Chapter 5. Vector random variables

- A **vector random variable** $\mathbf{X} = (X_1, X_2, \dots, X_n)$ is a collection of random numbers with probabilities assigned to outcomes.
- $\mathbf{X}$ can also be called a **multivariate random variable**.
- The case with $n = 2$ we call a **bivariate random variable**.
- Saying X and Y are **jointly distributed random variables** is equivalent to saying (X, Y) is a bivariate random variable.
- Vector random variables let us model relationships between quantities.

Example: midterm and final scores

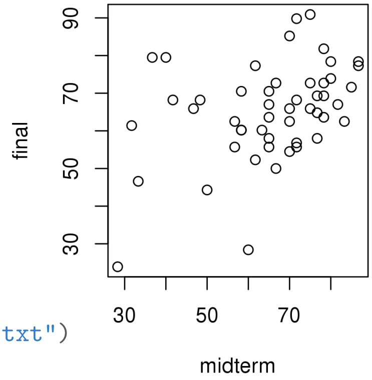
- We will look at the anonymized test scores for a previous course.

```
download.file(destfile="course_progress.txt",
 url="https://ionides.github.io/401f18/01/course_progress.txt")
```

```
# Anonymized scores for a random subset of 50 students
"final" "quiz" "hw" "midterm"
"1" 52.3 76.7 91 61.7
"2" 68.2 65.4 94.5 48.3
"3" 78.4 91.2 95.5 80
```

- A probability model lets us answer a question like, "What is the probability that someone gets at least 70% in both the midterm and the final"

```
x <- read.table("course_progress.txt")
plot(final~midterm,data=x)
```



The bivariate normal distribution and covariance

- Let $X \sim \mathrm{normal}(\mu_X, \sigma_X)$ and $Y \sim \mathrm{normal}(\mu_Y, \sigma_Y)$.
- If X and Y are bivariate random variables we need another parameter to describe their dependence. If X is big, does Y tend to be big, or small, or does the value of X make no difference to the outcome of Y?
- This parameter is the **covariance**, defined to be

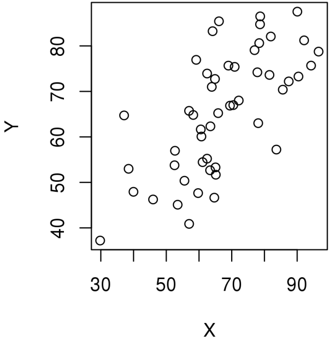
$$\mathrm{Cov}(X,Y) = \mathrm{E}\Big[\big(X - \mathrm{E}[X]\big)\big(Y - \mathrm{E}[Y]\big)\Big]$$

- The parameters of the bivariate normal distribution in matrix form are the **mean vector** $\boldsymbol{\mu} = (\mu_X, \mu_Y)$ and the **variance/covariance matrix**,

$$\mathbb{V} = \begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Var}(Y) \end{bmatrix}$$

- In R, the `mvtnorm` package lets us simulate the bivariate and multivariate normal distribution. It uses the vector and matrix form for the parameters.

Experimenting with the bivariate normal distribution



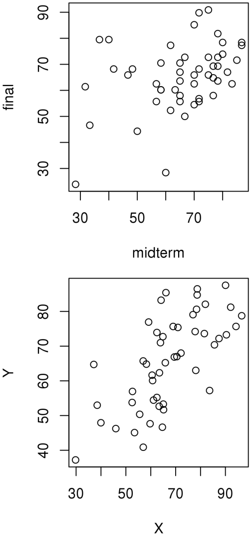

```r
library(mvtnorm)
mvn <- rmvnorm(n=50,
  mean=c(X=65,Y=65),
  sigma=matrix(
    c(200,100,100,150),
    2,2)
)
plot(Y~X,data=mvn)
```

- We write $(X, Y) \sim \text{MVN}(\boldsymbol{\mu}, \mathbb{V})$, where MVN is read "multivariate normal".

Question 5.1. What are μ_X, μ_Y, $\text{Var}(X)$, $\text{Var}(Y)$, and $\text{Cov}(X, Y)$ for this simulation?

The bivariate normal as a model for exam scores

Question 5.2. Compare the data on midterm and final scores with the simulation. Does a normal model seem to fit? Would you expect it to? Why, and why not?



More on covariance

- Covariance is **symmetric**: we see from the definition

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}\Big[\big(X-\mathrm{E}[X]\big)\big(Y-\mathrm{E}[Y]\big)\Big] \\ &= \mathrm{E}\Big[\big(Y-\mathrm{E}[Y]\big)\big(X-\mathrm{E}[X]\big)\Big] = \mathrm{Cov}(Y,X)\end{aligned}$$

- Also, we see from the definition that $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
- The **sample covariance** of n pairs of measurements $(x_1,y_1),\ldots,(x_n,y_n)$ is

$$\mathrm{cov}(\mathbf{x},\mathbf{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

where $\bar{x}$ and $\bar{y}$ are the sample means of $\mathbf{x} = (x_1,\ldots,x_n)$ and $\mathbf{y} = (y_1,\ldots,y_n)$.

