Multivariate Modelling and Simulation I

COSC/DATA 405/505

Multivariate Models: Handling More than One Measurement

- Joint distributions
- Expectation; Expected Value of Sums and Averages
- Covariance and Correlation (theoretical and empirical)
- Marginal pdfs
- Conditional pdfs
- Independence, mathematically and graphically
- Independence versus Correlation

- **Variances of Sums and Linear Combinations**
- **Central Limit Theorem**
- **Simple Linear Regression**

- **A single measurement is modelled using a continuous random variable X with probability density function $f(x)$, e.g. normal, exponential, etc.**

- **What if there are 2 or more measurements?**

- **Model each measurement with a random variable: $X_1, X_2, \ldots,$**

- **How do we evaluate $P(a < X_1 < b, c < X_2 < d)$?**

- **A basic probability result says that if random events A and B are independent, then** $P(A \text{ occurs and } B \text{ occurs}) = P(A \text{ occurs})P(B \text{ occurs})$**.**

- **This allows us to say that if the events $\{a < X_1 < b\}$ and $\{c < X_2 < d\}$ are independent, then we can write**

$$P(a < X_1 < b, c < X_2 < d) =$$

$$P(a < X_1 < b)P(c < X_2 < d)$$

$$= \int_a^b f_1(y_1)dy_1 \int_c^d f_2(y_2)dy_2$$

$$= \int_a^b \int_c^d f_1(y_1)f_2(y_2)dy_1dy_2$$

where $f_1(y_1)$ and $f_2(y_2)$ are the density functions for X_1 and X_2.

A Model for Two Independent Measurements

- **Set**

$$f(y_1, y_2) = f_1(y_1) f_2(y_2).$$

- $f(y_1, y_2)$ **is called the joint density function for** X_1 **and** X_2**.**

$$P(a < X_1 < b, c < X_2 < d) =$$

$$\int_a^b \int_c^d f(y_1, y_2) dy_1 dy_2$$

- **If the measurements** X_1 **and** X_2 **are not independent, we can still define** $f(y_1, y_2)$**, the joint density function for** X_1 **and** X_2**.**

- **Properties:**

  1. $f(y_1, y_2) \geq 0$

  2. $\int \int f(y_1, y_2) dy_1 dy_2 = 1$

- **Computation of probabilities:**

$$P(a < X_1 < b, c < X_2 < d) =$$

$$\int_a^b \int_c^d f(y_1, y_2) dy_1 dy_2$$

- **Joint density for more than two measurements ($X_1, X_2, \ldots, X_k$):**

$$f(y_1, y_2, \ldots, y_k)$$

Example.

A machine is used to automatically fill cylinders with propane gas.

* X **= amount of propane in a randomly selected cylinder (moles)**
* T **= temperature (C) at the time of filling**
* **model for X and T – joint density:**

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

$f(x,t) = 0$ **for other values of x and t**

Verify that $f(x,t)$ is a valid joint density function.

Example (cont'd).

We verify the two properties of the joint density function:

1. $f(x,t) \geq 0$ **for all** x, t

2. $\int \int f(x,t)dtdx = 1$:

$$\int_{10}^{11} \int_{15}^{20} \frac{x + t/5 - 13}{5} dtdx =$$

$$\int_{10}^{11} \frac{xt + t^2/10 - 13t}{5} \Big|_{15}^{20} dx =$$

$$\int_{10}^{11} (x - 9.5) dx = 1$$

Joint Probability Density Function

Example (cont'd).

Calculate the probability that the amount of propane is between 10.4 and 10.6 moles, and that the temperature is between 16 and 17 degrees.

$$P(10.4 < X < 10.6, 16 < T < 17) =$$

$$\int_{10.4}^{10.6} \int_{16}^{17} \frac{x + t/5 - 13}{5} dt dx = \int_{10.4}^{10.6} \frac{x - 9.7}{5} dx = .032$$

Example 2.

* **Suppose the reliability of an electric motor depends upon two critical components, having lifetimes X and Y which can be modelled with the joint probability density function**

$$f(x,y) = \begin{cases} xe^{-x(1+y)}, & x \geq 0, y \geq 0 \\ 0, & \textbf{otherwise} \end{cases}$$

Find the probability that both components last more than 1 unit of time.

