DATA 580

Modeling and Simulation I

Expected Value

Covariance and Correlation

Sums of Random Variables

Central Limit Theorem

Autoregressive Time Series Models

Expected Value

- **Definition: If X_1 and X_2 have joint density $f(y_1, y_2)$, then**

$$\mathbf{E}[g(X_1, X_2)] = \int \int g(y_1, y_2) f(y_1, y_2) dy_1 dy_2$$

- *Example.* **For the propane example of an earlier lecture, the joint pdf of temperature and amount is**

$$f(x, t) = \frac{x + \frac{t}{5} - 13}{5}, \; x \in (10, 11), t \in (15, 20)$$

The pressure in the cylinder is proportional to XT. Suppose the relation is

$$P = 3XT$$

Find $E[P]$.

Expected Value

$$E[P] = E[3XT] = \int_{10}^{11} \int_{15}^{20} 3xt \frac{x + \frac{t}{5} - 13}{5} dt dx$$

$$= \int_{10}^{11} \frac{105x^2 - 995x}{2} dx = 568.75$$

Expectations of Sums

$$\mathbf{E}[X_1 + X_2] = \int\int (y_1 + y_2) f(y_1, y_2) dy_1 dy_2$$

$$= \int\int y_1 f(y_1, y_2) dy_1 dy_2 + \int\int y_2 f(y_1, y_2) dy_1 dy_2$$

$$= \mathbf{E}[X_1] + \mathbf{E}[X_2]$$

$$\mathbf{E}[X_1 + X_2 + X_3] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \mathbf{E}[X_3]$$

$$\mathbf{E}[X_1 + X_2 + X_3 + X_4] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \mathbf{E}[X_3] + \mathbf{E}[X_4]$$

and so on ....

Expectations of Sums

Example. **Because of contaminants in the propane, and because of interactions among the propane gas molecules, etc., the pressure is more accurately modelled as**

$$P = 3XT + \varepsilon$$

where ε is a random variable representing all unaccounted for factors (noise). We assume $E[\varepsilon] = 0$.

Find $E[P]$.

$$E[P] = E[3XT + \varepsilon] = E[3XT] + E[\varepsilon]$$

$$= 568.75 + 0$$

$$= 568.75$$

Expected Values of Averages

Suppose $X_1, X_2, \ldots, X_n$ represent a sample of measurements from a population where $\mathbf{E}[X_1] = \cdots = \mathbf{E}[X_n] = \mu$. Then

$$\mathbf{E}[X_1 + X_2 + \cdots + X_n] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_n] = n\mu$$

$\Rightarrow$

$$\mathbf{E}[\bar{X}] = \mu$$

where

$$\bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k$$

Expected Values of Averages

Example. **Measurements were taken on the amount of vibration (in microns) produced by six electric motors all having the same type of bearings. Each such measurement has been modelled with the density function**

$$f(y) = \frac{1}{10} e^{-(y-5)/10}, \quad y > 5$$

Find the expected value of the average of the 6 vibration measurements.

Letting μ denote the common expected value, we have

$$\mu = \mathbf{E}[X_1] = \cdots = \mathbf{E}[X_6] = \int_5^\infty y \frac{1}{10} e^{-(y-5)/10} dy = 15$$

$\Rightarrow$

$$\mathbf{E}[\bar{X}] = \mu = 15$$

Covariance and Correlation

Covariance:

$$\textbf{Cov}(X_1, X_2) = E[X_1 X_2] - E[X_1]E[X_2]$$

This is a measure of *linear* dependence between two measurements.

Correlation:

$$\rho = \textbf{Corr}(X_1, X_2) = \frac{\textbf{Cov}(X_1, X_2)}{\sqrt{V(X_1)V(X_2)}}$$

This is a related measure. It can take values only between -1 and 1.

If ρ is positive, we say that there is a positive linear relationship between X_2 and X_1.

If ρ is negative, we say that there is a negative linear relationship between X_2 and X_1.

Dependent Exponential Random Variables

For example, consider the random variables with the following joint probability density function

$$f(x_1, x_2) = \frac{\lambda}{x_1} e^{-\lambda x_1 - x_2/x_1}, \qquad x_1, x_2 \geq 0$$

and 0, otherwise. We can see that X_1 and X_2 are positively associated as follows.

$$E[X_1 X_2] = \lambda \int_0^\infty \int_0^\infty \frac{x_1 x_2}{x_1} e^{-\lambda x_1 - x_2/x_1} dx_1 dx_2 = \frac{2}{\lambda^2}.$$

To compute this integral, it is necessary to use integration-by-parts several times.

