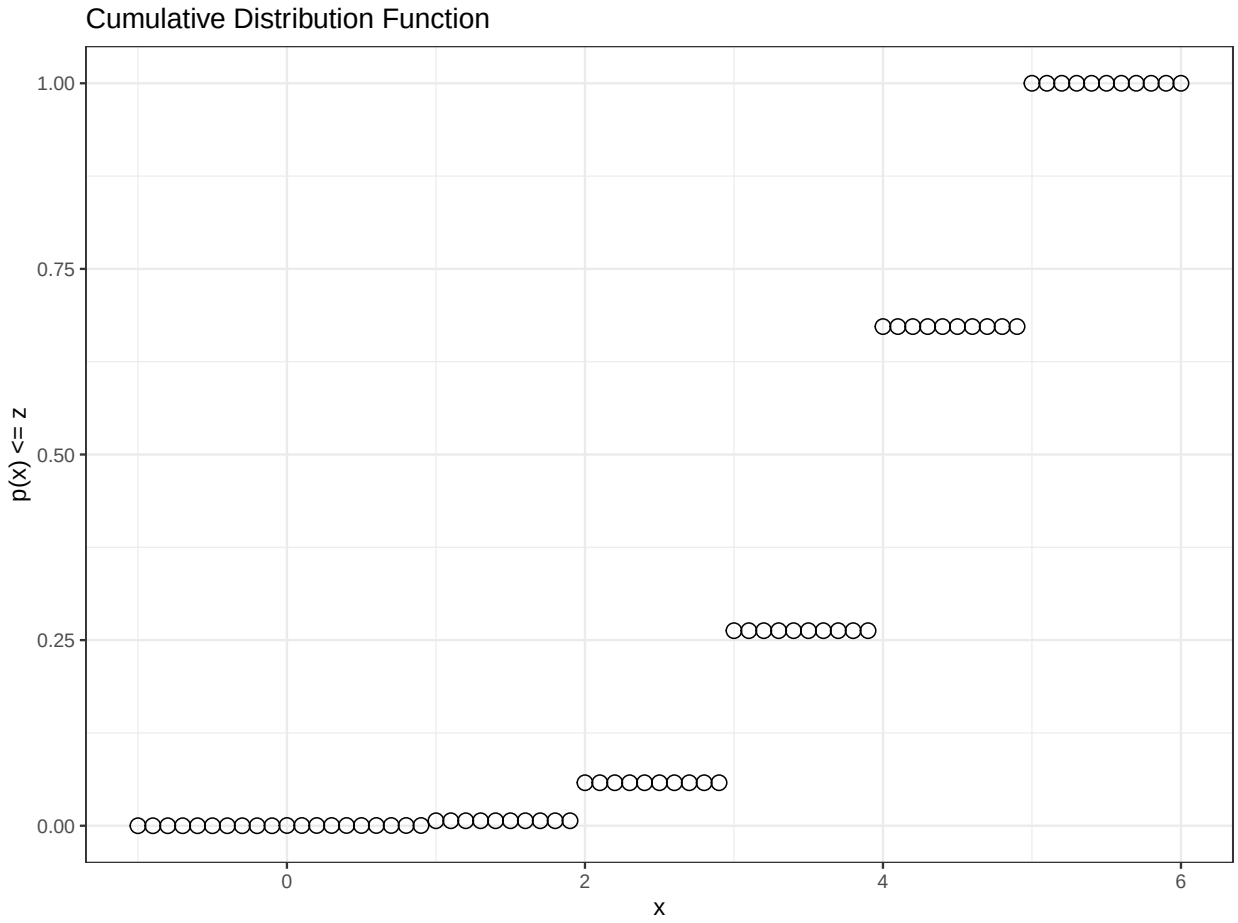
Plot the probability mass function.

```
p <- 0.8
x <- seq(0, 5, 1)
y <- dbinom(x = x, size = 5, prob = p)
ggplot(data.frame(x,y), aes(x = x, y = y)) +
  geom_point(shape = 1, size = 3) +
  labs(title = "Probability Mass Function") +
  xlab("x") +
  ylab("p(x)") +
  theme_bw()
```



Plot the cumulative distribution function for values between -1 and 6 in steps of .1 using pbinom. (Don't worry about the unfilled circles bit or making a fancy plot. Just get the concepts.)

```
p <- 0.8
x <- seq(-1, 6, 0.1)
y <- pbinom(q = x, size = 5, prob = p)
ggplot(data.frame(x,y), aes(x = x, y = y)) +
  geom_point(shape = 1, size = 3) +
  labs(title = "Cumulative Distribution Function") +
  xlab("x") +
  ylab("p(x) <= z") +
  theme_bw()
```



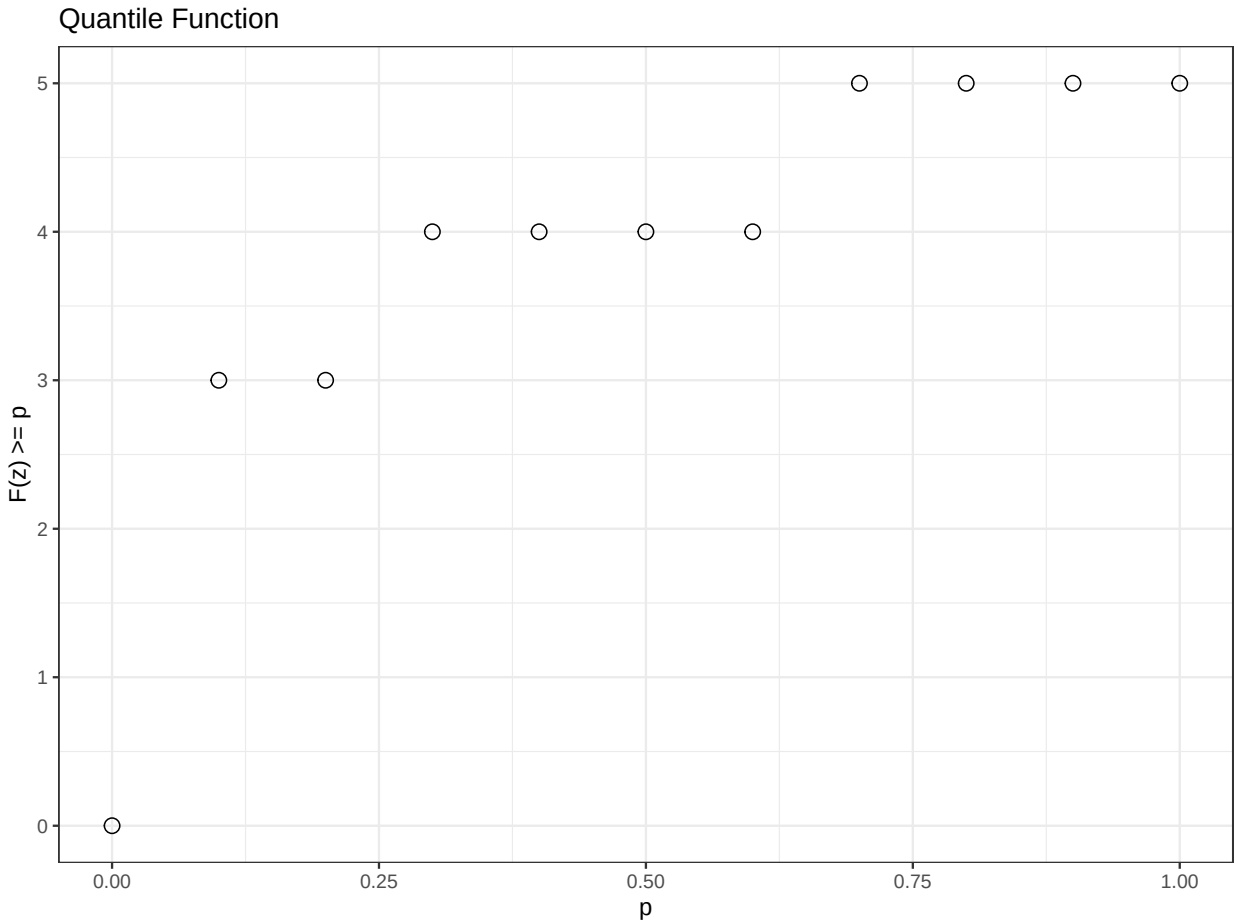
Write a mathematical expression for the the probability that there are fewer than four offspring sired by the female's mate and compute the probability using dbinom and pbinom.

```
p <- 0.8
x <- seq(0, 3, 1)

sum_dbinom <- sum(dbinom(x = x, size = 5, prob = p))
p_binom <- pbinom(q = 3, size = 5, prob = p)
```

The probability that there are fewer than four offspring sired by the female's mate: **26.3%**

Plot the quantile function over the range of zero to 1 in steps of .1 using qbinom.

```
p <- 0.8
x <- seq(0, 1, 0.1)
y <- qbinom(p = x, size = 5, prob = p)
ggplot(data.frame(x,y), aes(x = x, y = y)) +
  geom_point(shape = 1, size = 3) +
  labs(title = "Quantile Function") +
  xlab("p") +
  ylab("F(z) >= p") +
  theme_bw()
```



Compute the smallest number of offspring with a probability equal to or exceeding .3

```
p <- 0.8
x <- 0.3
q_binom <- qbinom(p = x, size = 5, prob = p)
```

The smallest number of offspring with a probability equal to or exceeding .3 is: **4**

Question 1

We commonly represent the following general framework for linking models to data:

$$[y_i \mid g(\theta, x_i), \sigma^2] \tag{1}$$

which represents the probability of obtaining the observation y_i given that our model predicts the mean of a distribution $g(\theta, x_i)$ with variance σ^2. Assume we have count data. What distribution would be a logical choice to model these data? Write out a model for the data.

