Exercise Set 5-1:

1) a) If X is a Bernoulli random variable, then the mass function is $f_X(x) = \mathrm{P}(X = x) = p^x(1-p)^{1-x}$ for $x \in \{0,1\}$. Because there are only two possible values of X, summing is easy. The expectation is

$$\mathrm{E}(X) = \sum_{x=0}^{1} x f_X(x) = \sum_{x=0}^{1} x p^x(1-p)^{1-x} = 0p^0(1-p)^1 + 1p^1(1-p)^0 = p.$$

b) This is one example where the linearity of expectation comes in handy. If X is a binomial random variable, then $f_X(x) = \mathrm{P}(X = x) = \binom{n}{x} p^x(1-p)^{n-x}$ for $x \in \{0,1,2,\dots,n\}$. This means that the expectation is

$$\mathrm{E}(X) = \sum_{x=0}^{n} x f_X(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x(1-p)^{n-x}.$$

It is possible to compute this directly, but there is an easier way. Recall that a binomial random variable is the number of successes out of n independent trials each with probability of success p. We already know that a Bernoulli random variable can model a single trial with probability of success p. Thus, we can re-imagine the binomial random variable as the sum of n independent Bernoulli random variables, which we can label $X_1, X_2, \dots, X_n$. So if $X = \sum_{i=1}^{n} X_i$, where the X_i are independent draws from a Bernoulli distribution with success probability p, then X is distributed as a binomial random variable with parameters n and p. We can now write

$$\mathrm{E}(X) = \mathrm{E}\left(\sum_{i=1}^{n} X_i\right).$$

And here is where linearity helps. Because the expectation is linear, the expectation of the sum is equal to the sum of the expectations. This lets us quickly finish the job using the result from part (a):

$$\mathrm{E}(X) = \mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i) = \sum_{i=1}^{n} p = np.$$

c) Let X be a discrete uniform random variable. Then the mass function is $f_X(x) = \mathrm{P}(X = x) = \frac{1}{b-a+1}$ for all for $x \in \{a, a+1, a+2, \dots, b\}$. (See Table 4-5). The expectation is

$$\mathrm{E}(X) = \sum_{x=a}^{b} x f_X(x) = \sum_{x=a}^{b} \frac{x}{b-a+1} = \frac{1}{b-a+1} \sum_{x=a}^{b} x.$$

Where the last step, moving $\frac{1}{b-a+1}$ outside the sum, is allowed because the sum is with respect to x and $\frac{1}{b-a+1}$ does not depend on x. Using the formula given in the hint to the problem,

$$\mathrm{E}(X) = \frac{1}{b-a+1}\sum_{x=a}^{b} x = \frac{(a+b)(b-a+1)/2}{b-a+1} = \frac{a+b}{2}.$$

d) If X is a continuous uniform random variable, then the density is $f_X(x) = \frac{1}{b-a}$ for $x \in [a,b]$. The expectation is

$$\mathrm{E}(X) = \int_a^b x f_X(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a}\int_a^b x dx.$$

Remembering that the integral $\int x\, dx = \frac{x^2}{2} + C$ and evaluating at a and b gives

$$\mathrm{E}(X) = \frac{(b^2 - a^2)/2}{b-a}.$$

The hint given in the problem lets us write this as

$$\mathrm{E}(X) = \frac{(b-a)(b+a)/2}{b-a},$$

which simplifies to

$$\mathrm{E}(X) = \frac{(a+b)}{2}.$$

2) a) When `n = 1`, the histogram ought to resemble the density function for a normal random variable with expectation 0 and standard deviation 1—we are just taking individual samples from a normal distribution and plotting them. That is to say it ought to be symmetric, centered around 0, and roughly bell-shaped. The great majority of the data should fall between -2 and 2 in this case. As we draw larger samples and take their means, the law of large numbers suggests that the sample means should generally be closer to the expectation than the individual observations are. Indeed, as we increase the sample size, we find that fewer samples have means that are far from the expectation. The shape of the distribution continues to look roughly normal, but the spread decreases.

b) Here is a modified version of the code that uses `rexp()` to simulate exponential random variables:

```
samp.size <- 20
n.samps <- 1000
```

```
samps <-matrix(rexp(samp.size*n.samps, rate = 1), ncol =
n.samps)
samp.means <- colMeans(samps)
hist(samp.means)
```

The shape of the exponential density is markedly different from the shape of the normal distribution. No observations smaller than 0 are allowed. When the expected value is set to 1, most of the observations are near 0, and the observations trail off far to the right. We say the distribution is “skewed right.” Again, when we take means of samples, we find that the sample means get closer to the expectation as the sample size increases. You ought to notice something odd here, though. The sample means cluster more tightly around the expectation as the sample size grows, but the shape of the distribution also changes. Namely, it starts to look more symmetric and bell-like—more normal. This is a preview of the central limit theorem, which will appear soon.