Thus,

$$\textbf{Cov}(X_1, X_2) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

This value is positive which means that X_1 and X_2 are positively related.

Calculation of covariance and correlation for a sample

For a sample $\{(x_{11}, x_{21}), (x_{12}, x_{22}), \ldots, (x_{1n}, x_{2n})\}$, the *sample covariance* is given by

$$c = \frac{1}{n-1} \sum_{j=1}^{n} (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2).$$

The *sample correlation* is given by

$$r = \frac{c}{s_1 s_2}$$

where s_1 and s_2 are the sample standard deviations of the samples of x_1's and x_2's respectively.

The `cor()` function calculates this quantity.

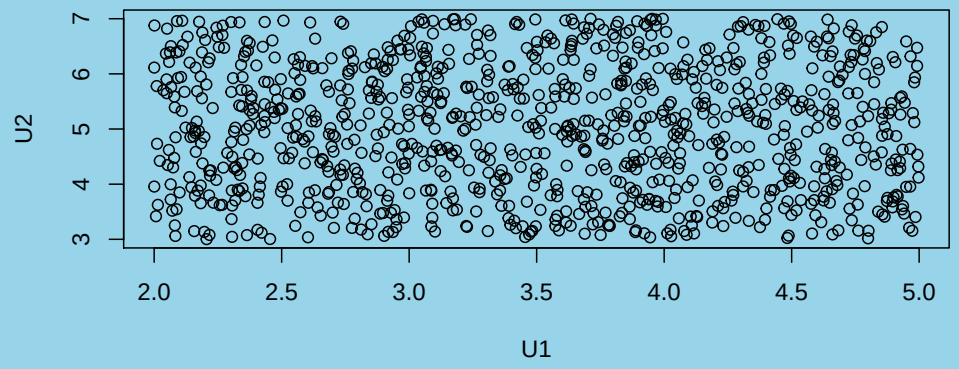
Correlation

Example.

```
## [1] -0.02292062
```

Simulated Pairs of Independent Uniforms:

```
U1 <- runif(1000, 2, 5)
U2 <- runif(1000, 3, 7)
cor(U1, U2); plot(U2 ~ U1)
```



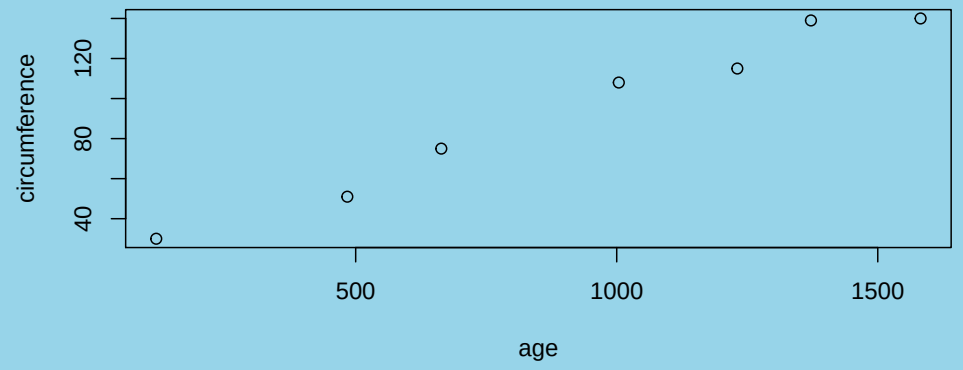
The correlation is small, close to 0, and the scatterplot shows no pattern.

Correlation

Example.

Orange tree circumference versus age (`Orange` data frame)

```
Orange3 <-
    subset(Orange, Tree == 3)
    # data on Tree No. 3
plot(circumference ~ age,
    data = Orange3)
with(Orange3,
    cor(circumference, age))
```




```
## [1] 0.9881766
```

The correlation is large and positive, and the points scatter about a line with positive slope.

Correlation versus Dependence

What is the difference between correlation and dependence?

Correlation ... is a measure of *linear* dependence.