Answer

The Poisson distribution is the classic distribution for (integer) counts; e.g., things in a plot, things that pass a point, etc.). The Poisson distribution has 1 parameter: λ = mean = variance; that is, the first and second central moments are equal.

$$y_i \sim \text{Poisson}(g(\theta, x_i)) \quad ()$$

```
pois <- rpois(n = 100000, lambda = 50)
pois[1:10]
```

```
##  [1] 50 48 40 45 52 41 48 45 57 55
```

Question 2

We commonly represent the following general framework for linking models to data:

- The mass of carbon in above ground biomass in square m plot.
- The number of seals on a haul-out beach in the gulf of AK.
- Presence or absence of an invasive species in forest patches.
- The probability that a white male will vote republican in a presidential election.
- The number of individuals in four mutually exclusive income categories.
- The number of diseased individuals in a sample of 100.
- The political party affiliation (democrat, republican, independent) of a voter.

Answer

Table 1: Question 2 Answers

Random Variable	Distribution	Support
The mass of carbon in above ground biomass in square m plot.	lognormal or gamma	non-negative real numbers
The number of seals on a haul-out beach in the gulf of AK.	Poisson or negative binomial	counts
Presence or absence of an invasive species in forest patches.	Bernoulli	0 or 1
The probability that a white male will vote republican in a presidential election.	beta	0 to 1 real numbers
The number of individuals in four mutually exclusive income categories.	multinomial	counts in >2 categories
The number of diseased individuals in a sample of 100.	binomial	counts in 2 categories
The political party affiliation (democrat, republican, independent) of a voter.	multinomial	counts in >2 categories

Question 3

Find the mean, variance, and 95% quantiles of 10000 random draws from a Poisson distribution with $\lambda = 33$.

```r
pois <- rpois(n = 10000, lambda = 33)
pois[1:10]
```

```
##  [1] 35 31 38 36 30 22 47 29 23 31
```

Answer

The mean of this example Poisson distribution is: **33.0**

The variance of this example Poisson distribution is: **33.4**

The 95% quantile of this example Poisson distribution is: **22.0, 45.0**

Question 4

Simulate `one` observation of survey data with five categories on a Likert scale, i.e. strongly disagree to strongly agree. Assume a sample of `80` respondents and the following probabilities:

a) Strongly disagree = 0.07
b) Disagree = .13
c) Neither agree nor disagree = .15
d) Agree = .23
e) Strongly agree = .42

```r
p <- c(0.07, 0.13, 0.15, 0.23, 0.42)
N <- 80
n <- 1
rmultinom(n = n, size = N, prob = p)
```

```
##      [,1]
## [1,]    5
## [2,]   11
## [3,]   10
## [4,]   22
## [5,]   32
```

Question 5

The average above ground biomass in a grazing allotment of sagebrush grassland is `103 g/m^2`, with a standard deviation of `23`. You clip a `1 m^2` plot. Write out the model for the probability density of the data point. What is the probability density of an observation of `94` assuming the data are normally distributed? Is there a problem using normal distribution? What is the probability that your plot will contain between `90 gm` and `110 gm` of biomass? See this question from Lab 4 assuming the data are lognormally and gamma distributed.

Answer

$$y \sim \mathsf{normal}(103, 23^2) \qquad ()$$

```r
x <- 94
mean <- 103
sd <- 23
d_norm <- dnorm(x = x, mean = mean, sd = sd)
scales::percent(d_norm, accuracy = .1)
```

```
## [1] "1.6%"
```

The probability density of an observation of 94 assuming the data are normally distributed is: **1.6%**

Is there a problem using normal distribution?

The normal distrubtion supports random variables from $-\infty$ to $+\infty$ and it is not possible to have negative biomass in a sample area.

```r
samp_low <- 90
samp_high <- 110
p_norm <- pnorm(q = c(samp_low, samp_high), mean = mean, sd = sd)

scales::percent(p_norm[2] - p_norm[1], accuracy = .1)
```

```
## [1] "33.4%"
```

The probability that the plot will contain between 90 gm and 110 gm of biomass is: **33.4%**

Question 6

The prevalence of a disease in a population is the proportion of the population that is infected with the disease. The prevalence of chronic wasting disease in male mule deer on winter range near Fort Collins, CO is `12 percent`. A sample of `24` male deer included `4` infected individuals. Write out a model that represents how the data arise. What is the probability of obtaining these data conditional on the given prevalence (`p=0.12`)? (For lab report)

Answer

$$y \sim \text{binomial}(24, 0.12) \qquad ()$$

```r
x <- 4
n <- 24
p <- 0.12
d_binom <- dbinom(x = x, size = n, prob = p)
```

The probability of obtaining these data conditional on the given prevalence (p=0.12) is: **17.1%**

Question 7

Researchers know that the true proportion of related age-sex classifications for elk in Rocky Mountain National Park are: Adult females (`p = 0.56`), Yearling males (`p = 0.06`), Bulls (`p = 0.16`), and Calves (`p = 0.22`). What is the probability of obtaining the classification data conditional on the known sex-age population proportions given the following counts?

a) Adult females (count = 65)
b) Yearling males (count = 4)
c) Bulls (count = 25)
d) Calves (count = 26)

Answer

```
p <- c(0.56, 0.06, 0.16, 0.22)
n <- c(65, 4, 25, 26)

d_multinom <- dmultinom(x = n, prob = p)
```

The probability of obtaining the classification data conditional on the known sex-age population proportions given the following counts is: **0.030%**

Question 8

Nitrogen fixation by free-living bacteria occurs at a rate of `1.9 g/N/ha/yr` with a standard deviation σ of `1.4`. What is the lowest fixation rate that exceeds `2.5 percent` of the distribution? Use a normal distribution for this problem, but discuss why this might not be a good choice.

Answer

```
p <- 0.025
mean <- 1.9
sd <- 1.4

q_norm <- qnorm(p = p, mean = mean, sd = sd)
```

The lowest fixation rate that exceeds 2.5% of the distribution is: **-84.4%**

Is there a problem using normal distribution?

The normal distrubtion supports random variables from $-\infty$ to $+\infty$ and it is not possible to have negative nitrogen fixation.

Probability Lab 3 - Marginal Distributions

Question 1

Fill in Table 2 to estimate the marginal probabilities of presence and absence of the two species. The cells show the joint probability of the events specified in the row and column. The right column and the bottom row show the marginal probabilities.

Table 2: Estimates of marginal probabilities for island occupancy

Events	S	S^c	Marginal
R	$\Pr(S, R) = \frac{2}{32}$	$\Pr(S^c, R) = \frac{9}{32}$	$\Pr(R) = \frac{11}{32}$
R^c	$\Pr(S, R^c) = \frac{18}{32}$	$\Pr(S^c, R^c) = \frac{3}{32}$	$\Pr(R^c) = \frac{21}{32}$
Marginal	$\Pr(S) = \frac{20}{32}$	$Pr(S^c) = \frac{12}{32}$	$\sum = \frac{32}{32}$

What is the sum of the marginal rows?

The sum of the marginal rows is 1

What is the sum of the marginal columns?

The sum of the marginal columns is 1

Why? Note, when we marginalize over R we are effectively eliminating S and vice versa.

When we marginalize over R we are effecitvely eliminating S because we are interested in making inference on a single parameter, R.

Question 2

Use the data in Table 1 and the probabilities in Table 2 to illustrate the rule for the union of two events, the probability that an island contains either species, $\Pr(R \cup S)$. You will need to derive the formula for the probability of event A or B to solve this problem. A Venn diagram might help you do so.

$$\Pr(R \cup S) = \Pr(R) + \Pr(S) - \Pr(S, R)$$

where:

- $\Pr(R) = \frac{11}{32}$
- $\Pr(S) = \frac{20}{32}$
- $\Pr(S, R) = \frac{2}{32}$

$$\Pr(R \cup S) = \frac{11}{32} + \frac{20}{32} - \frac{2}{32} = \frac{29}{32} = 90.6\%$$

Question 3

Use the marginal probabilities in Table 2 to calculate the probability that an island contains both species i.e., $\Pr(R, S)$, assuming that R and S are independent. Compare the results from those calculations with the probability that both species occur on an island calculated directly from the data in Table 1. Interpret the results ecologically. What is $\Pr(R \mid S)$ and $\Pr(S \mid R)$?

Assuming that R and S are independent

Calculate the probability that an island contains both species: $\Pr(R, S)$

$$\Pr(R, S) = \Pr(R) \cdot \Pr(S)$$

where:

- $\Pr(R) = \frac{11}{32}$
- $\Pr(S) = \frac{20}{32}$

$$\Pr(R, S) = \frac{11}{32} \cdot \frac{20}{32} = 0.215 = 21.5\%$$

What is $\Pr(R \mid S)$ and $\Pr(S \mid R)$?

Assuming that R and S are independent:

$$\Pr(R \mid S) = \Pr(R) = \frac{11}{32}$$
$$\Pr(S \mid R) = \Pr(S) = \frac{20}{32}$$

Assuming that R and S are not independent

Calculate the probability that an island contains both species: $\Pr(R, S)$

$$\Pr(R, S) = \frac{2}{32} = 0.0625 = 6.25\%$$

What is $\Pr(R \mid S)$ and $\Pr(S \mid R)$?

