# ESS 575: Probability Lab 4 - Moment Matching

Team England

16 September, 2022

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

## 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., $\alpha$ 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., $\beta$ 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

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

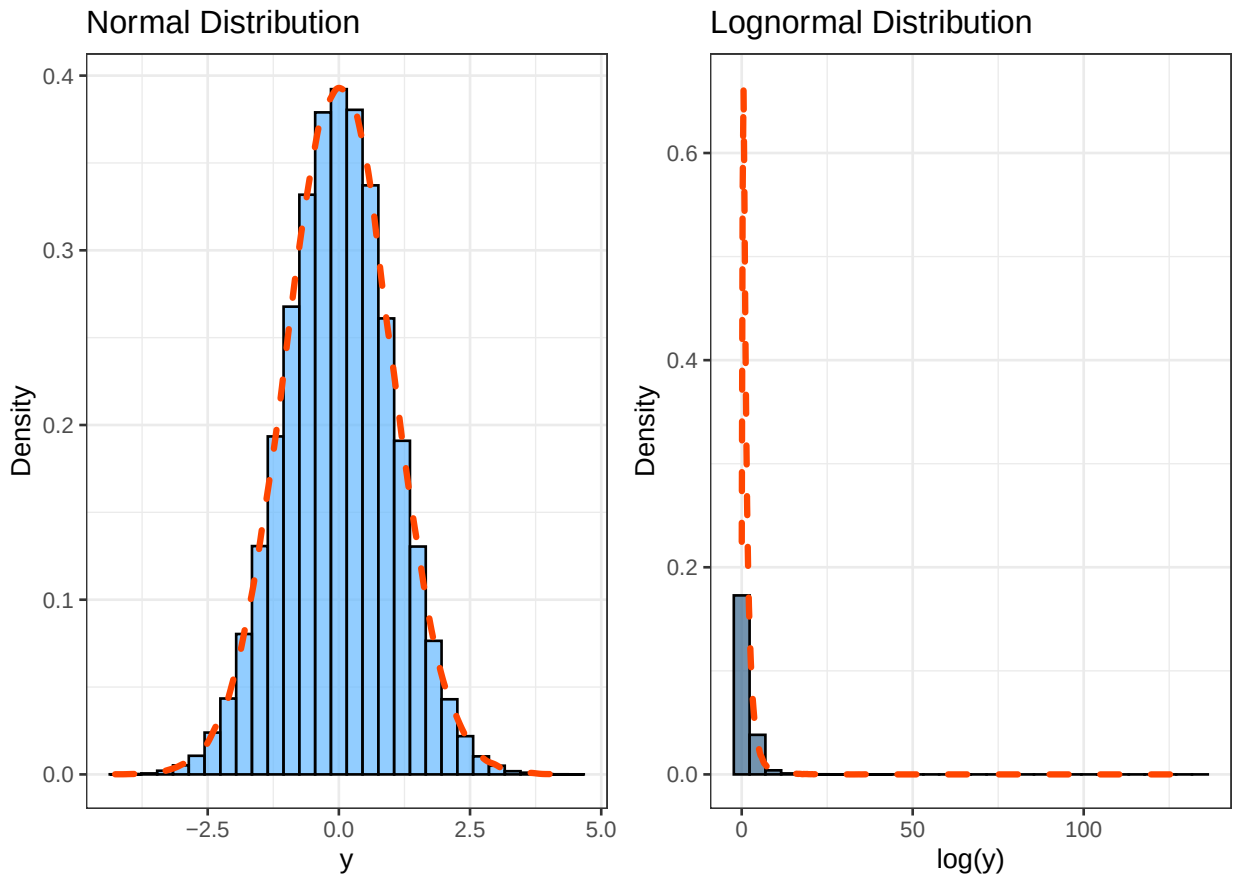
## Question 1

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: 1.637

The variance of the lognormal distribution is: 4.648

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 ($\sigma$) 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
#########################
# 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
```