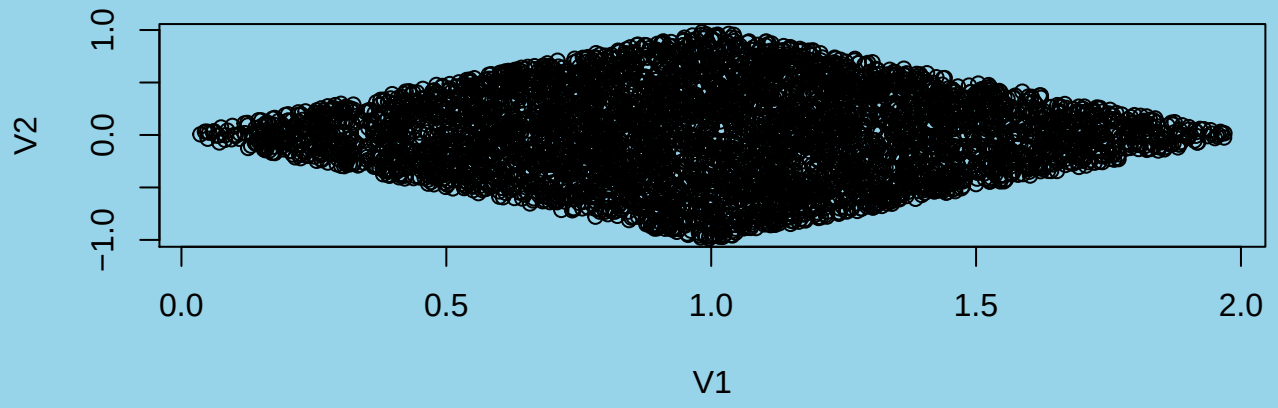
2 random variables may be dependent, but not correlated.

Correlation versus Dependence

Example.

```
U1 <- runif(5000); U2 <- runif(5000)
V1 <- U1 + U2;  V2 <- U1 - U2
cor(V1, V2); plot(V2 ~ V1)


## [1] -0.01021503
```



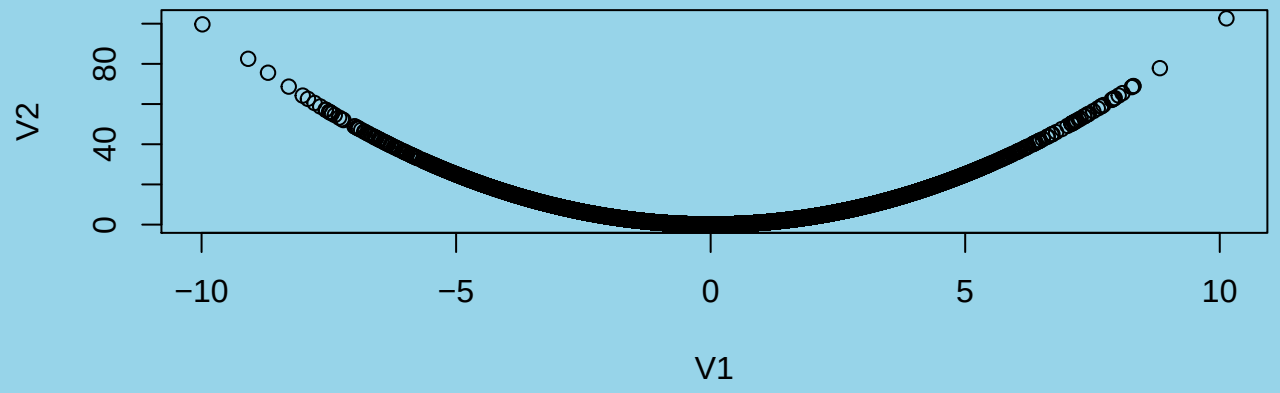
Are V_1 and V_2 dependent?

Are V_1 and V_2 correlated?

Example 2

```
U1 <- rexp(50000); U2 <- rexp(50000)
V1 <- U1 - U2;  V2 <- V1^2
cor(V1, V2);  plot(V2 ~ V1)


## [1] 0.01153105
```



Are V_1 and V_2 dependent?

Are V_1 and V_2 correlated?