Exercise Set 5-2:

1) Our definition says that $\mathrm{Var}(X) = \mathrm{E}([X - \mathrm{E}(X)]^2)$. We expand the expression to get:

$$\mathrm{Var}(X) = \mathrm{E}(X^2 - 2X\mathrm{E}(X) + [\mathrm{E}(X)]^2).$$

Applying the linearity of expectation,

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(2X\mathrm{E}(X)) + \mathrm{E}([\mathrm{E}(X)]^2).$$

$\mathrm{E}(X)$ is a constant, as is $[\mathrm{E}(X)]^2$. The expectation of a constant is just the constant itself, so we have:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - 2\mathrm{E}(X)\mathrm{E}(X) + [\mathrm{E}(X)]^2.$$

Finally, we collect the $[\mathrm{E}(X)]^2$ terms to get

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.$$

2) a) This is one of the rare situations where it is easier to start with the definition of the variance in equation 5.6 than with the identity in equation 5.7. Using the definition, we have

$$\mathrm{Var}(X + c) = \mathrm{E}([X + c - \mathrm{E}(X + c)]^2).$$

Linearity of expectation (equation 5.4) lets us write this as

$$\mathrm{Var}(X + c) = \mathrm{E}([X + c - \mathrm{E}(X) - c]^2),$$

which is

$$\mathrm{Var}(X + c) = \mathrm{E}([X - \mathrm{E}(X)]^2) = \mathrm{Var}(X).$$

Thus, adding a constant to a random variable does not change the variance of the random variable.

b) This time, we'll start with equation 5.7, which gives

$$\mathrm{Var}(cX) = \mathrm{E}([cX]^2) - [\mathrm{E}(cX)]^2.$$

Linearity of expectation lets us pull the constants out of the expectations, which gives

$$\mathrm{Var}(cX) = c^2\mathrm{E}(X^2) - [c\mathrm{E}(X)]^2 = c^2(\mathrm{E}(X^2) - [\mathrm{E}(X)]^2) = c^2\mathrm{Var}(X).$$

3) a) We want $f_{X,Y}(x,y) = \mathrm{P}(X = x \cap Y = y)$. If X and Y are independent, then $\mathrm{P}(X = x \cap Y = y) = \mathrm{P}(X = x)\mathrm{P}(Y = y)$. We already have $\mathrm{P}(X = x)$ and $\mathrm{P}(Y = y)$: these are $f_X(x)$ and $f_Y(y)$, respectively. Thus, if X and Y are independent, then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. That is, the joint mass function of independent discrete random variables is just the product of the *marginal* mass functions of the random variables. (Here, read "marginal" as "ignoring any other random variables.") You can use an analogous argument with cumulative distribution functions to prove a similar claim for continuous random variables.

b) Using equation 5.7,

$$\mathrm{Var}(X + Y) = \mathrm{E}[(X + Y)^2] - [\mathrm{E}(X + Y)]^2.$$

Expanding the first term and invoking the linearity of expectation on the second term gives

$$\mathrm{Var}(X + Y) = \mathrm{E}(X^2 + 2XY + Y^2) - [\mathrm{E}(X) + \mathrm{E}(Y)]^2,$$

Invoking linearity on the first term and expanding the second gives

$$\mathrm{Var}(X + Y) = \mathrm{E}(X^2) + 2\mathrm{E}(XY) + \mathrm{E}(Y^2) - [\mathrm{E}(X)]^2 - 2\mathrm{E}(X)\mathrm{E}(Y) - [\mathrm{E}(Y)]^2.$$

Recognizing $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$, $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - [\mathrm{E}(Y)]^2$, and $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$ lets us finish the proof:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2[\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)] = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

c) Remember that in part (a), we found that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. This means that

$$\mathrm{E}(XY) = \sum_x \sum_y xy\, f_X(x)f_Y(y).$$

Because $xf_X(x)$ is not a function of y and $yf_Y(y)$ is not a function of x, we can rewrite this as

$$\mathrm{E}(XY) = \left[\sum_x x f_X(x)\right]\left[\sum_y y\, f_Y(y)\right] = \mathrm{E}(X)\mathrm{E}(Y).$$

This result means that when X and Y are independent discrete random variables, $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. You can prove the same result for continuous random variables using integrals over the joint density instead of sums over the joint mass function.

d) The trick is to notice that $X - Y = X + cY$, where $c = -1$. If X and Y are independent, then X and cY are also independent. This means that $\mathrm{Var}(X + cY) = \mathrm{Var}(X) + \mathrm{Var}(cY)$. We already proved that $\mathrm{Var}(cY) = c^2\mathrm{Var}(Y)$. If $c = -1$, then $\mathrm{Var}(cY) = \mathrm{Var}(Y)$, and so $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = \mathrm{Var}(X + Y)$ if X and Y are independent.