Assuming that R and S are not independent:

$$\Pr(R \mid S) = \frac{\Pr(R, S)}{\Pr(S)} = \frac{\frac{2}{32}}{\frac{20}{32}} = 0.1 = 10\%$$

$$\Pr(S \mid R) = \frac{\Pr(S, R)}{\Pr(R)} = \frac{\frac{2}{32}}{\frac{11}{32}} = 0.182 = 18.2\%$$

Interpretation Interpret the results ecologically

Assuming that R and S are not independent means that we believe that the presenence of one species influences the presence of the other species and vice versa. The probability that an island contains both species, $\Pr(R, S) = 6.25\%$, means that there is a low chance of both species occupying the same site. Based on this low probability of co-habitation, it is likely that these two species compete with eachother for resources resulting in competitve exclusion.

Continuous random variables

We now explore marginal distributions for continuous random variables. This requires introducing a new distribution, the multivariate normal:

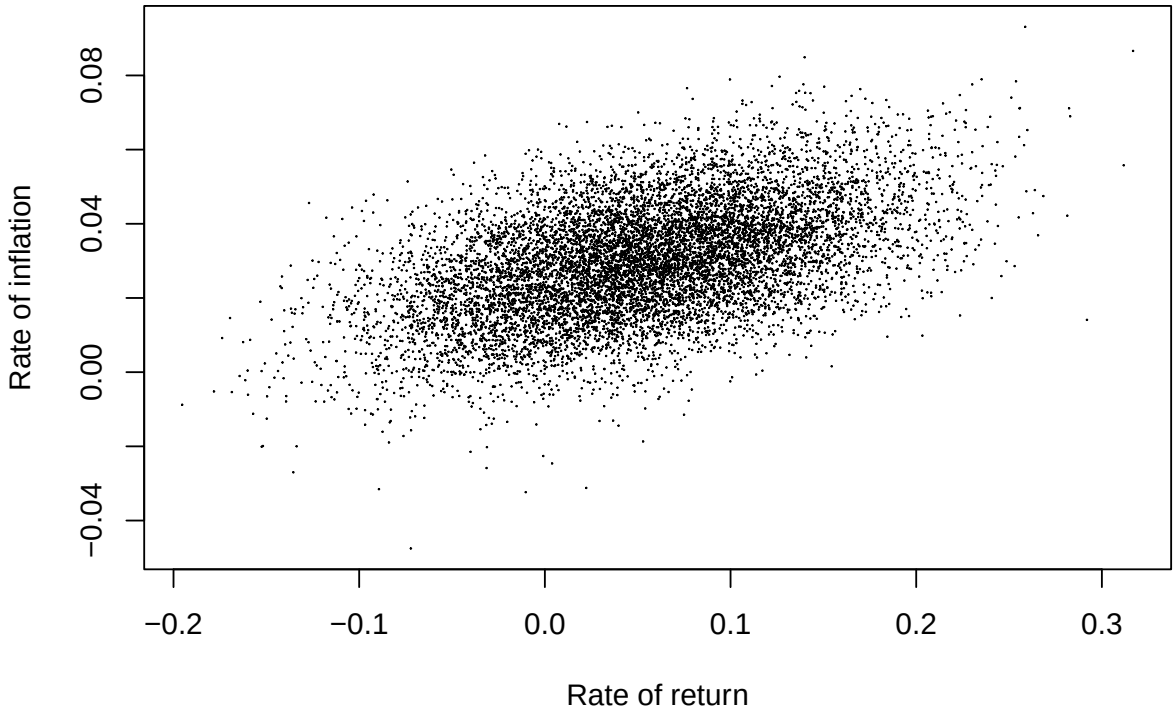
$$\mathbf{z} \sim \text{mulitvariate normal}(\mu, \Sigma)$$

where $\mathbf{z}_i$ is a vector of random variables, μ is a vector of means (which can be the output of a deterministic model) and $Sigma$ is a variance covariance matrix. The diagonal of Σ contains the variances and the off diagonal contains the covariance of $\Sigma[i, j]$. The covariance can be calculated as $\sigma_i \sigma_j \rho$ where σ_i is the standard deviation of the ith random variable, σ_j is the standard deviation of the jth random variable, and ρ is the correlation between the random variable i and j. The covariance matrix is square and symmetric. We will learn more about these matrices later in the course. For now, an example will go a long way toward helping you understand the multivariate normal distribution.

The rate of inflation and the rate of return on investments are know to be positively correlated. Assume that the mean rate of inflation is `.03` with a standard deviation of `0.015`. Assume that the mean rate of return is `0.0531` with a standard deviation of `0.0746`. Assume the correlation between inflation and rate of return is `0.5`.

You can simulate interest rate and inflation data reflecting their correlation using the following function:

```r
DrawRates = function(n, int,int.sd, inf, inf.sd, rho.rates) {
  covar = rho.rates * int.sd * inf.sd
  Sigma <- matrix(c(int.sd^2, covar, covar, inf.sd^2), 2, 2)
  mu = c(int,inf)
  x = (MASS::mvrnorm(n = n, mu = mu, Sigma))
  return(x)
}

mu.int = .0531
sd.int = .07 #.0746
mu.inf = .03
sd.inf = .015 #.015
rho=.5
n = 10000

x = DrawRates(n = n, int = mu.int, int.sd = sd.int, inf = mu.inf, inf.sd = sd.inf, rho.rates = rho)
par(mfrow=c(1,1))
plot(x[, 1], x[, 2], pch = 19, cex = .05, xlab = "Rate of return", ylab = "Rate of inflation")
```

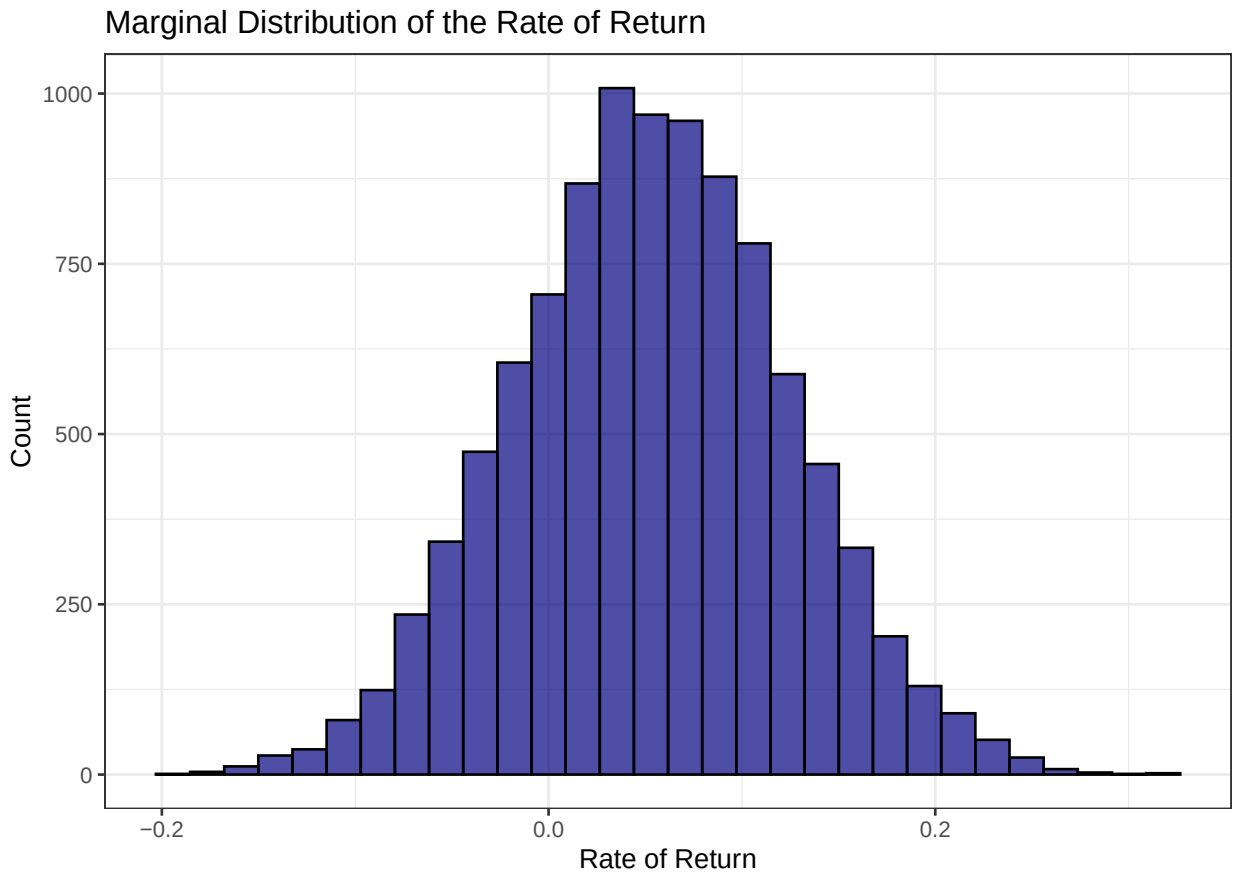


What would this cloud look like if the rates were not correlated?