Variances of Sums of Independent R.V.s

Suppose X_1 and X_2 are independent random variables. Then

$$\mathbf{E}[X_1 X_2] = \int\int y_1 y_2 f_1(y_1) f_2(y_2) dy_1 dy_2 = \mathbf{E}[X_1]\mathbf{E}[X_2]$$

and

$$\mathbf{E}[(X_1 + X_2)^2] = \mathbf{E}[X_1^2] + 2\mathbf{E}[X_1 X_2] + \mathbf{E}[X_2^2]$$

so

$$\mathbf{Var}(X_1 + X_2) = \mathbf{E}[(X_1 + X_2)^2] - (\mathbf{E}[X_1 + X_2])^2$$

$$= \mathbf{E}[X_1^2] + \mathbf{E}[X_2^2] - (\mathbf{E}[X_1])^2 - (\mathbf{E}[X_2])^2$$

$$= \mathbf{Var}(X_1) + \mathbf{Var}(X_2)$$

Variances of Sums of Independent R.V.s

For n independent random variables $X_1, X_2, \ldots, X_n$,

$$\textbf{Var}(X_1 + X_2 + \cdots + X_n) = \textbf{Var}(X_1) + \cdots \textbf{Var}(X_n)$$

Suppose $X_1, X_2, \ldots, X_n$ is a sample of independent measurements. If the variance of each is σ^2, then

$$\textbf{Var}(X_1 + X_2 + \cdots + X_n) = n\sigma^2$$

and

$$\textbf{Var}(\bar{X}) = \sigma^2/n$$

where

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

Example. **Find the variance of the average of the 6 vibration measurements.**

$$\sigma^2 = \mathbf{E}[X_1^2] - \mathbf{E}[X_1]^2$$

$$\mathbf{E}[X_1^2] = \frac{1}{10}\int_5^\infty y^2 e^{-(y-5)/10} dy = 325$$

$$\Rightarrow \sigma^2 = 100$$

$$\Rightarrow \mathbf{Var}(\bar{X}) = 100/6 = 16.7$$

The Distribution of the Sample Average

- **Suppose $X_1, X_2, \ldots, X_n$ are independent measurements coming from a normal population with mean μ and variance σ^2. (i.e. a normal random sample)**

- **What is the sampling distribution of $\bar{X}$?**

  The density function for $\bar{X}$ is normal with expected value μ and variance σ^2/n.

Example. **Drying times for a certain paint under certain temperature and humidity conditions can be modelled as a normally distributed random variable with expected value 75 minutes and variance 81 minutes^2.**

* **An additive might speed up drying.**

* **4 drying times were recorded using the additive.**

* **If the additive has no effect, find the probability that the average of these drying times would be less than 70 minutes.**

  Ans. $\bar{X}$ **is normally distributed with mean 75 and variance** $81/4$**, i.e.** $\sigma = 4.5$**.**

```
pnorm(70, 75, 4.5)

## [1] 0.1332603
```

Central Limit Theorem

- **Suppose $X_1, X_2, \ldots, X_n$ are independent measurements coming from a population with mean μ and variance σ^2. (i.e. the population does *not* have to be normal)**

- **What is the sampling distribution of $\bar{X}$?**

  The density function of Z is approximately standard normal, for large n.

  The density function for $\bar{X}$ is approximately normal with expected value μ and variance σ^2/n.

- **In practice, the magnitude of n needed for a reasonable approximation will depend on how skewed or heavy-tailed the underlying distribution is. For a uniform distribution, a sample of size 5 might be large enough, and for an exponential distribution, a sample of size 30 might be needed.**

Example. **The breaking strength of a rivet has an expected value of 10000 psi and a variance of 250000 psi. The breaking strengths of 40 rivets are measured. Find the probability that the average of the measurements is between 9900 and 10200.**

$$P(9900 < \bar{X} < 10200) \doteq$$

$$P(-1.26 < Z < 2.53) = .890$$

Find the probability that 1 of the rivets has a breaking strength between 9900 and 10200?

We don't know the distribution of the measurements.

Example. **The lifetime of a type of battery is approximately normally distributed with expected value 10 hours and standard deviation 3 hours. A package contains 4 batteries. When camping, I plan to use a flashlight that operates on 1 battery at a time for 35 hours.**

Find the probability of running out of power early.

$$P(\bar{X} < 35/4) = P(Z < -.833) \doteq .203$$

A more expensive battery has the same expected value, but a variance of 2.25 hours^2. Find the probability of running out of power early with this brand.

$$P(\bar{X} < 35/4) = P(Z < -1.6) \doteq .0548$$

Example. **The ACME elevator company uses cables which will break when carrying more than 1000 pounds. 7 men board an elevator. If adult male weight is normally distributed with expected value 150 pounds and standard deviation 15 pounds, find the probability that the elevator cable will break.**

$$P(\bar{X} > 1000/7) = P(Z > -1.26) = .896$$

The Distribution of a Linear Combination

- **Suppose X_1 and X_2 are independent normally distributed measurements.**
- **Set $Y = a_1X_1 + a_2X_2$**
- **$\Rightarrow Y$ is normally distributed with**
  * $E[Y] = a_1E[X_1] + a_2E[X_2]$
  * $\textbf{Var}(Y) = a_1^2\textbf{Var}(X_1) + a_2^2\textbf{Var}(X_2)$

The Distribution of a Linear Combination

Shaft in Sleeve Example.