Example 2 (cont'd).

The question asks for the probability that X is greater than 1 and Y is greater than 1:

$$P(X > 1, Y > 1) =$$

$$\int_1^\infty \int_1^\infty xe^{-x(1+y)} dydx = e^{-2}/2$$

- **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

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

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

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 ....

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

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

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:

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

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.

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.

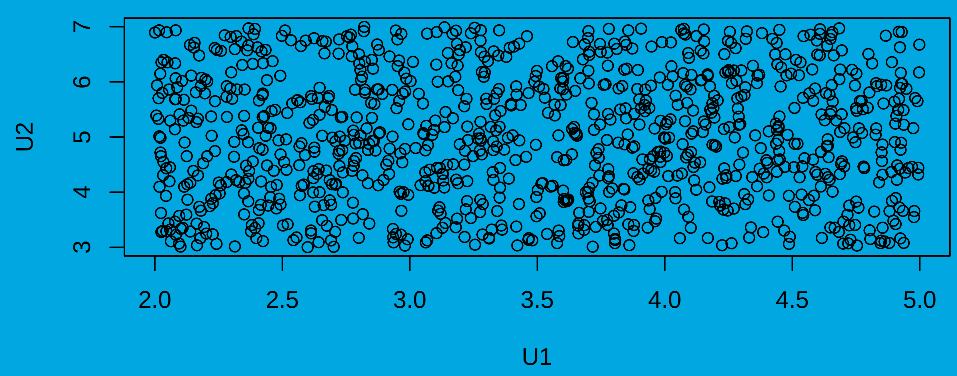
The `cor()` function calculates this quantity.

Example.

```
## [1] 0.02404286
```

Simulated Pairs of Independent Uniforms:

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

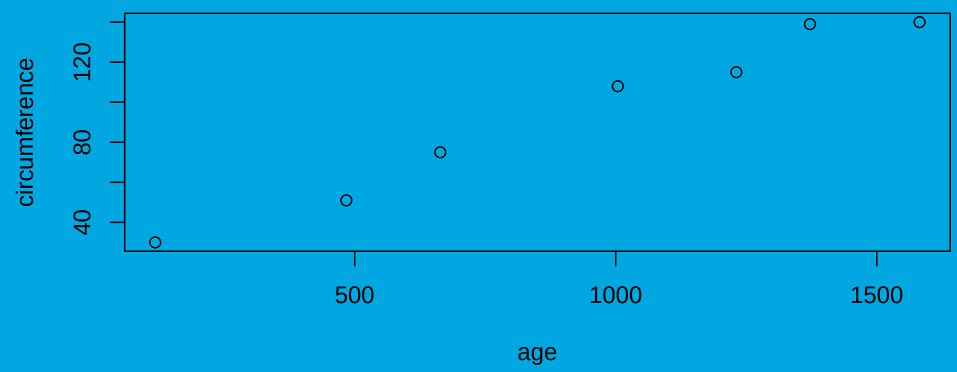


The correlation is small, close to 0, and the scatterplot shows no pattern.

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.

When we have more than one measurement, the joint density function summarizes the overall model. Each individual random variable still has a probability density function (as discussed earlier), but when in this larger context, the pdf is referred to as a marginal pdf.

- **If X_1 and X_2 have joint density function $f(y_1, y_2)$,**
  - **the density function for X_1 can be determined as**

$$f_1(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_2$$

That is, the pdf of X_1 is obtained by integrating over all possible values of the other variable.

  - **the density function for X_2 can be determined as**

$$f_2(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_1$$

That is, the pdf of X_2 is obtained by integrating over all possible values of the other variable.

Example.

Find the marginal densities of X and T for the propane example:

$$f_X(x) = \int_{15}^{20} \frac{x + \frac{t}{5} - 13}{5} dt = x - 9.5, \quad (x \in [10, 11])$$

$$f_T(t) = \int_{10}^{11} \frac{x + \frac{t}{5} - 13}{5} dx = \frac{t}{25} - \frac{1}{2}, \quad (t \in [15, 20])$$

Example. **Find the marginal density of X for the reliability example:**

$$f_1(x) = \int_0^\infty x e^{-x(1+y)} dy = e^{-x}, \quad x \geq 0$$

Find the marginal density function for Y:

$$f_2(y) = \int_0^\infty x e^{-x(1+y)} dx = \frac{1}{1+y^2} \quad y \geq 0$$

by integrating by parts.

