Chapter 5. Vector random variables

- A **vector random variable** $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 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.

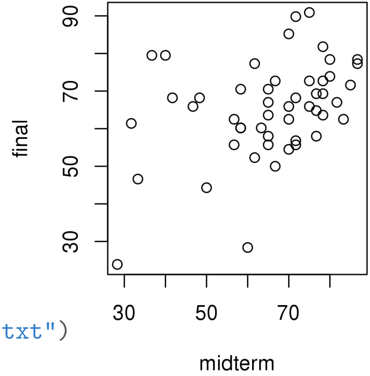
Example: midterm and final scores

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

```
download.file(destfile="course_progress.txt",
 url="https://ionides.github.io/401f18/05/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 $X \sim \text{normal}(\mu_X, \sigma_X)$ and $Y \sim \text{normal}(\mu_Y, \sigma_Y)$.
- 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

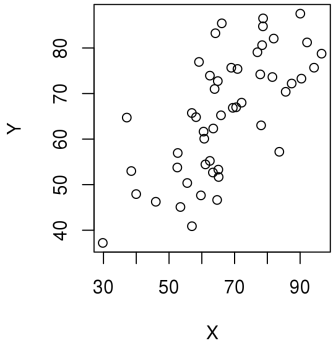
$$\text{Cov}(X,Y) = \text{E}\Big[\big(X - \text{E}[X]\big)\,\big(Y - \text{E}[Y]\big)\Big]$$

- 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**,

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

- In R, the `mvtnorm` package lets us simulate the bivariate and multivariate normal distribution via the `rmvnorm()` function. It has the mean vector and variance/covariance matrix as arguments.

Experimenting with the bivariate normal distribution



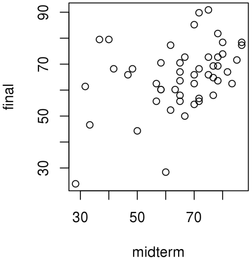

```r
library(mvtnorm)
mvn <- rmvnorm(n=50,
  mean=c(X=65,Y=65),
  sigma=matrix(
    c(200,100,100,150),
    2,2)
)
plot(Y~X,data=mvn)
```

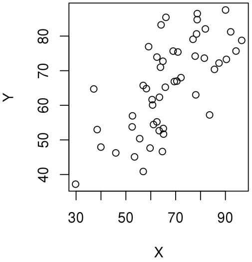
- We write $(X, Y) \sim \text{MVN}(\boldsymbol{\mu}, \mathbb{V})$, where MVN is read "multivariate normal".

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
```

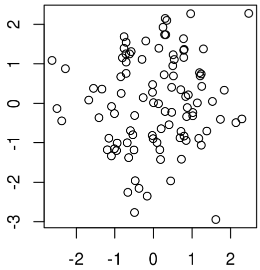
- Covariance is interpretable when scaled to give the **correlation**

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

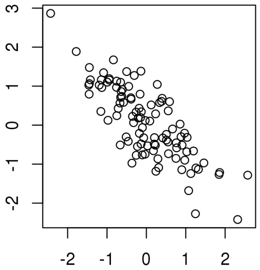
```
cor(x$midterm,x$final)

## [1] 0.3928662
```

```r
rho <- 0
```

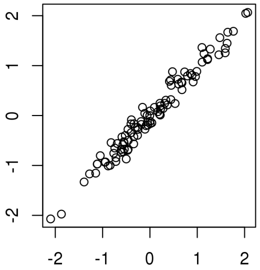



```r
rho <- -0.8
```




```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 $\operatorname{cov}(\mathbf{x}, \mathbf{y}) = 0.8$. What is the sample correlation, $\operatorname{cor}(\mathbf{x}, \mathbf{y})$?

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/covariance matrix of vector random variables

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

$$\mathbb{V} = \text{Var}(\mathbf{X}) = \begin{bmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \ldots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & & \text{Cov}(X_2, X_p) \\ \vdots & & \ddots & \vdots \\ \text{Cov}(X_p, X_1) & \text{Cov}(X_p, X_2) & \ldots & \text{Cov}(X_p, X_p) \end{bmatrix}$$

- The diagonal entries of $\mathbb{V}$ are $v_{ii} = \text{Cov}(X_i, X_i) = \text{Var}(X_i)$ for $i = 1, \ldots, p$ so the variance/covariance matrix can be written as

$$\mathbb{V} = \text{Var}(\mathbf{X}) = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \ldots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & & \text{Cov}(X_2, X_p) \\ \vdots & & \ddots & \vdots \\ \text{Cov}(X_p, X_1) & \text{Cov}(X_p, X_2) & \ldots & \text{Var}(X_p) \end{bmatrix}$$

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 $\mathrm{Cor}(X_i, X_i) = 1$.

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

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

The sample variance/covariance matrix

- The **sample variance/covariance matrix** places all the sample variances and covariances in a matrix.
- Let $\mathbb{X} = [x_{ij}]_{n \times p}$ be a data matrix made up of p data vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x_p}$ each of length n.

$$\mathrm{var}(\mathbb{X}) = \begin{bmatrix} \mathrm{var}(\mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_2) & \ldots & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_p) \\ \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_1) & \mathrm{var}(\mathbf{x}_2) & & \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_p) \\ \vdots & & \ddots & \vdots \\ \mathrm{cov}(\mathbf{x}_p, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_p, \mathbf{x}_2) & \ldots & \mathrm{var}(\mathbf{x}_p) \end{bmatrix}$$

- R uses the same notation. If `x` is a matrix or dataframe, `var(x)` returns the sample variance/covariance matrix.