* **Let X denote the cross-sectional diameter of a steel rod**
* **Let Y denote the cross-sectional diameter of a hollow cylinder.**
* **Suppose X is normally distributed with expected value 15 mm and variance 3 mm**
* **Suppose Y is normally distributed with expected value 16 mm and variance 2 mm.**
* **What is the probability that a randomly selected steel rod will fit into the cylinder?**

$$P(Y - X > 0) = ?$$

The Distribution of a Linear Combination

Example (cont'd). $Y - X$ **has a normal distribution with expected value**

$$E[Y - X] = E[Y] - E[X] = 16 - 15 = 1$$

and variance

$$\textbf{Var}(Y - X) = \textbf{Var}(Y) + \textbf{Var}(X) = 5$$

$a_1 = 1$ **and** $a_2 = -1$**.**

$$P(Y - X > 0) = P(Z > -.45) = .674$$

The Distribution of a Linear Combination of n Random Variables

If $Y = a_1 X_1 + \cdots + a_m X_m$, then

* $E[Y] = \sum_{i=1}^{m} a_i E[X_i]$ **and**
* **Var**$(Y) = \sum_{i=1}^{m} a_i^2$**Var**$(X_i)$

If the X's are independent normal random variables, then Y will be normally distributed.

The Distribution of a Linear Combination

Example. **When manufacturing a certain component, 3 different machining operations are required.**

* **Each machining time is normally distributed and is independent of the other times.**
* **The expected machining times are 15, 30 and 20 minutes, resp.**
* **The standard deviations are 1, 2, and 1.5, resp.**
* **The cost of using machine 1 is 2 dollars per minute.**
* **Machine 2 costs 3 dollars per minute.**
* **Machine 3 costs 4 dollars per minute.**
* **Find the probability that the machining cost of producing one component is more than 220 dollars.**

$$P(2X_1 + 3X_2 + 4X_3 > 220) = P(Z > 2.29) = 0.011$$

An autoregressive time series model

Suppose

- ε_1 **and** Z_0 **are independent normal random variables**
- ϕ_1 **is a constant**
- **the expected value of** ε_1 **is 0, and its variance is** $\sigma_\varepsilon^2 > 0$**.**
- $\mu_{Z_0} = E[Z_0]$ **and** $\sigma_{Z_0}^2 = \mathsf{Var}(Z_0)$ **and**

$$Z_1 = \phi_1 Z_0 + \varepsilon_1.$$

Then Z_1 **is normally distributed, and has mean** $\phi_1 \mu_{Z_0}$ **and variance** $\sigma_{Z_0}^2 \phi_1^2 + \sigma_\varepsilon^2$**.**

We now want to find conditions on ϕ_1 and μ_{Z_0} so that the distribution of Z_1 will be exactly the same as the distribution of Z_0.

This kind of stationarity condition is often useful in modelling of processes that occur in time (or in space, for that matter).

$$E[Z_0] = \mu_{Z_0} = E[Z_1] = \phi_1 \mu_{Z_0}$$

implies that either $\phi_1 = 1$ or $\mu_{Z_0} = 0$.

$$V(Z_0) = \sigma^2_{Z_0} = V(Z_1) = \sigma^2_\varepsilon + \phi_1^2 \sigma^2_{Z_0}.$$

If $\phi_1 = 1$, then $\sigma_\varepsilon = 0$, which is not possible. Therefore, $\phi_1 \neq 1$. This means $\mu_{Z_0} = 0$.