Observe that X does not have the same density function as Y, and we say that X and Y are not identically distributed.

If X and Y had the same marginal distributions, we would say they are identically distributed.

Exercise.

Suppose the joint density function of X_1 and X_2 is

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

Find the marginal density function for X_1 by integrating over all possible values of x_2 (i.e. $x_2 > 0$).

Ans.

$$f_{X_1}(x_1) = \int_0^\infty \frac{\lambda}{x_1} e^{-\lambda x_1 - x_2/x_1} dx_2 = \lambda e^{-\lambda x_1}, \quad x_1 \geq 0$$

and 0, otherwise. We recognize this as the exponential density function.

Exercise 2.

Try to obtain the marginal density function for X_2.

Ans.

$$f_{X_2}(x_2) = \int_0^\infty \frac{\lambda}{x_1} e^{-\lambda x_1 - x_2/x_1} dx_1$$

which can only be evaluated numerically. Monte Carlo integration, anyone?

To summarize:

When in the context of several random variables, the marginal probability density function of a single one of the random variables can be obtained by integrating the joint density function over all possible values of all of the random variables apart from the one of interest.

Suppose we know the value of X_1, and we would like to predict the value of X_2, using this information.

The joint probability density function tells us how X_1 and X_2 are related, so we might think that $f(x_1, x_2)$ is the probability density function of X_2 for each given value of X_1, but this would be an incorrect interpretation.

This is because

$$\int_{-\infty}^{\infty} f(x_1, x_2) dx_2 = f_{X_1}(x_1)$$

If $f(x_1, x_2)$ were a probability density function for X_2, given X_1, the above integral should evaluate to 1.

Conditional Density Functions

However, if we divide $f(x_1, x_2)$ by $f_{X_1}(x_1)$, we obtain a function of x_2 which integrates to 1:

$$\int_{-\infty}^{\infty} \frac{f(x_1, x_2)}{f_{X_1}(x_1)} dx_2 = \frac{f_{X_1}(x_1)}{f_{X_1}(x_1)} = 1.$$

It turns out that this gives a very useful predictive density function for X_2, given knowledge of X_1.

We write

$$f_{X_2|X_1}(x_1, x_2) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)}$$

as the *conditional density function* of X_2 given X_1. This density function predicts the probability density of X_2, when we know the value of X_1.

Conditional Density Functions

Similar reasoning tells us that the predictive probability density of X_1 or its conditional density function is given by

$$f_{X_1|X_2}(x_1, x_2) = \frac{f(x_1, x_2)}{f_{X_2}(x_2)}$$

This completely summarizes how knowledge of X_2 can help us make predictions about X_1.

Conditional Density Functions

Example. **Suppose X_1 and X_2 have the joint pdf**

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

and we know the value of X_1 as x_1, say 10, or 3, etc. Earlier, we found that

$$f_{X_1}(x_1) = \lambda e^{-\lambda x_1}, \quad x_1 \geq 0.$$

Then the conditional density for X_2, given X_1

$$f_{X_2|X_1}(x_1, x_2) = \frac{1}{x_1} e^{-x_2/x_1}, \qquad x_2 \geq 0.$$

This is an exponential density function, but now the rate is $1/x_1$. e.g. If we know that $x_1 = 10$, then the expected value of X_2 would be 10, and if $x_1 = 3$, the expected value of X_2 is 3.

The conditional density function of X_2, given X_1, is not the same as the marginal density of X_2. Thus, X_1 gives predictive information about X_2. The two random variables are not independent.

We have spoken earlier of cases where X_1 and X_2 are independent. In such cases, X_1 provided no predictive information about X_2; technically, this means that the conditional density function of X_2 given X_1 is identical to the marginal density function of X_2:

$$f_{X_2|X_1}(x_1, x_2) = f_{X_2}(x_2).$$

Multiplying both sides of this by $f_{X_1}(x_1)$ gives the joint density function on the left and the product of the marginal density function on the right. Random variables are independent if their joint distribution can be factored in this way. That is, X_1 and X_2 are independent, if their joint density is

$$f(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) \tag{1}$$

where the two functions are the respective marginal densities of X_1 and X_2.