```r
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

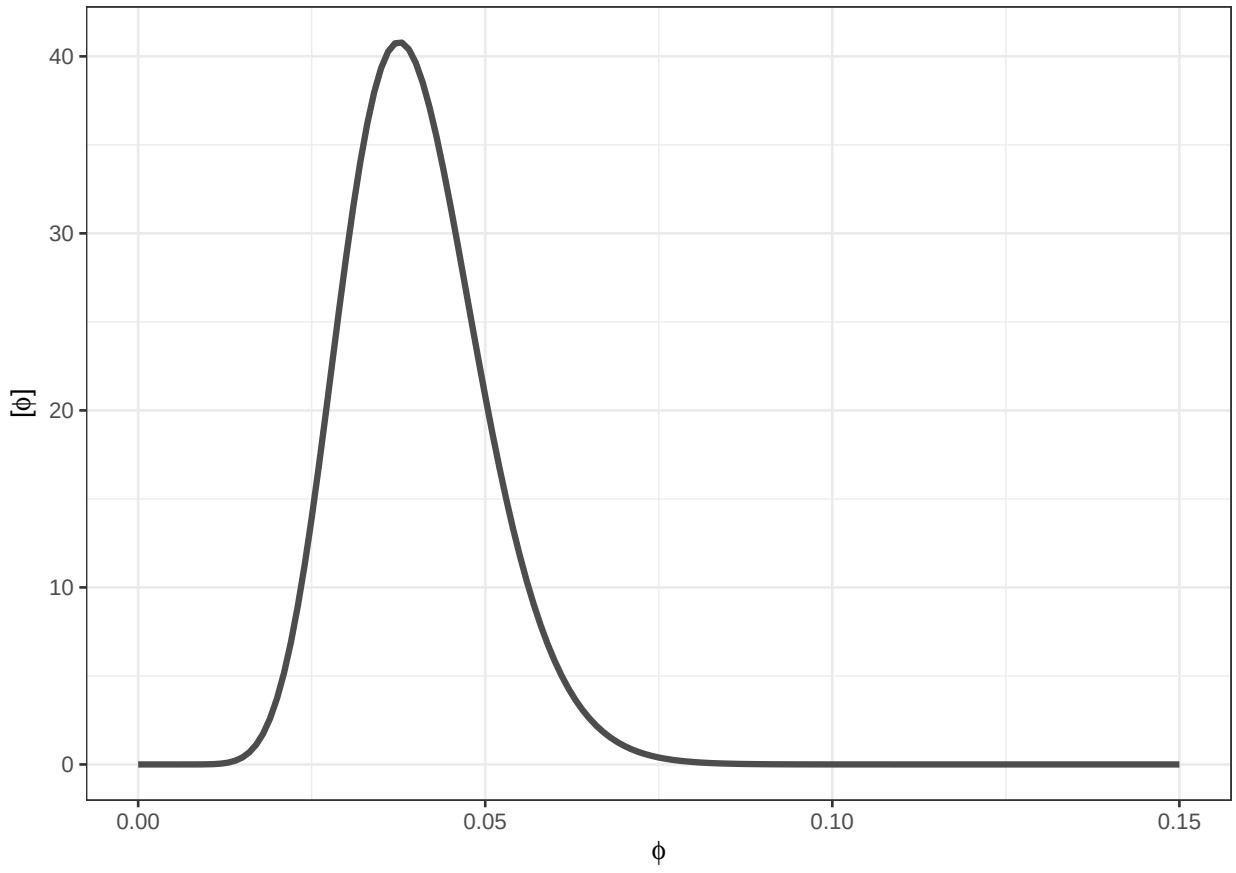
We are interested in the proportion ($\phi$) of Maryland counties that contain a coal fired power plant. Existing literature shows that that this proportion has a mean of $\mu = 0.04$ with a standard deviation of $\sigma = 0.01$. Write out a model for the distribution of $\phi$, conditional on $\mu$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

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

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

```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)
```

```r
# 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)
```