4) a) Bernoulli random variables only take two values, 0 and 1. Note that because $0^2 = 0$ and $1^2 = 1$, if X is a Bernoulli random variable, then $X = X^2$ in every case, which implies that $\mathrm{E}(X^2) = \mathrm{E}(X) = p$. By equation 5.7, this means that the variance is $\mathrm{Var}(X) = p - p^2 = p(1 - p)$.

b) A binomial random variable with parameters n and p is just the sum of n independent Bernoulli random variables with parameter p. Because the variance of the sum of independent random variables is just the sum of the variances of the random variables, the variance of the binomial random variable is $\mathrm{Var}(X) = np(1-p)$. (This is so much easier than it is to calculate $\mathrm{E}(X^2) = \sum_{x=0}^{n} x^2 \binom{n}{x} p^x (1-p)^{n-x}$.)

5) First, using equation 5.8, we note that

$$\mathrm{Var}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \cdots + X_n)$$

because $1/n$ is a constant. We also know that because the Xs are independent, the variance of the sum is the sum of the individual variances (equation 5.9), which is $n\sigma^2$. Thus, the variance is

$$\mathrm{Var}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

The standard deviation is the square root of the variance:

$$\mathrm{SD}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{\sigma}{\sqrt{n}}$$

This is an important result in statistics. As we add observations to a sample, the sample mean becomes less and less variable, as long as we are taking observations from a distribution with finite variance. The result about the standard deviation of the sample mean shows that our increases in precision per additional sample—when considered in the original units—diminish as the sample size increases. Every time we double the sample size, the standard deviation decreases by a factor of $\sqrt{2} \approx 1.41$.

6) a) We define Z such that $Z = 1$ if $X \geq c$ and $Z = 0$ if $X < c$. X may have any distribution so long as it cannot take negative values. Z is a Bernoulli random variable with parameter $p = \mathrm{P}(X \geq c)$. Because the expectation of a Bernoulli random variable with parameter p is p (Exercise set 5-1, problem 1), we have

$$\mathrm{P}(X \geq c) = \mathrm{E}(Z).$$

If $X \geq c$, then $X/c \geq 1$ and $Z = 1$, so $Z \leq X/c$. Similarly, if $X < c$, then $Z = 0$, and because X is nonnegative and c is positive, once again, $Z \leq X/c$. This means that $Z \leq X/c$ in all cases, so

$$\mathrm{E}(Z) \leq \mathrm{E}(X/c).$$

Using equation 5.4 to pull c out of the expectation and combining gives

$$\mathrm{P}(X \geq c) = \mathrm{E}(Z) \leq \frac{\mathrm{E}(X)}{c},$$

which proves Markov's inequality.

b) $(Y - \mu)^2$ is a non-negative random variable, so Markov's inequality applies to it. $\mathrm{E}[(Y - \mu)^2] = \mathrm{Var}(Y)$ by definition, so Markov's inequality gives us

$$\mathrm{P}((Y - \mu)^2 \geq c) \leq \frac{\mathrm{Var}(Y)}{c}.$$

Now define another constant d as $d = \sqrt{c}$. Notice that the statement $(Y - \mu)^2 \geq c$ is equivalent to $|Y - \mu| \geq d$: whenever one statement is true, the other statement is guaranteed to be true. This means that $\mathrm{P}((Y - \mu)^2 \geq c) = \mathrm{P}(|Y - \mu| \geq d)$. Making this replacement and replacing the c on the right with d^2 gives Chebyshev's inequality:

$$\mathrm{P}(|Y - \mu| \geq d) \leq \frac{\mathrm{Var}(Y)}{d^2}.$$

c) Chebyshev's inequality gets us most of the way there. We need to prove that for positive δ, $\lim_{n\to\infty} \mathrm{P}(|\bar{X}_n - \mu| > \delta) = 0$. Chebyshev's inequality gives

$$\mathrm{P}(|\bar{X}_n - \mu| > \delta) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\delta^2}.$$