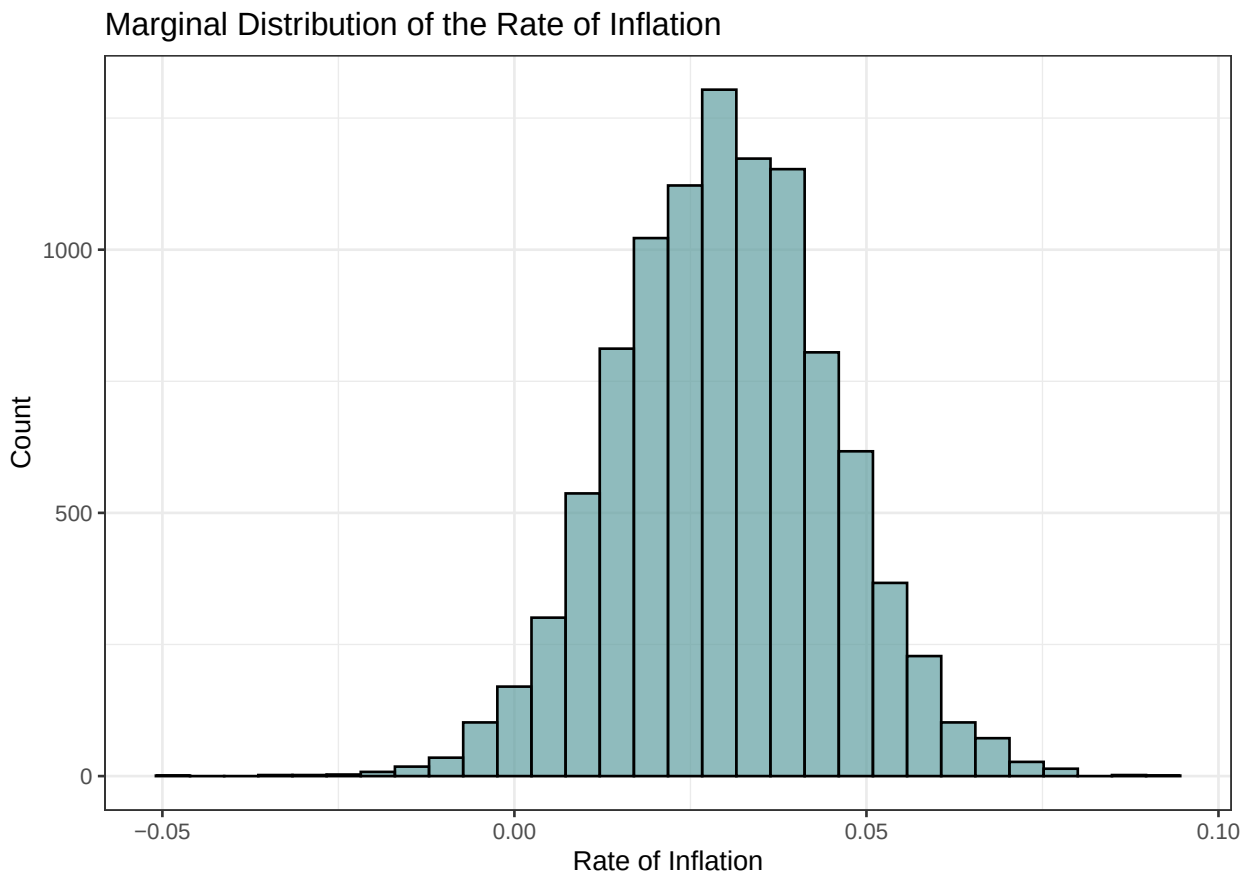
If the rates were not correlated the data would be dispersed randomly in the plot.

Show an approximate plot of the marginal distribution of each random variable. It turns out this is the way we will do it with MCMC.

```r
## create data frame
dta <- data.frame(
  return_rate = x[,1]
  , inflation_rate = x[,2]
)
## return rate distribution
ggplot(data = dta) +
  geom_histogram(aes(return_rate), fill = "navy", lwd = 0.5, color = "black", alpha = 0.7) +
  xlab("Rate of Return") +
  ylab("Count") +
  labs(title = "Marginal Distribution of the Rate of Return") +
  theme_bw()
```

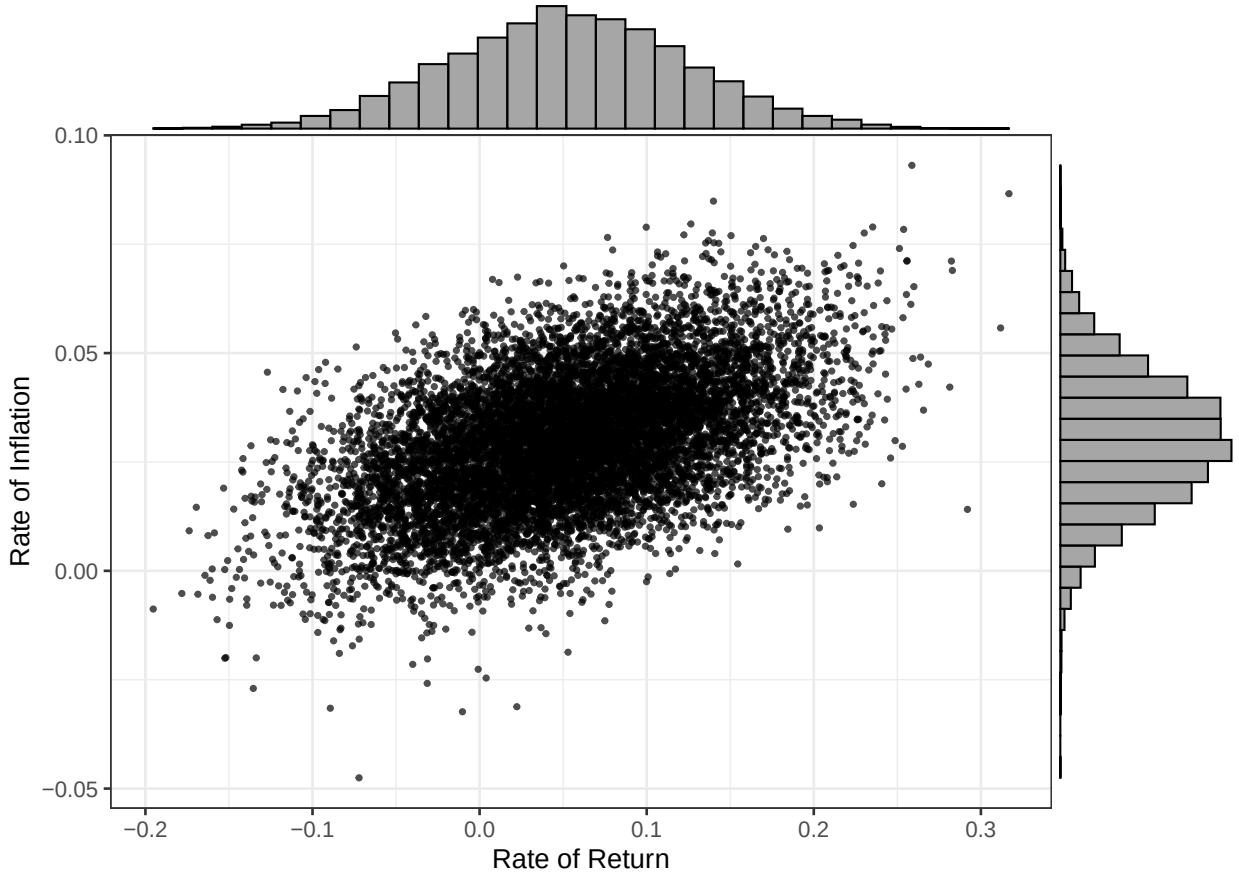



```r
## inflation rate distribution
ggplot(data = dta) +
  geom_histogram(aes(inflation_rate), fill = "cadetblue", lwd = 0.5, color = "black", alpha = 0.7) +
  xlab("Rate of Inflation") +
  ylab("Count") +
  labs(title = "Marginal Distribution of the Rate of Inflation") +
  theme_bw()
```



Can you combine the histogram with the scatter plot to more clearly reveal the marginalization? Take a look at the `ggExtra` package and `ggMarginal` function.

```
p1 <- ggplot(dta, aes(x = return_rate, y = inflation_rate)) +
  geom_point(alpha = 0.7, size = 0.7) +
  xlab("Rate of Return") +
  ylab("Rate of Inflation") +
  theme_bw()

p1 %>%
  ggExtra::ggMarginal(type = "histogram", fill = "gray50", alpha = 0.7, lwd = 0.4)
```



Probability Lab 4 - Moment Matching

Introduction

When we say *support*, we are referring to the values of a random variable for which probability density or probability exceed 0 and are defined. The support of lognormal distribution is continuous and strictly non-negative, which makes it particularly useful in ecology. Moreover, it is often useful because it is asymmetric, allowing for values that are extreme in the positive direction. Finally, it is useful for representing products of random variables. The central limit theorem would predict that the distribution of sums of random variables will be normal, no matter how each is individually distributed. The products of random variables will be lognormally distributed regardless of their individual distributions.