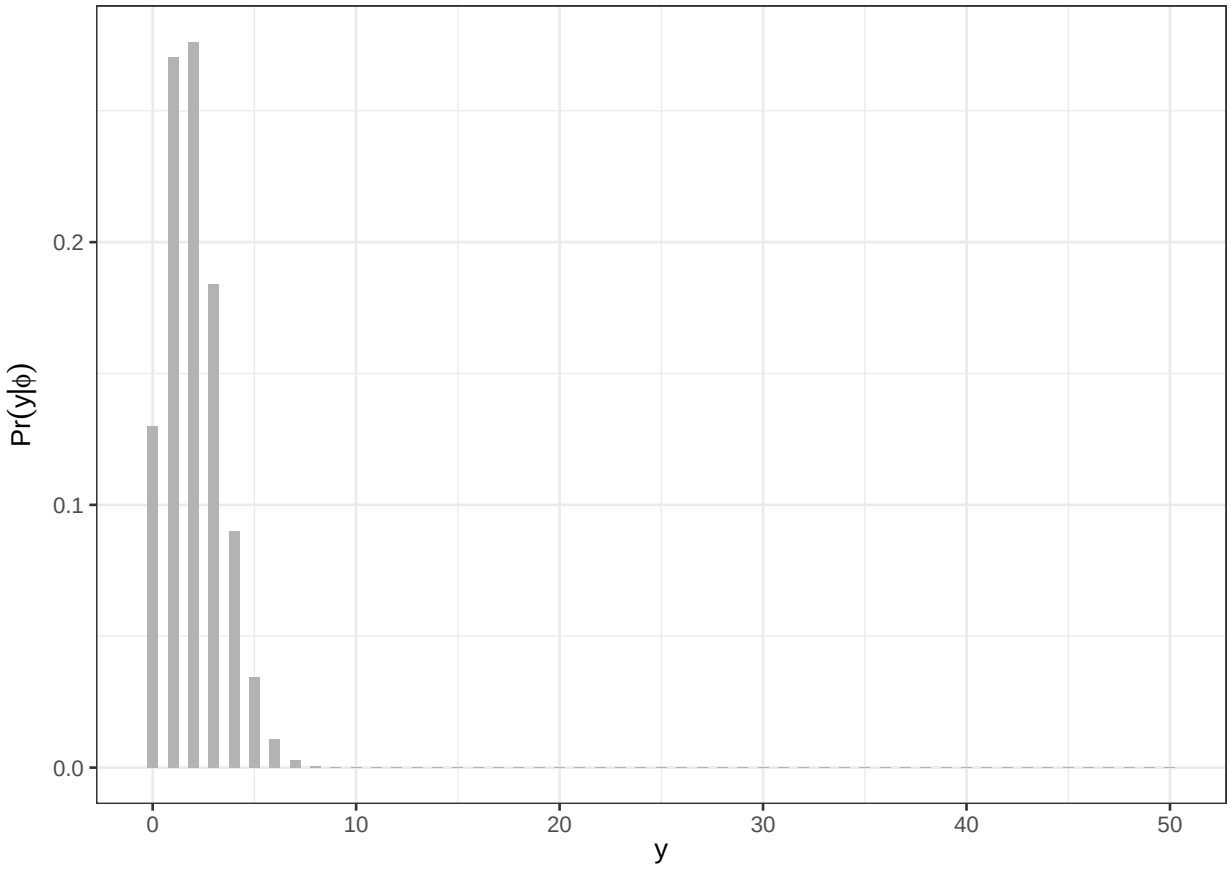
### Answer

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)
```

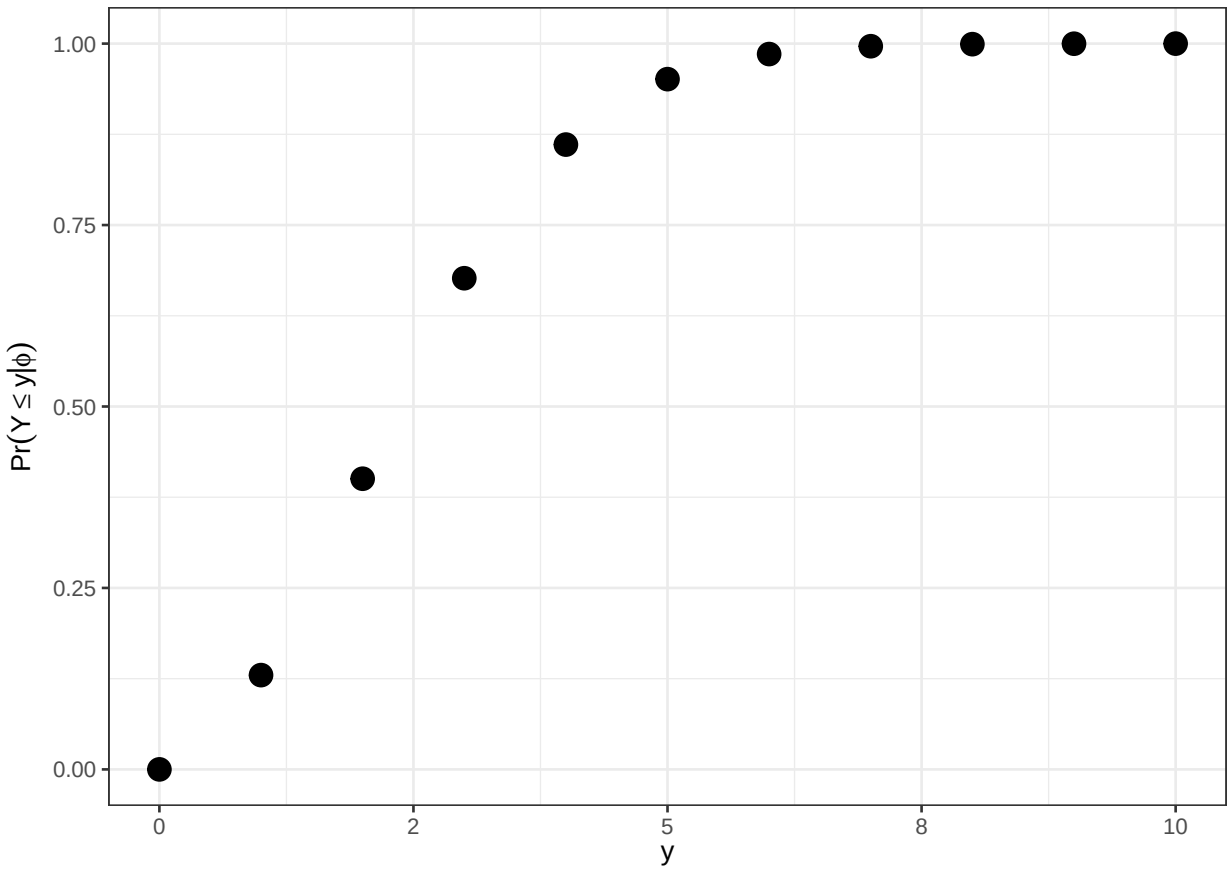
### Answer

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 43 has a coal plant"
```