Scaling covariance to give correlation

- The standard deviation of a random variable is interpretable as its scale.
- Variance is interpretable as the square of standard deviation

```
var(x$midterm)

## [1] 218.2155

var(x$final)

## [1] 169.7518

cov(x$midterm,x$final)

## [1] 75.61269
```

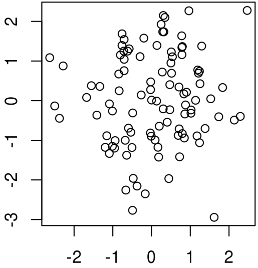
- Covariance is interpretable when scaled to give the **correlation**

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \qquad \mathrm{cor}(\mathbf{x},\mathbf{y}) = \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\sqrt{\mathrm{var}(\mathbf{x})\mathrm{var}(\mathbf{y})}}$$

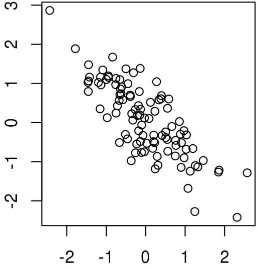
```
cor(x$midterm,x$final)

## [1] 0.3928662
```

```r
rho <- 0
```

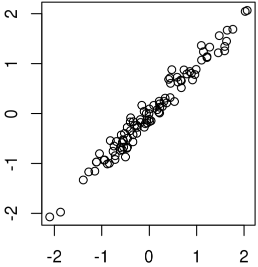



```r
rho <- -0.8
```




```r
library(mvtnorm)
mvn <- rmvnorm(n=100,
  mean=c(X=0,Y=0),
  sigma=matrix(
    c(1,rho,rho,1),
    2,2)
)
```

```r
rho <- 0.95
```



More on interpreting correlation

- Random variables with a correlation of ± 1 (or data with a sample correlation of ± 1) are **linearly dependent**.
- Random variables with a correlation of 0 (or data with a sample correlation of 0) are **uncorrelated**.
- Random variables with a covariance of 0 are also uncorrelated!

Question 5.3. Suppose two data vectors $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ have been **standardized**. That is, each data point has had the sample mean substracted and then been divided by the sample standard deviation. You calculate $\mathrm{cov}(\mathbf{x}, \mathbf{y}) = 0.8$. What is the sample correlation, $\mathrm{cor}(\mathbf{x}, \mathbf{y})$?

The variance of a sum

- A basic property of covariance is

(Eq. C1) $$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$$

- Sample covariance has the same formula,

(Eq. C2) $$\mathrm{var}(\mathbf{x}+\mathbf{y}) = \mathrm{var}(\mathbf{x}) + \mathrm{var}(\mathbf{y}) + 2\,\mathrm{cov}(\mathbf{x},\mathbf{y})$$

- These nice formulas mean it can be easier to calculate using variances and covariances rather than standard deviations and correlations.

Question 5.4. Rewrite (Eq. C1) to give a formula for $\mathrm{SD}(X+Y)$ in terms of $\mathrm{SD}(X)$, $\mathrm{SD}(Y)$ and $\mathrm{Cor}(X,Y)$.

More properties of covariance

- Covariance is not affected by adding constants to either variable

(Eq. C3) $$\mathrm{Cov}(X+a, Y+b) = \mathrm{Cov}(X,Y)$$

- Recall the definition $\mathrm{Cov}(X,Y) = \mathrm{E}\big[\big(X - \mathrm{E}[X]\big)\,\big(Y - \mathrm{E}[Y]\big)\big]$. In words, covariance is the mean product of deviations from average. These deviations are unchanged when we add a constant to the variable.

- Covariance scales **bilinearly** with each variable

(Eq. C3) $$\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X,Y)$$

- Covariance distributes across sums

(Eq. C4) $$\mathrm{Cov}(X, Y+Z) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,Z)$$

- Sample covariances also have these properties. You can test them in R using bivariate normal random variables, constructed as previously using ‘rmvnorm()‘.

The variance matrix of a vector random variable

- Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector random variable.
- For any pair of elements, say X_i and X_j, we can compute the usual scalar covariance, $v_{ij} = \mathrm{Cov}(X_i, X_j)$.
- The variance matrix $\mathbb{V} = [v_{ij}]_{n \times n}$ collects together all these covariances.

$$\mathbb{V} = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & & \mathrm{Cov}(X_2, X_n) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \ldots & \mathrm{Cov}(X_n, X_n) \end{bmatrix}$$

- The diagonal entries of $\mathbb{V}$ are $v_{ii} = \mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$ for $i = 1, \ldots, n$. This lets us write the variance matrix as

$$\mathbb{V} = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & & \mathrm{Cov}(X_2, X_n) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \ldots & \mathrm{Var}(X_n) \end{bmatrix}$$

The correlation matrix

- Covariance is harder to interpret than correlation, but easier for calculations.
- We can put together all the correlations into a correlation matrix, using the fact that $\text{Cor}(X_i, X_i) = 1$.

$$\text{Cor}\mathbf{X} = \begin{bmatrix} 1 & \text{Cor}(X_1, X_2) & \dots & \text{Cor}(X_1, X_n) \\ \text{Cor}(X_2, X_1) & 1 & & \text{Cor}(X_2, X_n) \\ \vdots & & \ddots & \vdots \\ \text{Cor}(X_n, X_1) & \text{Cor}(X_n, X_2) & \dots & 1 \end{bmatrix}$$

- Multivariate distributions can be very complicated.
- The variance/covariance and correlation matrices deal only with **pairwise** relationships between variables.
- Pairwise relationships can be graphed.