If a random variable is lognormally distributed then the log of that random variable is normally distributed (conversely, if you exponentiate a normal random variable it generates a lognormal random variable). The first parameter of the lognormal distribution is the mean of the random variable on the log scale (i.e., α on cheat sheet, `meanlog` in R) and the second parameter is the variance (or sometimes the standard deviation) of the random variable on the log scale (i.e., β on cheatsheet, `sdlog` in R). We often predict the median of the distribution with our deterministic model, which means that we can use the log of the median as the first parameter because

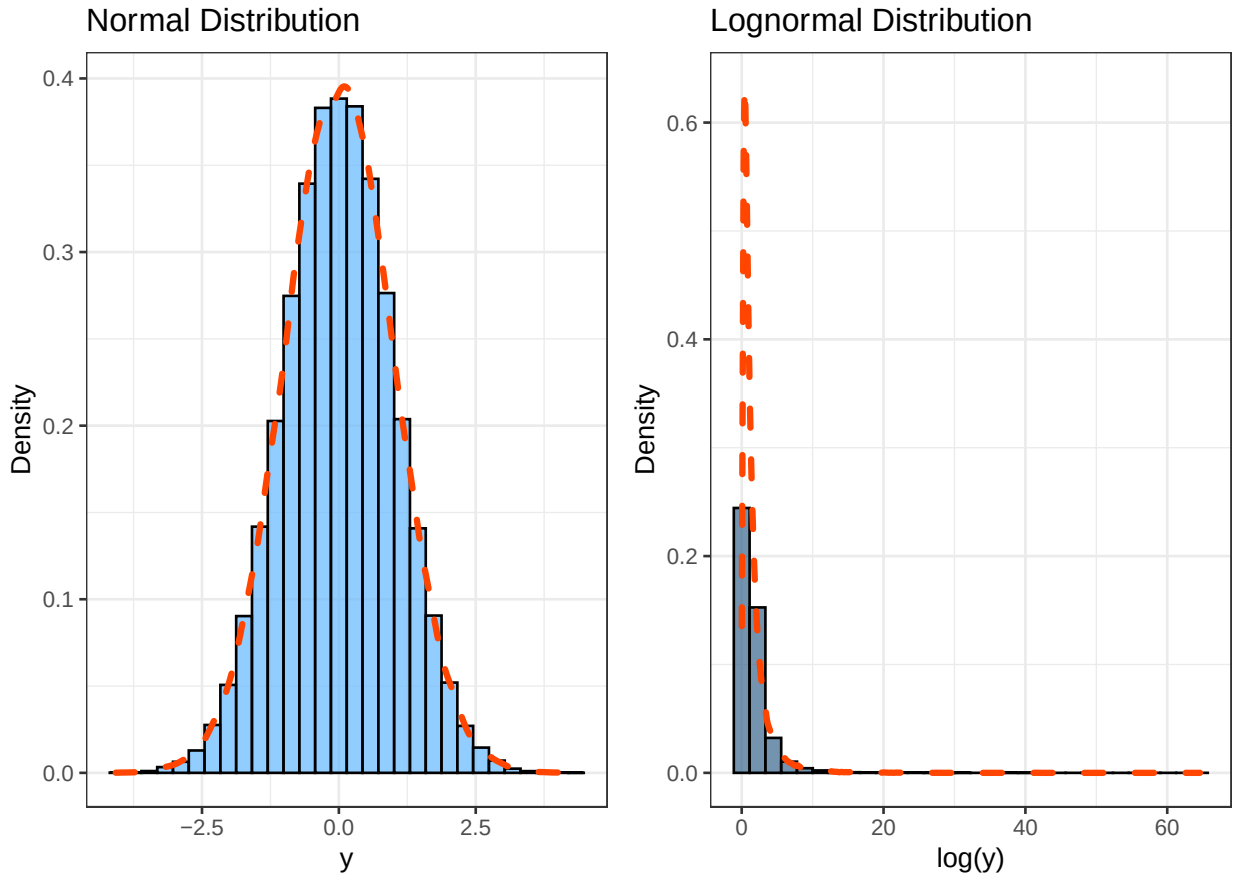
$$\begin{aligned} z &\sim \mathsf{lognormal}(\alpha, \beta) \\ \mathsf{median}(\mathrm{z}) &= \mathrm{e}^{\alpha} \\ \mathsf{log}(\mathsf{median}(\mathsf{z})) &= \alpha \end{aligned}$$

Question 1

Simulate 10,000 data points from a normal distribution with mean 0 and standard deviation 1 and another 10,000 data points from a log normal distribution with first parameter (the mean of the random variable on the log scale) = 0 and second parameter (the standard deviation of the parameter on the log scale) = 1. Display side-by-side histograms scaled to the density. Find the mean and variance of the lognormal distribution using moment matching. Check your moment-matched values empirically with the simulated data. The moment-matched values and the empirical values are close for the mean, but less so for the variance. Why? What happens when you increase the number or draws? Explore the two distributions by repeating with different means and standard deviations of your choice.

```
mean <- 0
sd <- 1
n <- 100000
# normal
dist_norm <- rnorm(n = n, mean = mean, sd = sd)

# log-normal
dist_lognorm <- rlnorm(n = n, meanlog = mean, sdlog = sd)

# plot historgram norm
  plt_norm <-
    ggplot(data = data.frame(dist_norm), aes(x = dist_norm)) +
    geom_histogram(
      aes(y = ..density..)
      , fill = "steelblue1"
      , alpha = 0.7
      , color = "black"
      , lwd = 0.5
    ) +
    geom_density(
      linetype = 2
      , lwd = 1.2
      , color = "orangered"
    ) +
    xlab("y") +
    ylab("Density") +
    labs(
      title = "Normal Distribution"
    ) +
    theme_bw()

# plot historgram log-norm
  plt_lognorm <-
    ggplot(data = data.frame(dist_lognorm), aes(x = dist_lognorm)) +
    geom_histogram(
      aes(y = ..density..)
      , fill = "steelblue4"
      , alpha = 0.7
      , color = "black"
      , lwd = 0.5
    ) +
    geom_density(
      linetype = 2
```

```
      , lwd = 1.2
      , color = "orangered"
    ) +
    xlab("log(y)") +
    ylab("Density") +
    labs(
      title = "Lognormal Distribution"
    ) +
    theme_bw()

# combine charts
  ggpubr::ggarrange(
      plt_norm, plt_lognorm
      , nrow = 1
      , ncol = 2
    )
```




```
# empirically simulate data
mean_lognorm <- exp(mean + sd^2/2)
var_lognorm <- (exp(sd^2) - 1) * exp(2*mean + sd^2)
```

Answers

Find the mean and variance of the lognormal distribution using moment matching. Check your moment-matched values empirically with the simulated data.

The mean of the lognormal distribution is: 1.655

The variance of the lognormal distribution is: 4.734

The empirical mean of the lognormal distribution ($\mu = e^{\alpha+\frac{\beta^2}{2}}$) is: 1.649

The empirical variance of the lognormal distribution ($\sigma^2 = (e^{\beta^2}-1)\cdot e^{2\alpha+\beta^2}$) is: 4.671

The moment-matched values and the empirical values are close for the mean, but less so for the variance. Why? What happens when you increase the number or draws? Explore the two distributions by repeating with different means and standard deviations of your choice.

The moment-matched values and the empirical values for the mean and variance are close but less so for the variance than the mean. This occurs because the variance of the lognormal distribution is a product of two exponential functions. When we simulate the data using rlnorm in R, increasing the number of simulated values results in the moment-matched values approaching the emprical values as n approaches $+\infty$.

Question 2

The average above ground biomass in a grazing allotment of sagebrush grassland is `103.4 g/m^2^`, with a standard deviation (σ) of `23`. You clip a `1 m^2^` plot. You assumed a normal distribution for this problem in Probability Lab #2. Now redo the problem with a more appropriate distribution. Write out the model for the probability density of the data point. What is the probability density of an observation of `94` assuming the data are gamma distributed? What is the probability that your plot will contain between `90 and 110 gm` of biomass? Now assume the data are lognormally distributed and redo this problem. Compare to the results where you assumed a gamma distribution.

Gamma distribution

$$y_i \sim \text{gamma}\left(\frac{\mu^2}{\sigma^2}, \frac{\mu}{\sigma^2}\right)$$

$$y_i \sim \text{gamma}\left(\frac{103.4^2}{23^2}, \frac{103.4}{23^2}\right)$$

Lognormal distribution

$$y_i \sim \text{lognormal}\left(\log\left(\frac{\mu}{\sqrt{\frac{\sigma^2}{\mu^2}+1}}\right), \sqrt{\log\left(\frac{\sigma^2}{\mu^2}+1\right)}\right)$$

$$y_i \sim \text{lognormal}\left(\log\left(\frac{103.4}{\sqrt{\frac{23^2}{103.4^2}+1}}\right), \sqrt{\log\left(\frac{23^2}{103.4^2}+1\right)}\right)$$

```
x <- 94
mu <- 103.4
sd <- 23
########################
## assume gamma dist
########################
```

```r
## calculate the shape alpha for gamma dist
alpha <- (mu^2)/(sd^2)
## calculate the rate beta for gamma dist
beta <- (mu)/(sd^2)
## probability density f'n
d_gamma <- dgamma(x = x, shape = alpha, rate = beta)
## cumulative density f'n
x_low <- 90
x_high <- 110
p_gamma <- pgamma(q = c(x_low, x_high), shape = alpha, rate = beta)
p_gamma_cdf <- p_gamma[2] - p_gamma[1]
#########################
## assume lognormal dist
#########################
## meanlog
alpha = log( mu / sqrt((sd^2/mu^2)+1) )
## sdlog
beta = sqrt( log( (sd^2/mu^2)+1 ) )
## probability density f'n
d_lnorm <- dlnorm(x = x, meanlog = alpha, sdlog = beta)
## cumulative density f'n
p_lnorm <- plnorm(q = c(x_low, x_high), meanlog = alpha, sdlog = beta)
p_lnorm_cdf <- p_lnorm[2] - p_lnorm[1]
```

Answers

What is the probability density of an observation of `94`?

The probability density of an observation of '94' assuming the data are gamma distributed: 0.0174

The probability density of an observation of '94' assuming the data are lognormally distributed: 0.0183

What is the probability that your plot will contain between `90 and 110 gm` of biomass?

