Lab 2: Probability Labs

Team USA: Yuting Deng, Kendra Gilbertson, Lily Durkee, Bennett Hardy

Hermione.Deng@colostate.edu, kendra01@colostate.edu, L.Durkee@colostate.edu,Bennett.Hardy@colosta

2022-09-14

Probability lab 1

The probability of an observation y depends on a true ecological state of interest, z, and the parameters in a data model, θ_d. The probability of the true state z depends on the parameters in an ecological process model, θ_p. We know that θ_d and θ_p are independent. Draw the DAG and write a factored expression for the joint distribution, $Pr(y, z, \theta_d, \theta_p)$.

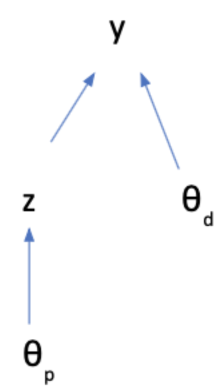

Figure 1: Directed acyclic graph for the joint distribution $Pr(y, z, theta_d, theta_p)$.

The factored expression for the joint distribution is:

$$Pr(y, z, \theta_p, \theta_d) = Pr(y \mid z, \theta_d)Pr(z \mid \theta_p)Pr(\theta_p)Pr(\theta_d)$$

Probability lab 2:

5. The average above ground biomass in a grazing allotment of sagebrush grassland is $103g/m^2$, with a standard deviation of 23. You clip a $1m^2$ plot. Assume the data are normally distributed. Write a model for the probability density of the data point. What is the probability density of an observation of 94? Is there a problem using normal distribution? What is the probability that your plot will contain between 90 and 110 gm of biomass? (For lab report)

$$Y \sim \text{normal } (103, 23)$$

The population density for an observation of 94 is

```
## [1] 0.01606693
```

The problem with the normal is that it takes negative numbers, and you can't have a negative density.

The probability that our plot will contain between 90 and 110 gm of biomass is

```
## [1] 0.3336056
```

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)

$$Y \sim \text{binomial}\,(24, .12)$$

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

```
## [1] 0.1709024
```

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% of the distribution? Use a normal distribution for this problem, but discuss why this might not be a good choice. (For lab report)

```
## [1] -0.8439496
```

Probability lab 3:

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

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

where z_i is a vector of random variables, μ is a vector of means (which can be the output of a deterministic model) and Σ is a variance covariance matrix. The diagonal of Σ contains the variances and the off diagonal contains the covariance of $\sum[i, j]$. 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.

If the rates were not correlated the points would be dispersed throughout the plot, there there would be no clumping along the 1:1 line.

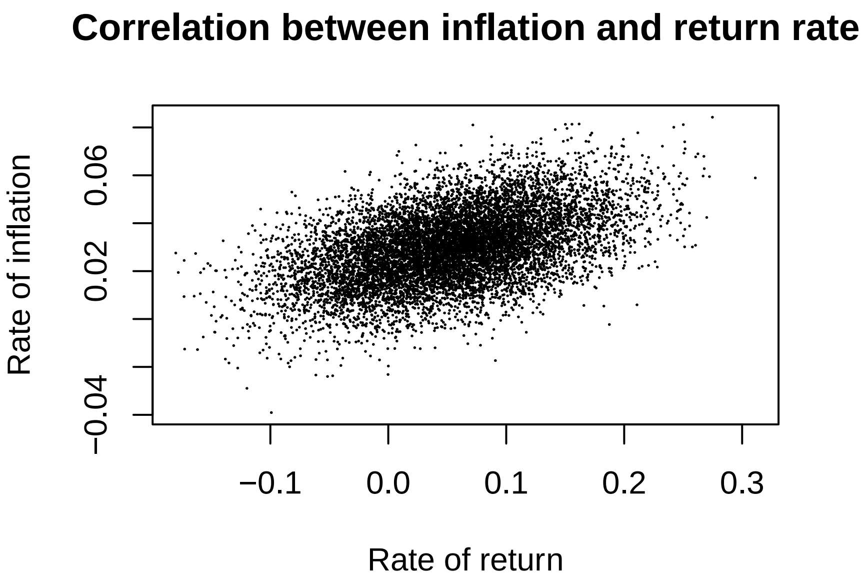

Figure 2: Correlation between the rate of inflation and rate of return.

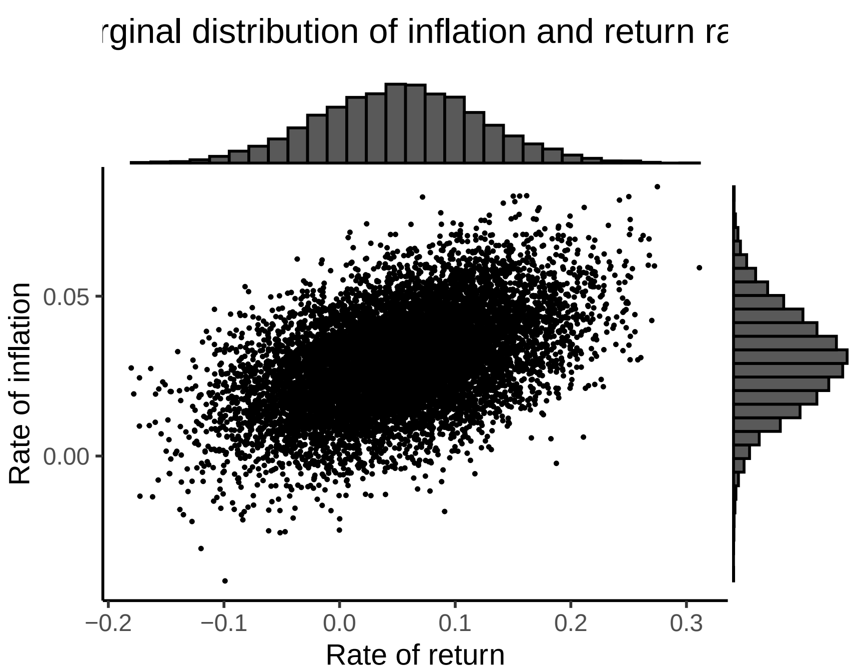

Figure 3: Marginal distributions of inflation and rate of return.

Probability lab 4

Probability mass functions and probability density functions have alternative parameterizations, which means they have different parameters, but yield the same probability or probability density for the same value of a random variable.

The Poisson distribution is often used for count data, despite the fact that one must assume the mean and variance are equal. The negative binomial distribution is a more robust alternative, allowing the variance to differ from the mean. There are two parameterizations for the negative binomial. The first is more frequently used by ecologists:

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

,

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. Note that $\phi = r/(\lambda + r)$.

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.

Code

```
knitr::opts_chunk$set(echo = FALSE,
  attr.source = '.numberLines',
  warning = FALSE,              # don't show warnings
  message = FALSE,               # don't show messages (less serious warnings)
  fig.width=4.5,
  fig.height=3.5
)
library(ggplot2)
library(gridExtra)
library(dplyr)
library(ggpubr)
library(mathjaxr)
dnorm(94,mean=103,sd=23)
pnorm(110,103,23)-pnorm(90,103,23)
dbinom(4,24,.12)
qnorm(.025,1.9,1.4)
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",main='Cor
# show an approximate plot of the marginal distribution of each rv
library(ggExtra)
x<-as.data.frame(x)
p <- ggplot(x, aes(V1, V2)) + geom_point(size=.3) + theme_classic()+
  labs(title="Marginal distribution of inflation and return rates",x="Rate of return",y="Rate of inflat

ggExtra::ggMarginal(p, type = "histogram")
# this R markdown chunk generates a code appendix
```