```
var(x)

##              final      quiz        hw   midterm
## final   169.75184  78.14294  51.27143  75.61269
## quiz     78.14294 224.39664 103.57755 107.32550
## hw       51.27143 103.57755 120.13265  61.44694
## midterm  75.61269 107.32550  61.44694 218.21553
```

The sample correlation matrix

- The **sample correlation matrix** places all the sample correlations in a matrix.
- Let $\mathbb{X} = [x_{ij}]_{n \times p}$ be a data matrix made up of p data vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x_p}$ each of length n.

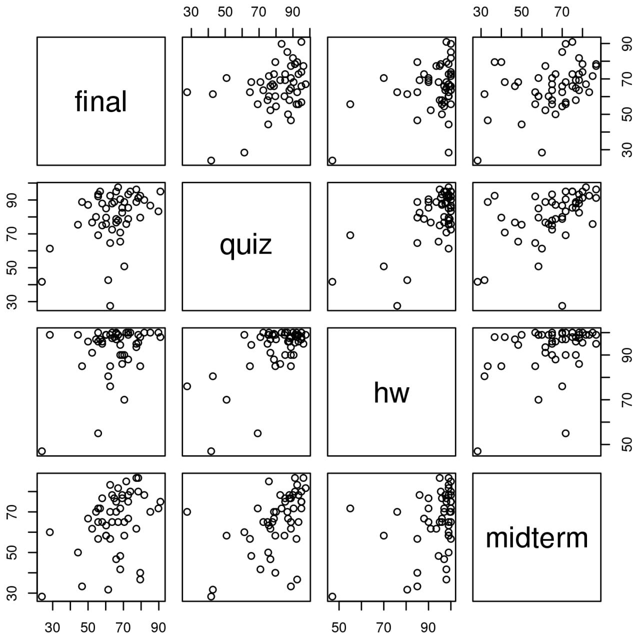
$$\mathrm{cor}(\mathbb{X}) = \begin{bmatrix} 1 & \mathrm{cor}(\mathbf{x}_1, \mathbf{x}_2) & \ldots & \mathrm{cor}(\mathbf{x}_1, \mathbf{x}_p) \\ \mathrm{cor}(\mathbf{x}_2, \mathbf{x}_1) & 1 & & \mathrm{cor}(\mathbf{x}_2, \mathbf{x}_p) \\ \vdots & & \ddots & \vdots \\ \mathrm{cor}(\mathbf{x}_p, \mathbf{x}_1) & \mathrm{cor}(\mathbf{x}_p, \mathbf{x}_2) & \ldots & 1 \end{bmatrix}$$

- R uses the same notation. If `x` is a matrix or dataframe, `cor(x)` returns the sample correlation matrix.

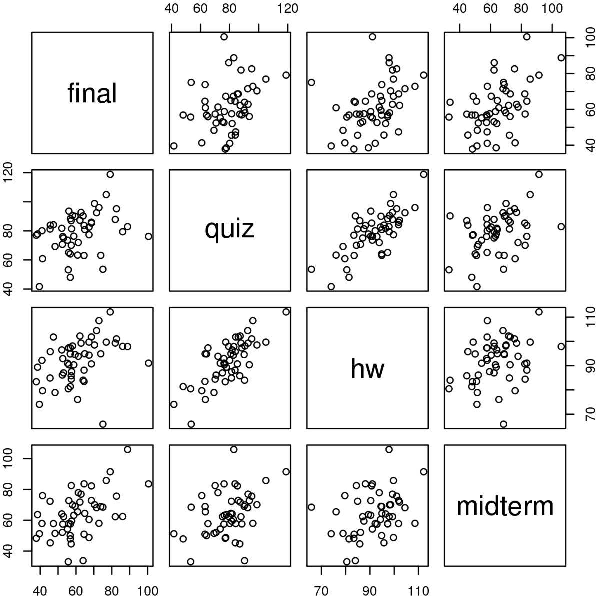
```
cor(x)

##             final      quiz        hw   midterm
## final   1.0000000 0.4003818 0.3590357 0.3928662
## quiz    0.4003818 1.0000000 0.6308512 0.4850114
## hw      0.3590357 0.6308512 1.0000000 0.3795132
## midterm 0.3928662 0.4850114 0.3795132 1.0000000
```