Example.

Consider the propane example.

$$f(x,t) = \frac{x + \frac{t}{5} - 13}{5}$$

$$f_X(x) f_T(t) = (x - 9.5)\left(\frac{t}{25} - \frac{1}{2}\right)$$

$\Rightarrow$

$$f(x,t) \neq f_X(x) f_T(t)$$

so X and T are not independent.

Exercise. **For the reliability problem, show that X and Y are not independent.**

Example. **An electronic surveillance system has one of each of two types of components in joint operation. The joint density function of the lifelengths X_1 and X_2 of the two components is**

$$f(y_1, y_2) = (1/8) y_1 e^{-\frac{1}{2}(y_1+y_2)}$$

for $y_1 > 0$ and $y_2 > 0$, and it is 0, otherwise.

Are X_1 and X_2 independent?

$$f_1(y_1) = \int_0^\infty \frac{1}{8} y_1 e^{-\frac{1}{2}(y_1+y_2)} dy_2 = \frac{1}{4} y_1 e^{-\frac{1}{2}y_1}, \qquad y_1 > 0$$

and

$$f_2(y_2) = \frac{1}{2} e^{-\frac{1}{2}y_2}, \qquad y_2 > 0$$

$\Rightarrow$

$$f(y_1, y_2) = f_1(y_1) f_2(y_2)$$

so X_1 and X_2 are independent.

Example.

- **The time X_1 until failure of a fuel pump in an internal combustion engine can be modelled as a normal random variable with expected value 2000 hours and standard deviation 400 hours.**
- **The lifetime X_2 of a timing belt can be modelled as an exponential random variable with expected value 2800 hours.**
- **Supposing that these parts operate independently, find the probability that both fail before 1000 hours of operation.**

$$P(X_1 < 1000, X_2 < 1000) = P(X_1 < 1000)P(X_2 < 1000)$$

$$= P(Z < -2.5)(1 - e^{-.357}) = .0062(.300) = .0019$$

Here $Z = \frac{X_1 - 2000}{400}$ is a standard normal random variable. The `pnorm` function can be used to determine that $P(Z < -2.5) = .0062$.

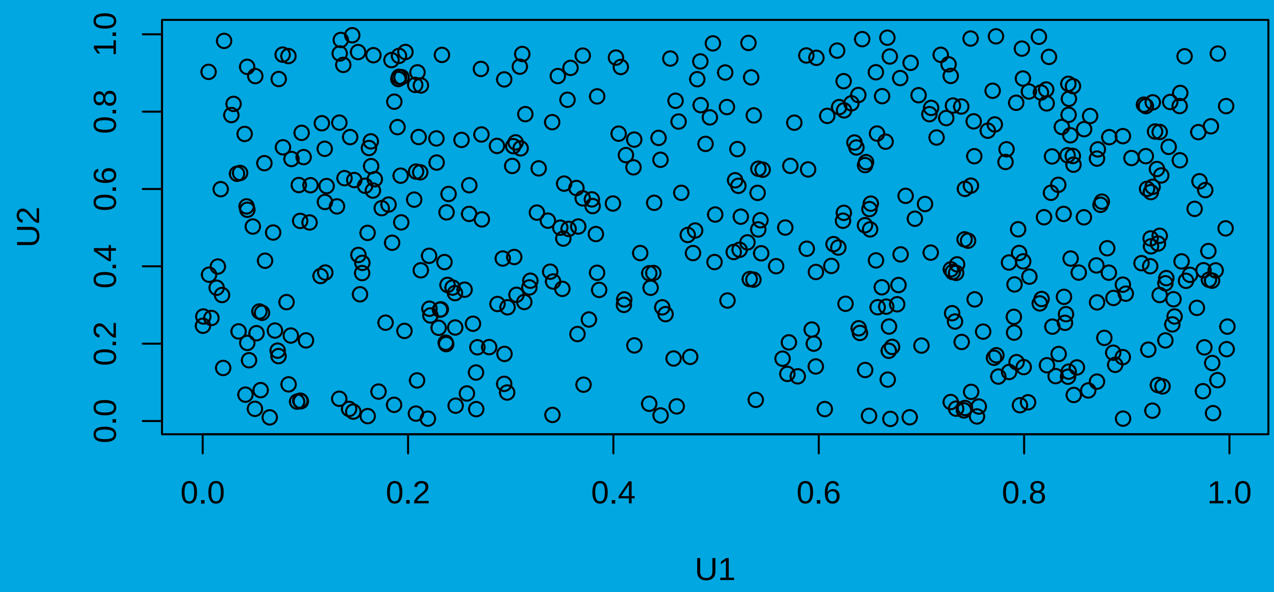
We can get an intuitive feel for the independence modelling assumption (1) by graphing some simulated random variables, both independent and for some forms of dependence.