But we also have

$$\sigma^2_{Z_0} = \sigma^2_\varepsilon + \sigma^2_{Z_0}\phi_1^2$$

so that

$$\sigma^2_{Z_0}(1 - \phi_1^2) = \sigma^2_\varepsilon$$

which implies that $\phi_1^2 < 1$**, and**

$$\sigma^2_{Z_0} = \frac{\sigma^2_\varepsilon}{1 - \phi_1^2}.$$

Summarizing the results of the example, we observe that if Z_0 has a normal distribution with mean 0 and variance $\frac{\sigma_\varepsilon^2}{1-\phi_1^2}$, independent of ε_1 which also has a normal distribution with mean 0 and variance σ_ε^2, then

$$Z_1 = \phi_1 Z_0 + \varepsilon_1$$

has the same normal distribution as Z_0.

An autoregressive time series model

Now, let ε_2 be another normal random variable, independent of ε_1

but with the same mean and variance as ε_1. Then

$$Z_2 = \phi_1 Z_1 + \varepsilon_2$$

must have the same distribution as Z_1.

In fact, for $n = 2, 3, \ldots$,

$$Z_n = \phi_1 Z_{n-1} + \varepsilon_n$$

defines a sequence of normal random variables all having mean 0 and variance $\frac{\sigma_\varepsilon^2}{1-\phi_1^2}$, when $\phi_1^2 < 1$ and the ε's are independent of each other.

The Z's define an autoregressive model of order 1: AR(1).

Simulating from an AR(1) Model

Suppose $\phi = -.7$ and $\sigma_\varepsilon^2 = 0.1$. Then $\sigma_{Z_0}^2 = \frac{0.1}{1-(-.7)^2} = 0.196$.

Therefore, we can simulate Z_0 as a normal random variable with mean 0 and variance 0.196:

```
Z0 <- rnorm(1, sd = sqrt(.196))
Z0
```

```
## [1] -0.2440506
```

With Z_0 in hand, we can simulate Z_1 as $\phi_1 Z_0 + \varepsilon_1$,

```
phi1 <- -0.7
eps1 <- rnorm(1, sd = sqrt(.1))
Z1 <- phi1*Z0 + eps1
Z1
```

```
## [1] 0.2669045
```

Simulating from an AR(1) Model

With Z_1 in hand, we can simulate Z_2 as $\phi_1 Z_1 + \varepsilon_2$:

```
eps2 <- rnorm(1, sd = sqrt(.1))
Z2 <- phi1*Z1 + eps2
Z2


## [1] -0.7951741
```

and Z_3

```
eps3 <- rnorm(1, sd = sqrt(.1))
Z3 <- phi1*Z2 + eps3
Z3


## [1] 0.3895803
```

and so on ... for this kind of calculation, we cannot avoid the use of a `for()` loop.

Simulating from an AR(1) Model

Set up a vector which has enough elements to store the result

```
n <- 20 # we will simulate 20 values
Z <- numeric(n)
```

Generate enough normal errors, ε with mean 0 and variance σ_ε^2:

```
eps <- rnorm(n, sd = sqrt(.1))
```

Using the `for()` loop for $i = 1, \ldots, n$, repeatedly overwrite Z_0 with the value of $\phi_1 Z_0 + \varepsilon$, and store the result in successive entries of Z:

```
for (i in 1:n) {
    Z0 <- phi1*Z0 + eps[i]
    Z[i] <- Z0
}
Z # the results are stored here

##  [1] -0.18275726  0.77482530 -0.30308626  0.40287202
##  [5] -0.69237779  0.97410962  0.29612378 -0.01663420
##  [9]  0.08575178 -0.06866682  0.22158700 -1.34430271
## [13]  0.75853027 -0.82279048  0.88179616 -0.57839665
## [17]  1.09481349 -0.35167708  0.58186923 -0.04699310
```

Simulating from the AR(1) Model

A better way to simulate from an AR model is to use the `arima.sim()` function.