The probability that your plot will contain between '90 and 110 gm' of biomass assuming the data are gamma distributed: 0.3423

The probability that your plot will contain between '90 and 110 gm' of biomass assuming the data are lognormally distributed: 0.3513

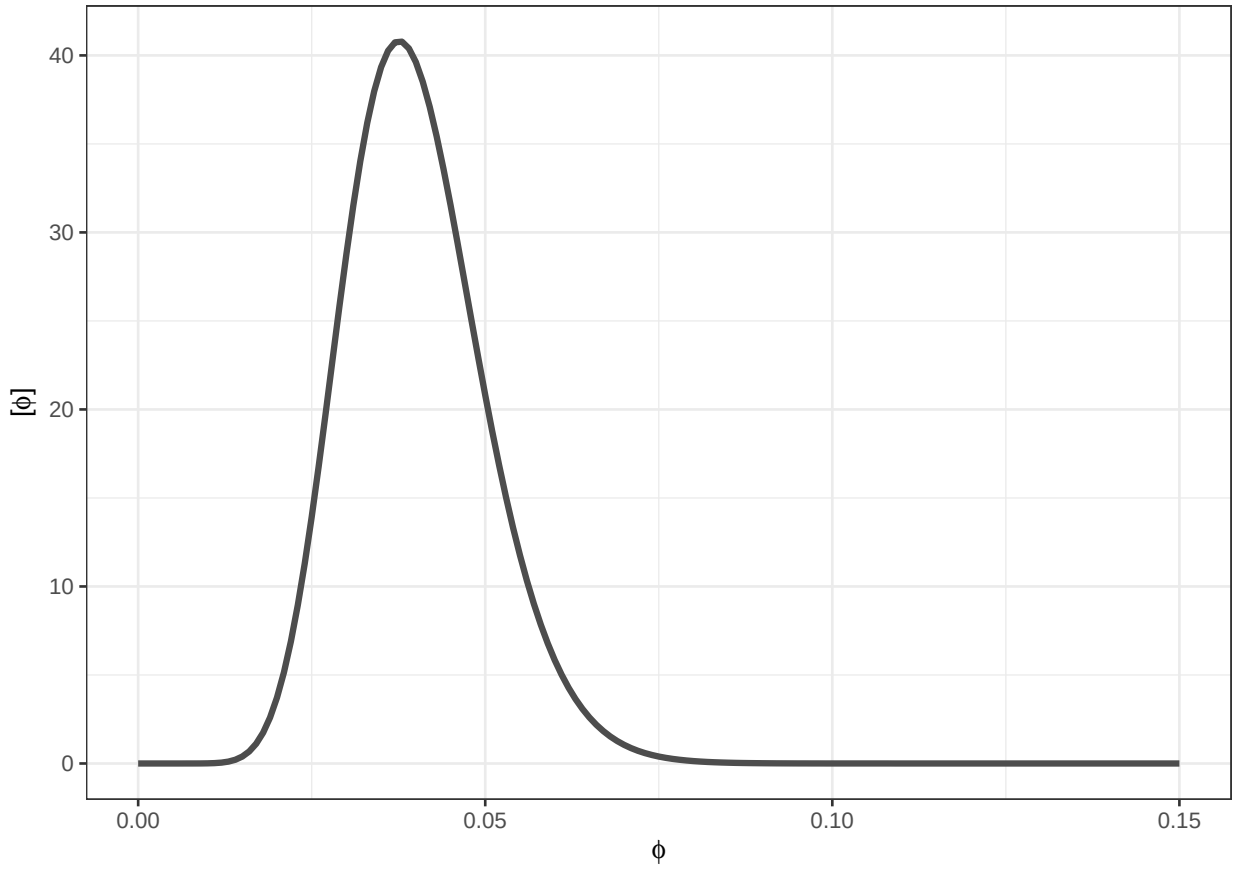
Question 3

We are interested in the proportion (ϕ) of Maryland counties that contain a coal fired power plant. Existing literature shows that that this proportion has a mean of $\mu = 0.04$ with a standard deviation of $\sigma = 0.01$. Write out a model for the distribution of ϕ, conditional on μ and σ. The challenge here is to use moment matching for a random variable with support between 0-1. Plot the probability distribution of ϕ.

$$\phi \sim \text{beta}\left(\frac{\left(\mu^2 - \mu^3 - \mu \cdot \sigma^2\right)}{\sigma^2}, \frac{\left(\mu - 2\mu^2 + \mu^3 - \sigma^2 + \mu \cdot \sigma^2\right)}{\sigma^2}\right)$$

$$\phi \sim \text{beta}\left(\frac{\left(0.04^2 - 0.04^3 - 0.04 \cdot 0.01^2\right)}{0.01^2}, \frac{\left(0.04 - 2 \cdot 0.04^2 + 0.04^3 - 0.01^2 + 0.04 \cdot 0.01^2\right)}{0.01^2}\right)$$

```r
mu <- 0.04
sigma <- 0.01
## shape1
alpha <- (mu^2 - mu^3 - mu * sigma^2) / sigma^2
## shape2
beta <- (mu - 2*mu^2 + mu^3 - sigma^2 +  mu * sigma^2) / sigma^2
## x
x <- seq(0, .15, .001)
## pbeta
d_beta <- dbeta(x = x, shape1 = alpha, shape2 = beta)
## data
dta <- data.frame(
  x = x
  , y = d_beta
)
## plot
ggplot(data = dta, mapping = aes(x = x, y = y)) +
  geom_line(
    color = "gray30"
    , lwd = 1.2
  ) +
  xlab(latex2exp::TeX("$\\phi$")) +
  ylab(latex2exp::TeX("\\[$\\phi$\\]")) +
  theme_bw()
```



Question 4

If you visited 50 counties, what is the probability that 5 would contain a plant, conditional on the hypothesis that $\phi = 0.04$?

$$\begin{aligned} \Pr(y = 5 \mid \phi, n = 50) = \\ \text{binomial}\big(\text{y} = 5 \mid \phi = 0.04, \text{n} = 50\big) = \\ \binom{50}{5} \cdot 0.04^5 \cdot (1 - 0.04)^{50-5} \end{aligned}$$

```
x <- 5
phi <- 0.04
n <- 50
d_binom <- dbinom(x = x, p = phi, size = n)
```

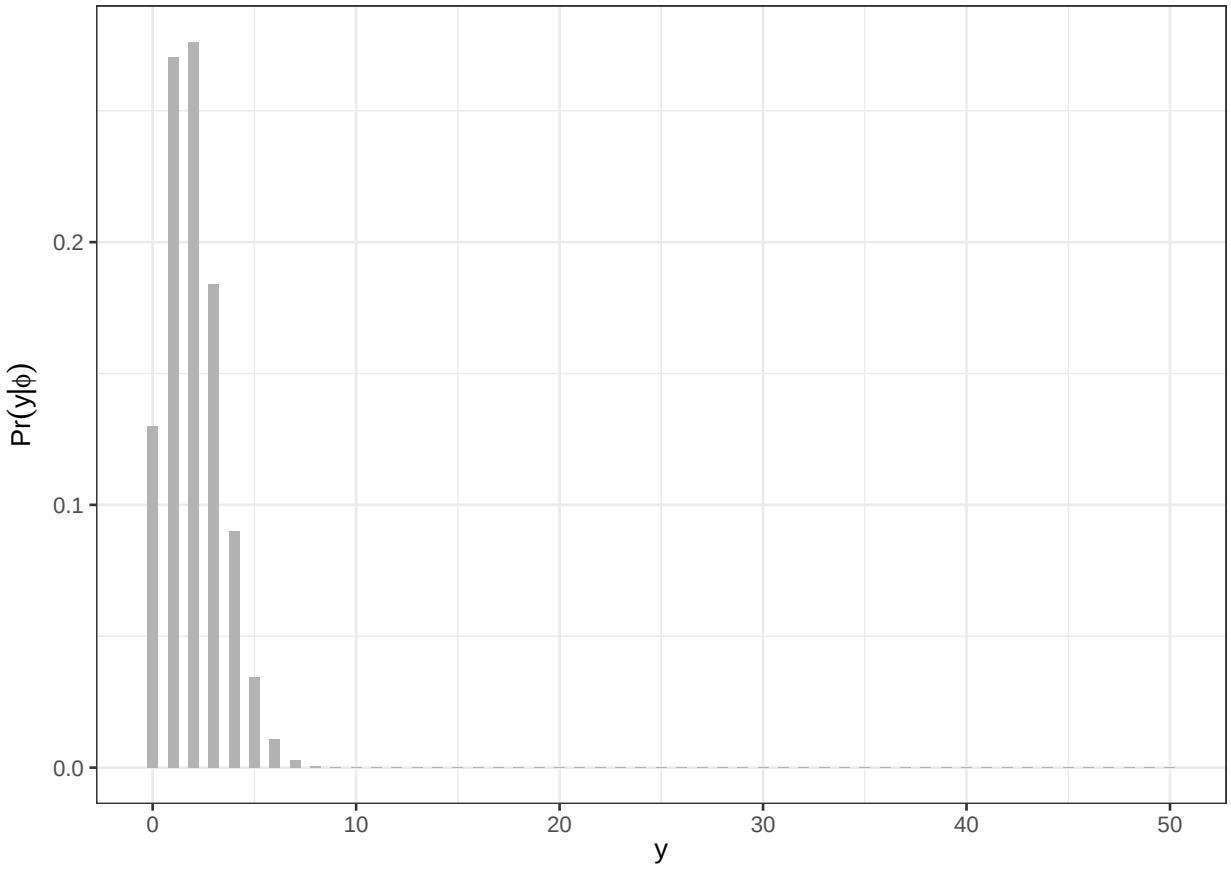
Answer

If you visited 50 counties the probability that 5 would contain a plant, conditional on the hypothesis that $\phi = 0.04$ and assuming the data are binomially distributed: 0.0346