First, let's consider two independent uniform random variables which take values in the interval $[0, 1]$.

The next figure displays a scatterplot of 500 values taken from the distributions of U_2 and U_1, where are both uniformly distributed.

Graphical Views of Independence and Dependence

```
U1 <- runif(500)
U2 <- runif(500)
plot(U2 ~ U1)
```

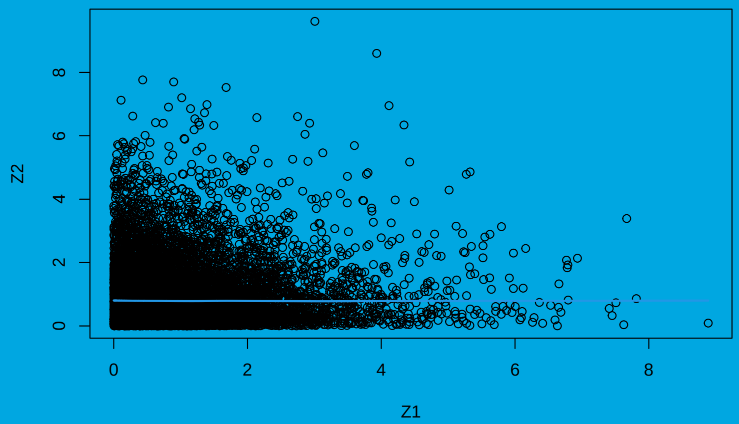


There is no structure to the patterns that might be discerned from this picture. This is the clearest possible illustration of variables which have no relation.

Independence manifests itself in other ways.

Next, we consider exponential random variables, Z_1 and Z_2. In both cases, we will sample 10000 points from their respective distributions and look at a scatterplot of the corresponding pairs of data points.

```
Z1 <- rexp(10000)
Z2 <- rexp(10000)
plot(Z2 ~ Z1)
lines(lowess(Z1, Z2),
      col=4, lwd=2)
```



What you should observe in the figure is that it is not possible to predict the value of Z_2 from knowledge of Z_1.

This is a random collection of points, even though it might appear that there is a pattern (points are bunched up towards the lower left corner of the plot).

What characterizes independence is that neither variable gives predictive information about the other variable.