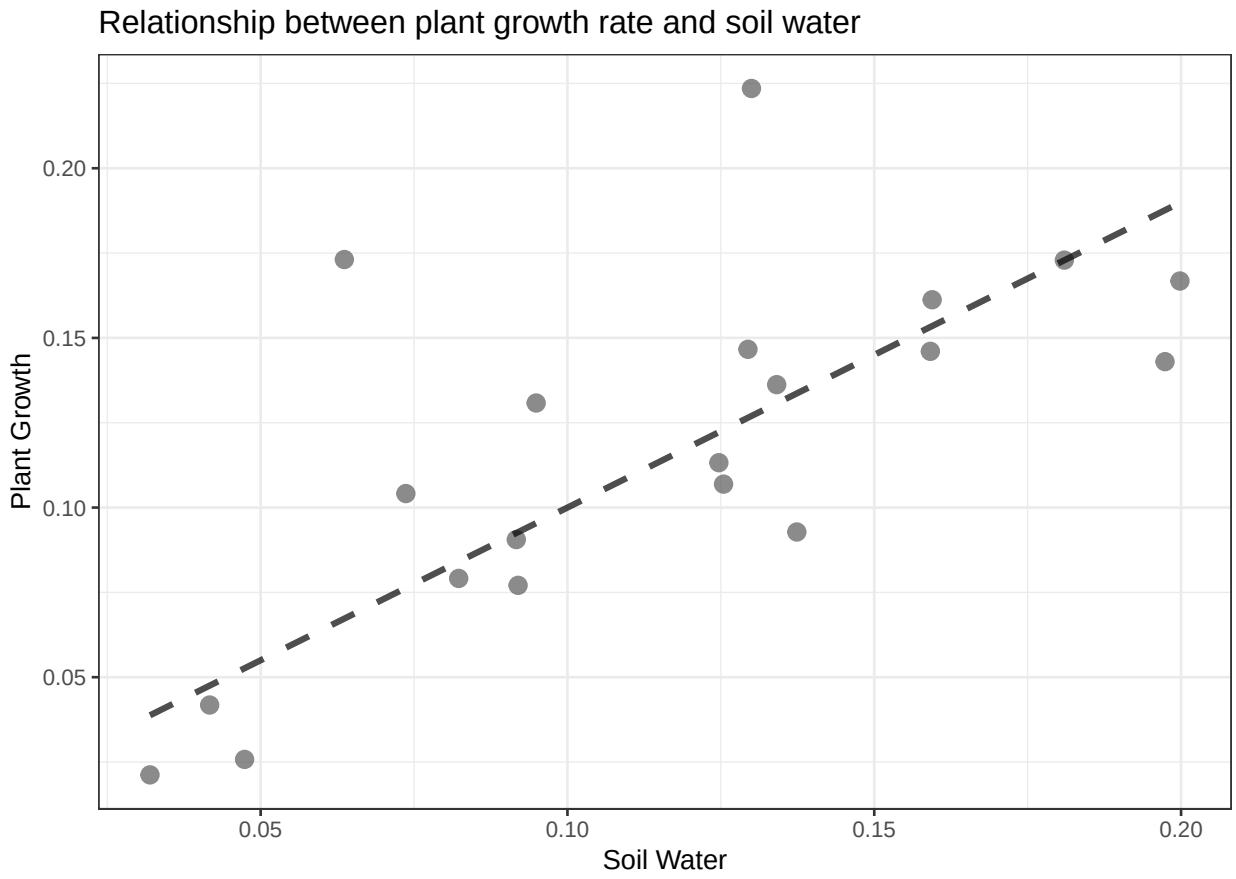
## Question 9

You are modeling the relationship between plant growth rate and soil water. Represent plant growth ($\mu_i$) 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, $\lambda$ 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 $\phi$ is the probability of success on a given trial. Where $\phi$, 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 $\mu$=100 and variance of $\sigma^2$=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: 100.021