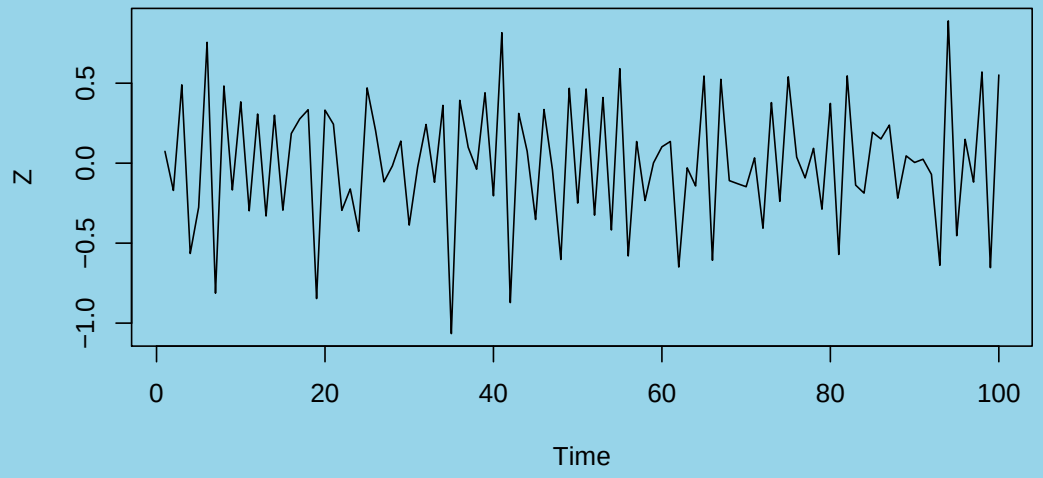
In the following, we simulate a time series of 100 observations from the AR(1) process with $\phi = -0.7$:

```
Z <- arima.sim(100, model = list(ar = phi1), sd = sqrt(0.1))
```

The Trace Plot for Time Series Data

```
ts.plot(Z)
```

The trace plot is just a graph of the time series measurements against time.

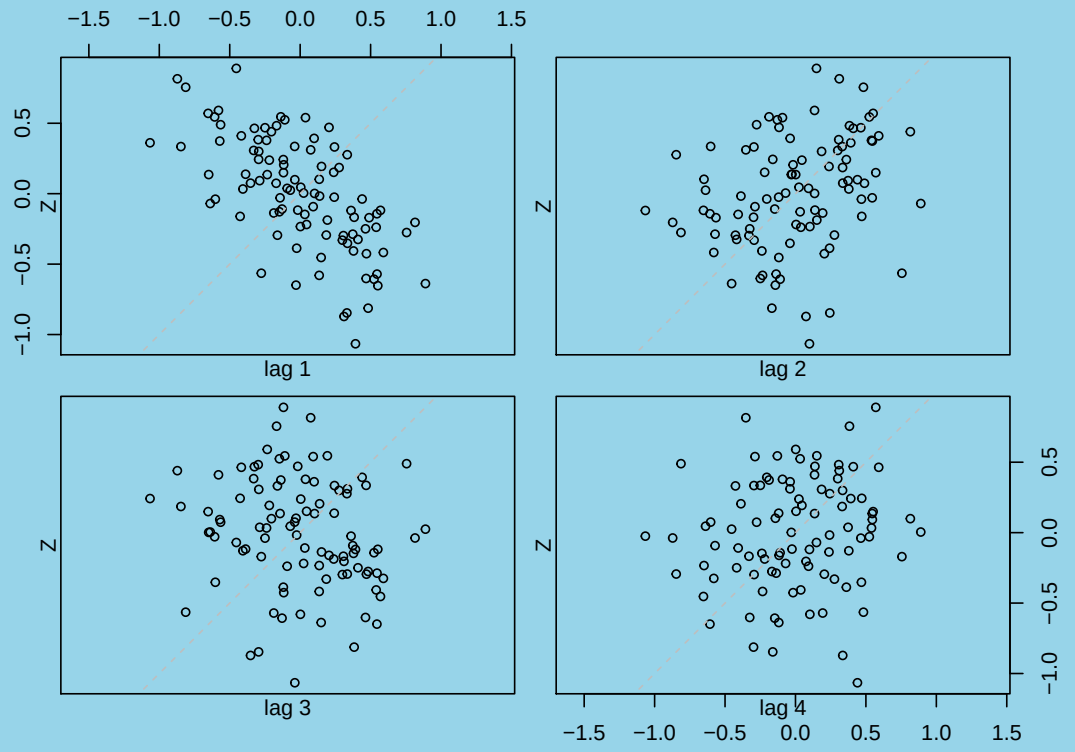


Observe that the observations are not completely random. They appear to oscillate: a large positive value is often followed by a large negative value.

The Lag Plot for Time Series Data

The lag plot is just a graph of the successive time series measurements against values at previous lags.

```
lag.plot(Z, lags=4,
    do.lines=FALSE)
```



Here, we are looking at the first 4 lag plots. Note how the relation at lag 1 is negative, and positive at lag 2, and negative again at lag 3, and so on.