A Case of Dependence

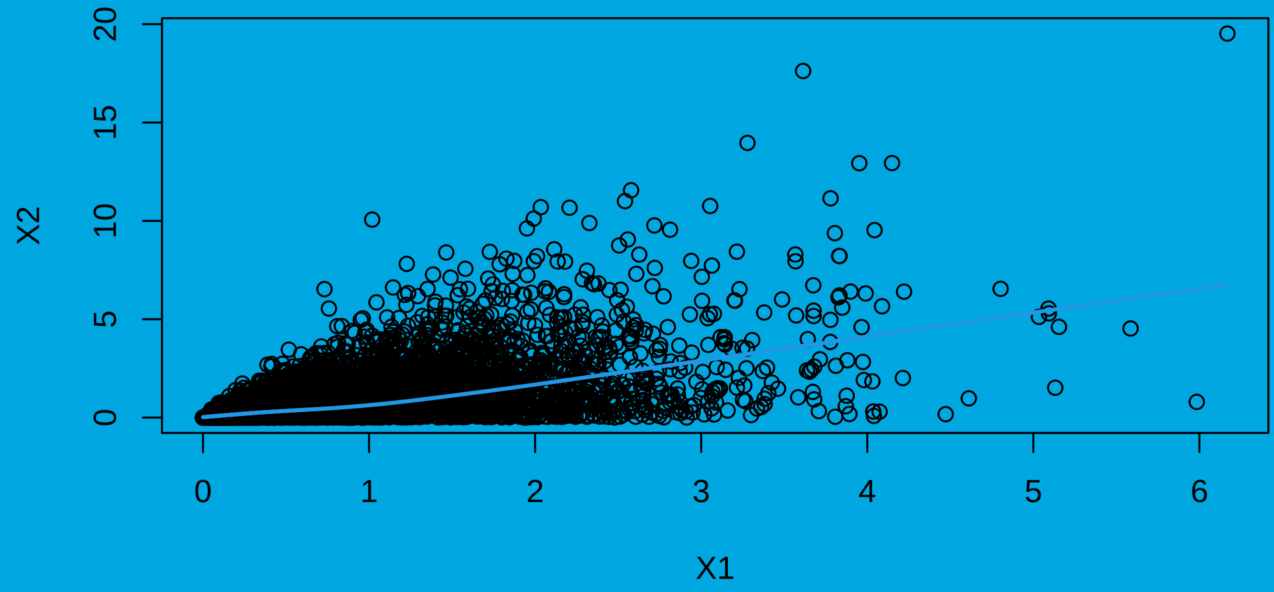
Let's change the story a bit now.

Suppose X_1 and X_2 are related. In particular, suppose X_1 is exponentially distributed with rate $\lambda = 1.5$ and X_2 is exponential with rate $1/X_1$.

Next, we will apply our simulation-based graphical analysis.

Graphical Views of Independence and Dependence

```r
X1 <- rexp(10000, rate = 1.5)
X2 <- rexp(10000, rate = 1/X1)
plot(X2 ~ X1)
lines(lowess(X1, X2), col=4, lwd=2)
```



Correlation versus Dependence

What is the difference between correlation and dependence?

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

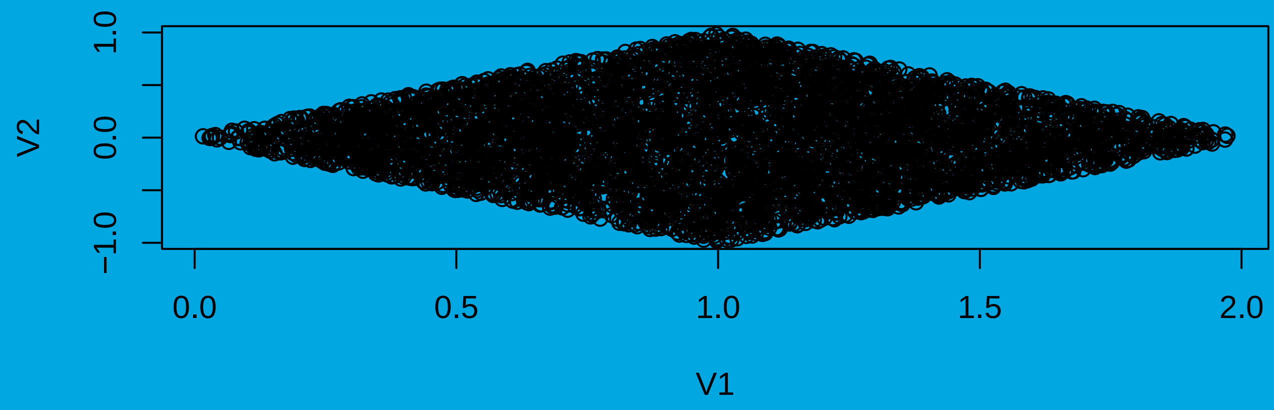
2 random variables may be dependent, but not correlated.

Correlation versus Dependence

Example.