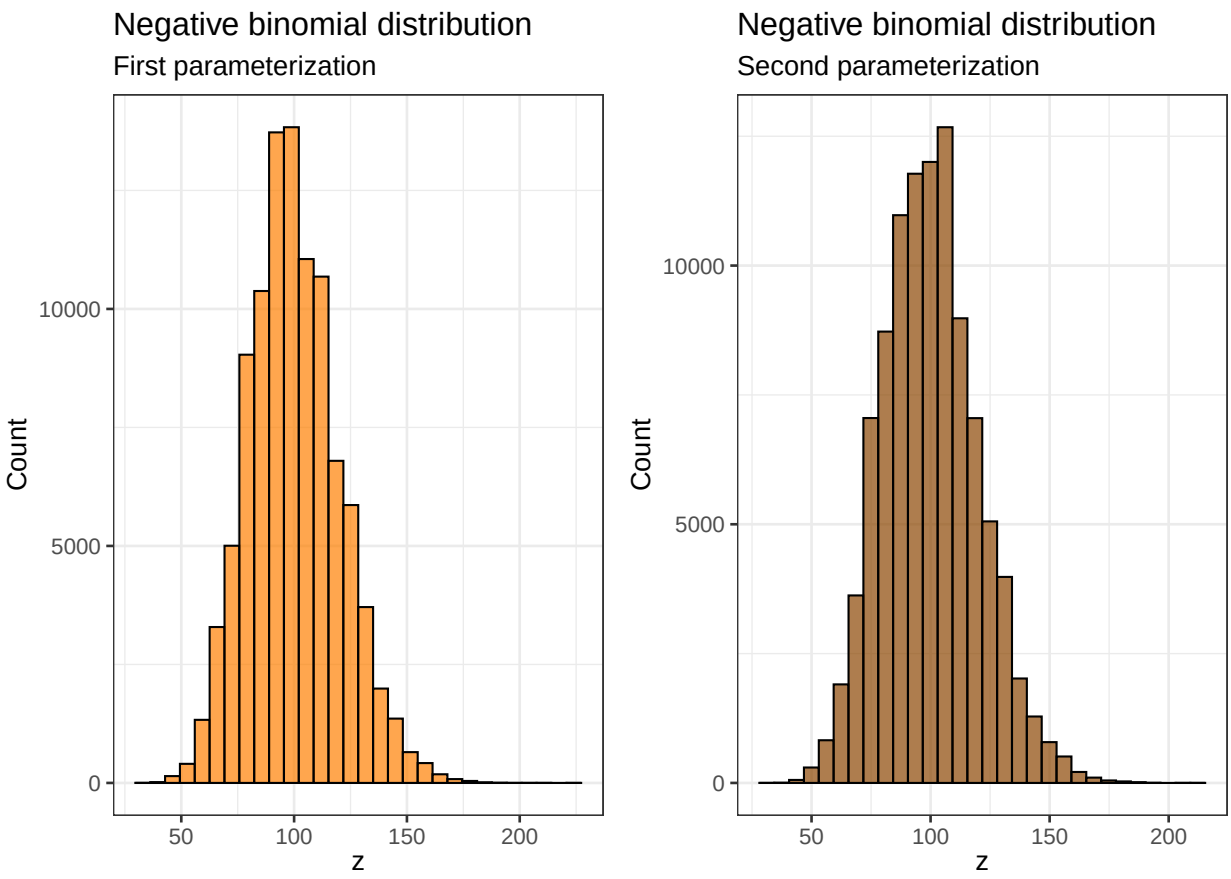
The variance of the first parameterization negative binomial distribution as simulated is: 400.11

The mean of the second parameterization negative binomial distribution as simulated is: 99.957

The variance of the second parameterization negative binomial distribution as simulated is: 399.141

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