Question 5

Plot the probability of the data for $y = 1...50$ counties with coal plants out of 50 counties visited conditional on the hypothesis $\phi = 0.04$.

```
x <- seq(0, 50)
phi <- 0.04
n <- 50
## PMF
y <- dbinom(x = x, p = phi, size = n)
## plot
ggplot(data = data.frame(x, y), mapping = aes(x = x, y = y)) +
  geom_col(
    fill = "gray70"
    , width = 0.5
  ) +
  xlab(latex2exp::TeX("$y$")) +
  ylab(latex2exp::TeX("$Pr(y | \\phi)$")) +
  scale_x_continuous(labels = scales::label_comma(accuracy = 1) ) +
  theme_bw()
```



Question 6

What is the probability that at least 5 counties have a coal plant, conditional on the hypothesis that $\phi = 0.04$?

$$\begin{aligned} \Pr\left(y \geq 5 \mid \phi, n = 50\right) = \\ \text{binomial}\left(\text{y} \geq 5 \mid \phi = 0.04, \text{n} = 50\right) = \\ \sum_{y_i \in (5,6,\ldots,50)} \binom{50}{y_i} (0.04)^{y_i} (1 - 0.04)^{50 - y_i} \end{aligned}$$

```
q <- 4
p <- 0.04
n <- 50
## CDF
p_binom_inv = 1 - pbinom(q = q, p = p, size = n)
```

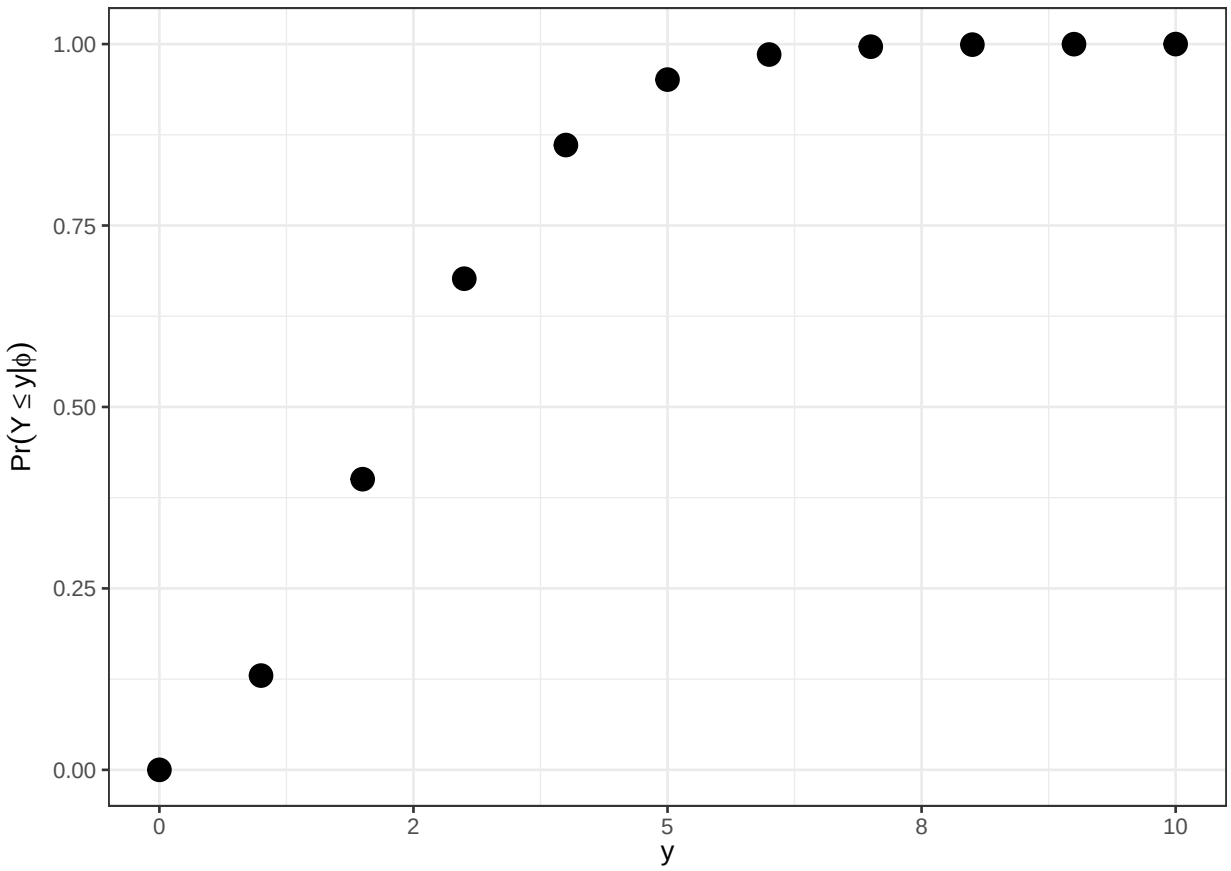
Answer

The probability that at least 5 counties have a coal plant, conditional on the hypothesis that $\phi = 0.04$ and assuming the data are binomially distributed: 0.049

Question 7

Plot the probability that fewer than y counties contain plants where y takes on values between 1 and 10. Condition of the probability of occupancy $\phi = 0.04$?

```
x <- seq(0, 10)
phi <- 0.04
n <- 50
## CDF
y <- pbinom(q = x-1, p = phi, size = n)
## plot
ggplot(data = data.frame(x, y), mapping = aes(x = x, y = y)) +
  geom_point(
    fill = "gray15"
    , size = 4
  ) +
  xlab(latex2exp::TeX("$y$")) +
  ylab(latex2exp::TeX("$Pr(Y \\leq y | \\phi)$")) +
  scale_x_continuous(labels = scales::label_comma(accuracy = 1) ) +
  theme_bw()
```



Question 8

Simulate data for 75 counties (no coal plant = 0, coal plant = 1).

$$y \sim \text{binomial}(1, \phi) \equiv \text{y} \sim \text{Bernoulli}(\phi)$$

```
n <- 75
size <- 1
phi <- 0.04
## random generation
rand <- rbinom(n = n, size = size, prob = phi)
has_plant <- ifelse(rand == 1, "coal plant", "no coal plant")
table(has_plant)
```

```
## has_plant
##    coal plant no coal plant
##             1            74
```

```
if(length(rand == 1)>0){print(paste0("county ", which(rand %in% c(1)), " has a coal plant" ))}
```

```
## [1] "county 67 has a coal plant"
```

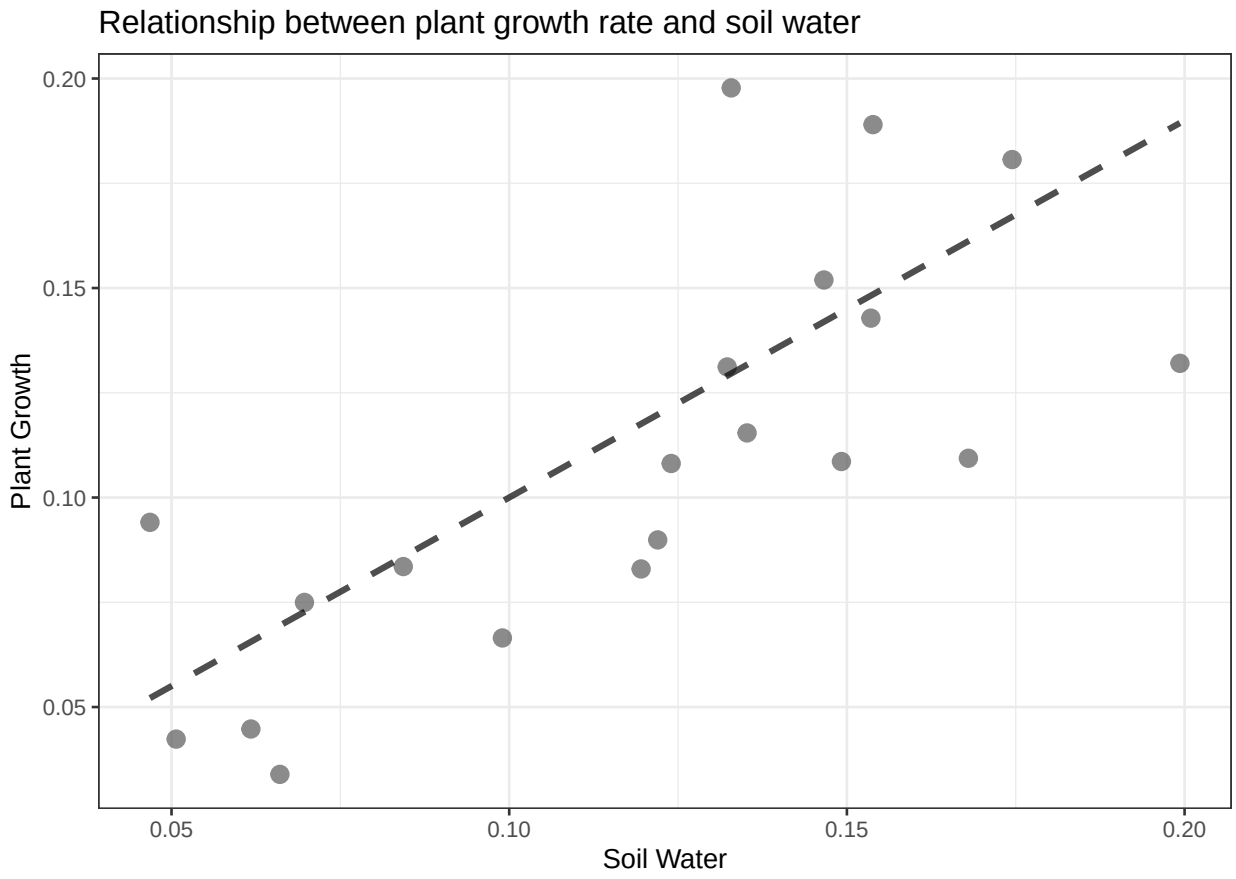
Question 9

You are modeling the relationship between plant growth rate and soil water. Represent plant growth (μ_i) as a linear function of soil water, $\mu_i = \beta_0 + \beta_1 x_i$. Write out the model for the data. Simulate a data set of 20, strictly non-negative pairs of y and x values. Plot the data and overlay the generating model. Assume that:

- Soil water, the x value, varies randomly and uniformly between 0.01 and 0.2
- $\beta_0 = 0.01$ and $\beta_1 = 0.9$
- the standard deviation of the model prediction is $\sigma = 0.03$

```
n <- 20
x <- runif(n = n, min = 0.01, max = 0.2)
b0 <- 0.01
b1 <- 0.9
sd <- 0.03
## calculate plant growth mu
mu <- b0 + b1 * x
## calculate the shape alpha for gamma dist
alpha <- (mu^2)/(sd^2)
## calculate the rate beta for gamma dist
beta <- (mu)/(sd^2)
## simulate y of strictly non-negative values using the gamma dist.
y <- rgamma(n = n, shape = alpha, rate = beta)
## data
dta <- data.frame(
  x = x
  , mu = mu
  , y = y
)
## plot
ggplot(data = dta) +
  geom_point(
```

```
  aes(x = x, y = y)
  , alpha = 0.7
  , size = 3
  , color = "gray35"
) +
geom_line(
  aes(x = x, y = mu)
  , linetype = 2
  , lwd = 1.2
  , alpha = 0.7
  , color = "black"
) +
xlab("Soil Water") +
ylab("Plant Growth") +
labs(
  title = "Relationship between plant growth rate and soil water"
) +
theme_bw()
```



Question 10

The negative binomial distribution is a more robust alternative to the Poisson distribution, allowing the variance to differ from the mean. There are two parameterizations for the negative binomial.

The first parameterization of the negative binomial distribution is more frequently used by ecologists:

$$[z \mid \lambda, r] = \frac{\Gamma(z+r)}{\Gamma(r)z!}\Big(\frac{r}{r+\lambda}\Big)^r\Big(\frac{\lambda}{r+\lambda}\Big)^z,$$

where z is a discrete random variable, λ is the mean of the distribution, and r is the *dispersion parameter*, also called the size. The variance of z is:

$$\sigma^2 = \lambda + \frac{\lambda^2}{r}$$

The second parameterization is more often implemented in coding environments (i.e. JAGS):

$$[z \mid r, \phi] = \frac{\Gamma(z+r)}{\Gamma(r)z!}\phi^r(1-\phi)^z,$$

where z is the discrete random variable representing the number of failures that occur in a sequence of Bernoulli trials before r successes are obtained. The parameter ϕ is the probability of success on a given trial. Where ϕ, the probability of success on a given trial is:

$$\phi = \frac{r}{(\lambda + r)}$$

Use the `rnbinom` function in R to simulate 100,000 observations from a negative binomial distribution with mean of μ=100 and variance of σ^2=400 using the **first** parameterization that has a mean and a dispersion parameter. (Hint: find an expression for r and moment match.) Do the same simulation using the **second** parameterization. Plot side-by-side histograms of the simulated data.

```
n <- 100000
mean <- 100
var <- 400
## the dispersion param
r <- mean^2/(var - mean)
## first parameterization
dist_negbinom1 <- rnbinom(n = n, mu = mean, size = r)

## phi = probability of success in given trial
phi <- r/(mean+r)
## second parameterization
dist_negbinom2 <- rnbinom(n, prob = phi ,size = r)
```

The mean of the first parameterization negative binomial distribution as simulated is: 99.866

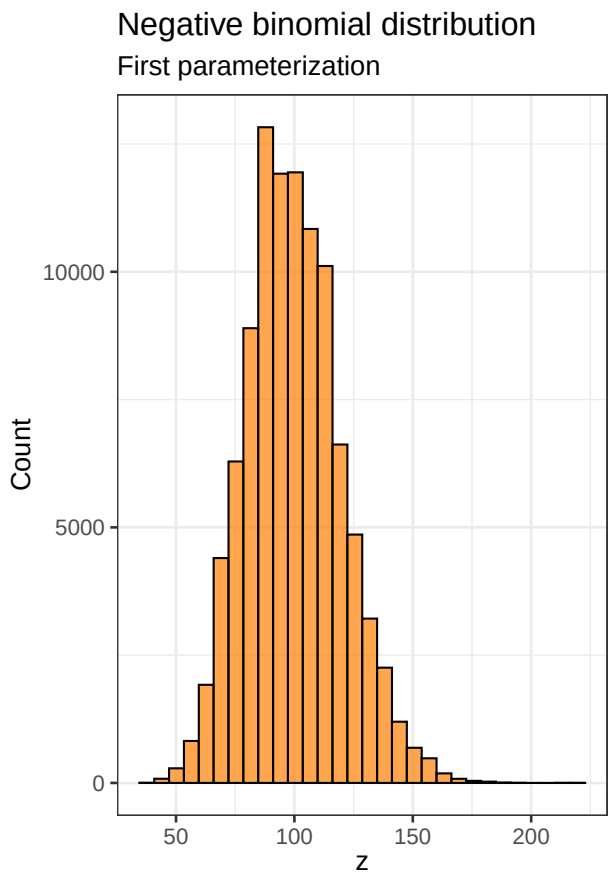
The variance of the first parameterization negative binomial distribution as simulated is: 397.698

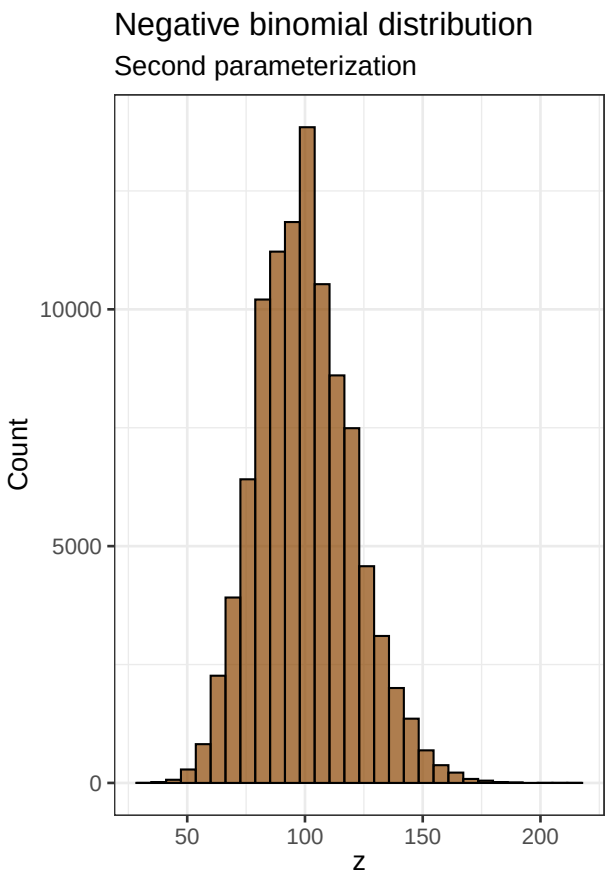
The mean of the second parameterization negative binomial distribution as simulated is: 100.099

The variance of the second parameterization negative binomial distribution as simulated is: 401.765

Plot side-by-side histograms of the simulated data.

```r
## plot historgram param 1
  p_dist_negbinom1 <-
    ggplot(data = data.frame(dist_negbinom1), aes(x = dist_negbinom1)) +
    geom_histogram(
      fill = "darkorange1"
      , alpha = 0.7
      , color = "black"
      , lwd = 0.4
    ) +
    xlab(latex2exp::TeX("$z$")) +
    ylab("Count") +
    labs(
      title = "Negative binomial distribution"
      , subtitle = "First parameterization"
    ) +
    theme_bw()
## plot historgram param 2
  p_dist_negbinom2 <-
    ggplot(data = data.frame(dist_negbinom2), aes(x = dist_negbinom2)) +
    geom_histogram(
      fill = "darkorange4"
      , alpha = 0.7
      , color = "black"
      , lwd = 0.4
    ) +
    xlab(latex2exp::TeX("$z$")) +
    ylab("Count") +
    labs(
      title = "Negative binomial distribution"
      , subtitle = "Second parameterization"
    ) +
    theme_bw()

## combine charts
  ggpubr::ggarrange(
      p_dist_negbinom1, p_dist_negbinom2
      , nrow = 1
      , ncol = 2
    )
```


Negative binomial distribution
First parameterization
Count
10000
5000
0
50
100
150
200
z


Negative binomial distribution
Second parameterization
Count
10000
5000
0
50
100
150
200
z