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


## [1] -0.007104551
```



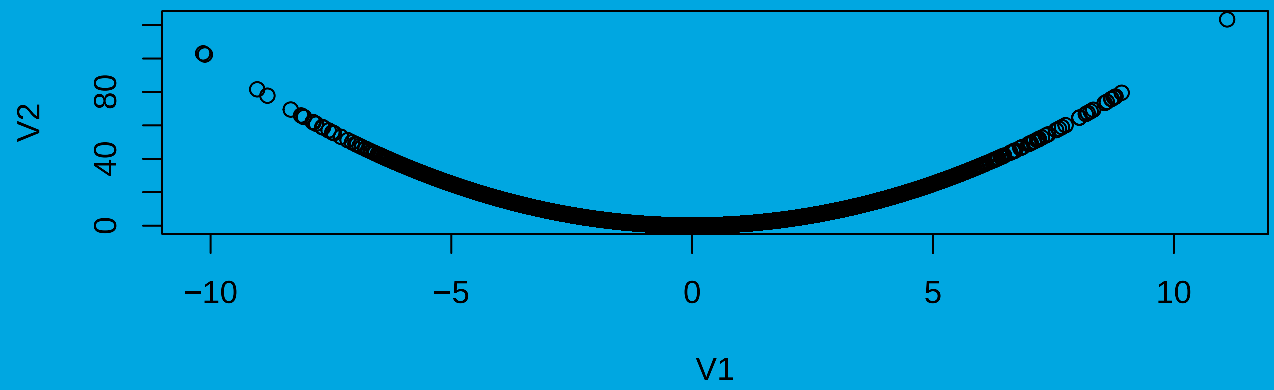
Are V_1 and V_2 dependent?

Are V_1 and V_2 correlated?

Example 2

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

## [1] 0.003538419
```



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

$$\mathbf{E}[X_1X_2] = \int\int y_1y_2f_1(y_1)f_2(y_2)dy_1dy_2 = \mathbf{E}[X_1]\mathbf{E}[X_2]$$

and

$$\mathbf{E}[(X_1+X_2)^2] = \mathbf{E}[X_1^2] + 2\mathbf{E}[X_1X_2] + \mathbf{E}[X_2^2]$$

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 Random Variables

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
```

- 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**
  * $E[Y] = a_1E[X_1] + a_2E[X_2]$
  * **Var**$(Y) = a_1^2$**Var**$(X_1) + a_2^2$**Var**(X_2)

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?**

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

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** $a_2 = -1$**.**

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

If $Y = a_1X_1 + \cdots + a_mX_m$, then

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

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

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.**

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

Suppose X and Y are random variables which have a joint density function given by

$$f(x, y) = \frac{e^{-(y-\beta_0-\beta_1 x)^2/(2\sigma^2)-x^2/2}}{2\pi\sigma}.$$

This is an example of a bivariate normal pdf: X is normal with mean 0, and Y is normal with mean $\beta_0 + \beta_1 x$. In other words, the mean of Y is now a linear function of x.

β_0 and β_1 are unknown intercept and slope parameters.

If we want to predict Y from X, we should use the conditional density function $f_{Y|X}(x, y)$. We can obtain that density function in 2 steps:

1. Find $f_X(x)$ by integrating over all y.
2. Divide $f(x, y)/f_X(x)$.

1.

$$f_X(x) = \int f(x,y)dy = \frac{\mathbf{e}^{-x^2/2}}{\sqrt{2\pi}}.$$

We have used the fact that

$$\frac{\mathbf{e}^{-(y-\beta_0-\beta_1 x)^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma}$$

is a normal pdf and integrates to 1.

2. Dividing $f(x,y)$ by $f_X(x)$ gives

$$f_{Y|X}(x,y) = \frac{\mathbf{e}^{-(y-\beta_0-\beta_1 x)^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma}.$$

The conditional distribution that we just obtained is a normal pdf with mean

$$\beta_0 + \beta_1 x$$

and variance σ^2.

The mean is an expected value (called the *conditional expectation*) and has the notation

$$E[Y|X = x] = \beta_0 + \beta_1 x. \tag{2}$$

The conditional expectation of Y, given $X = x$ is also referred to as the regression function, a function of x.

The variance is actually a *conditional variance*:

$$\textbf{Var}(Y|X = x) = \sigma^2. \tag{3}$$

This is often referred to as the noise variance.

Simple Regression

The regression function at (2) and the variance function at (3), which is just the constant function, tell us that, given $X = x$, we could view Y as the random variable

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

where ε is the noise random variable - a normal random variable with mean 0 and variance σ^2. The $\beta_0 + \beta_1 x$ terms are not random.

This is the usual form of the simple linear regression model which relates Y to x in the presence of noise.

We will return to this model in a later lecture.