```
pairs(x)
```



```
mvn <- rmvnorm(50,mean=apply(x,2,mean),sigma=var(x))
pairs(mvn)
```



Question 5.5. From looking at the scatterplots, what are the strengths and weaknesses of a multivariate normal model for test scores in this course?

Question 5.6. To what extent is it appropriate to summarize the data by the mean and variance/covariance matrix (or correlation matrix) when the normal approximation is dubious?

Linear combinations

- Let $\mathbf{X} = (X_1, \dots, X_p)$ be a vector random variable with mean vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)$ and $p \times p$ variance/covariance matrix $\mathbb{V}$.
- Let $\mathbb{X}$ be a $n \times p$ data matrix.
- Let $\mathbb{A}$ be a $q \times p$ matrix.
- $\mathbf{Z} = \mathbb{A}\mathbf{X}$ is a collection of q linear combinations of the p random variables in the vector $\mathbf{X}$, viewed as a **column** vector.
- $\mathbb{Z} = \mathbb{X}\mathbb{A}^{\mathrm{T}}$ is an $n \times q$ collection of linear combinations of the p data points in each **row** of $\mathbb{X}$.
- Mental gymnastics are required: vectors are often interpreted as **column vectors** (e.g., $p \times 1$ matrices) but the vector of measurements for each unit is a **row vector** when considered as a row of an $n \times p$ data matrix.

Question 5.7. How would you construct a simulated data matrix $\mathbb{Z}_{\text{sim}}$ from n realizations $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ of the random column vector $\mathbf{Z} = \mathbb{A}\mathbf{X}$? Be careful with transposes and keep track of dimensions.

Solution:

- There is a useful matrix variance/covariance formula for a linear combination.

$$\mathrm{Var}(\mathbb{A}\,\mathbf{X}) = \mathbb{A}\,\mathrm{Var}(\mathbf{X})\,\mathbb{A}^{\mathrm{T}} \qquad\qquad \mathrm{var}(\mathbb{X}\,\mathbb{A}^{\mathrm{T}}) = \mathbb{A}\,\mathrm{var}(\mathbb{X})\,\mathbb{A}^{\mathrm{T}}$$

Question 5.8. Add dimensions to each term in these equations to check they make sense.

Testing the variance/covariance formula

- Suppose that the overall course score is weighted 40% on the final and 20% on each of the miterm, homework and quiz.
- We can find the sample variance of the overall score two different ways.

(i) Directly computing the overall score for each student.

```
weights <- c(final=0.4,quiz=0.2,hw=0.2,midterm=0.2)
overall <- as.matrix(x) %*% weights
var(overall)

##          [,1]
## [1,] 104.2624
```

(ii) Using $\mathrm{var}(\mathbb{X}\,\mathbb{A}^{\mathrm{T}}) = \mathbb{A}\,\mathrm{var}(\mathbb{X})\,\mathbb{A}^{\mathrm{T}}$.

```
weights %*% var(x) %*% weights

##          [,1]
## [1,] 104.2624
```

- R interprets the vector ‘weights‘ as a row or column vector as necessary.

Independence

- Two events E and F are **independent** if

$$\mathrm{P}(E \text{ and } F) = \mathrm{P}(E) \times \mathrm{P}(F)$$

Worked example 5.1. Suppose we have a red die and a blue die. They are idea fair dice, so the values should be independent. What is the chance they both show a six?

(a) Using the definition of independence.

(b) By considering equally likely outcomes, without using the definition.

- The multiplication rule agrees with an intuitive idea of independence.

Independence of random variables

- X and Y are **independent random variables** if, for any intervals $[a, b]$ and $[c, d]$,

$$\mathrm{P}(a < X < b \text{ and } c < Y < d) = \mathrm{P}(a < X < b) \times \mathrm{P}(c < Y < d)$$

- This definition extends to vector random variables. $\mathbf{X} = (X_1, \ldots, X_n)$ is a **vector of independent random variables** if for any collection of intervals $[a_i, b_i]$, $1 \leq i \leq n$,

$$\mathrm{P}(a_1 < X_1 < b_1, \ldots, a_n < X_n < b_n) = \mathrm{P}(a_1 < X_1 < b_1) \times \cdots \times \mathrm{P}(a_n < X_n < b_n)$$

- $\mathbf{X} = (X_1, \ldots, X_n)$ is a **vector of independent identically distributed (iid) random variables** if, in addition, each element of $\mathbf{X}$ has the same distribution.

- "$X_1, \ldots, X_n$ are n random variables with the $\mathrm{normal}(\mu, \sigma)$ distribution" is written more formally as "Let $X_1, \ldots, X_n \sim \mathrm{iid\ normal}(\mu, \sigma)$."

Independent vs uncorrelated

- If X and Y are independent they are uncorrelated.
- The converse is not necessarily true.
- **For normal random variables, the converse is true**.
- If X and Y are bivariate normal random variables, and $\mathrm{Cov}(X, Y) = 0$, then X and Y are independent.
- The following slide demonstrated the possibility of being uncorrelated but not independent (for non-normal random variables).