Because the expressions on both sides of the inequality are positive, if $\lim_{n\to\infty} \frac{\mathrm{Var}(\bar{X}_n)}{\delta^2} = 0$, then $\lim_{n\to\infty} \mathrm{P}(|\bar{X}_n - \mu| > \delta) = 0$. (This follows from the squeeze theorem.) Thus, we can prove the weak law of large numbers by proving that

$$\lim_{n\to\infty} \frac{\mathrm{Var}(\bar{X}_n)}{\delta^2} = 0.$$

The result of problem 5 lets us write

$$\frac{\mathrm{Var}(\bar{X}_n)}{\delta^2} = \frac{\sigma^2}{n\delta^2}$$

Because σ^2 and δ^2 are finite constants, $(\sigma^2/\delta^2)/n$ approaches 0 as n approaches infinity, which implies that $\mathrm{P}(|\bar{X}_n - \mu| > \delta)$ approaches 0 and proves the weak law of large numbers.

Exercise Set 5-3:

1) a) We have to find $\mathrm{E}(XY)$, $\mathrm{E}(X)$, and $\mathrm{E}(Y)$. We do this by summing over the joint mass function. We can do this with a table like the following:

Outcome	Probability	X	Y	XY
$X = -1, Y = 1$	1/3	-1	1	-1
$X = 1, Y = 1$	1/3	1	1	1
$X = 0, Y = -2$	1/3	0	-2	0
		$\mathrm{E}(X) = 0$	$\mathrm{E}(Y) = 0$	$\mathrm{E}(XY) = 0$

In the bottom row, we compute the expectations by taking an average weighted by the probabilities, which, in this case, are all equal. Thus, $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = 0$.

b) $\mathrm{Cov}(X, Y) = 0$, but X and Y are not independent. For example, $\mathrm{P}(X = 0 \cap Y = -2) = 1/3 \neq \mathrm{P}(X = 0)\mathrm{P}(Y = -2) = 1/9$. You can compute similar examples for the other outcomes, but only one is needed to show that the two random variables are not independent.

2) First, the covariance:

$$\mathrm{Cov}(Z, Y) = \mathrm{Cov}(a + bX, Y) = \mathrm{E}[(a + bX)Y] - \mathrm{E}(a + bX)\mathrm{E}(Y)$$

Remember that expectation is linear (equation 5.4) and that the expectation of a constant is just that constant:

$$= \mathrm{E}[aY + bXY] - \mathrm{E}(a)\mathrm{E}(Y) - b\mathrm{E}(X)\mathrm{E}(Y) = a\mathrm{E}(Y) + b\mathrm{E}(XY) - a\mathrm{E}(Y) - b\mathrm{E}(X)\mathrm{E}(Y)$$
$$= b[\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)] = b\gamma.$$

To get the correlation, we take the covariance and divide by $\sqrt{\mathrm{Var}(Z)\mathrm{Var}(Y)}$. Remember that if X is a random variable a and b are constants, then $\mathrm{Var}(a + bX) = b^2\mathrm{Var}(X)$. Thus,

$$\mathrm{Cor}(Z, Y) = \frac{b\gamma}{\sqrt{\mathrm{Var}(Z)\mathrm{Var}(Y)}} = \frac{b\gamma}{\sqrt{\mathrm{Var}(a + bX)\mathrm{Var}(Y)}} = \frac{b\gamma}{\sqrt{b^2\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

The b^2 can come outside the square root as b, where it will cancel with the b in the numerator, leaving

$$\mathrm{Cor}(Z, Y) = \frac{\gamma}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

From the definition of correlation and the fact that $\gamma = \mathrm{Cov}(X, Y)$, it follows that this is

$$\mathrm{Cor}(Z, Y) = \mathrm{Cor}(X, Y) = \rho.$$

This shows the covariance changes with linear scaling of one (or both) of the random variables being considered, but the correlation does not change with scaling.

3) Start by distributing the joint distribution and then splitting up the sum,

$$\begin{aligned}\mathrm{E}(X+Y) &= \sum_{i=1}^{k}\sum_{j=1}^{m}(x_i + y_j) f_{X,Y}(x_i, y_j) = \sum_{i=1}^{k}\sum_{j=1}^{m}[x_i f_{X,Y}(x_i, y_j) + y_j f_{X,Y}(x_i, y_j)] \\ &= \sum_{i=1}^{k}\sum_{j=1}^{m} x_i f_{X,Y}(x_i, y_j) + \sum_{i=1}^{k}\sum_{j=1}^{m} y_j f_{X,Y}(x_i, y_j)\end{aligned}$$

Now in the first double-sum, the x_i can be pulled outside the sum over j (because it does not depend on j), and by equation 5.13, $\sum_{j=1}^{m} f_{X,Y}(x_i, y_j) = f_X(x_i)$, giving

$$\sum_{i=1}^{k}\sum_{j=1}^{m} x_i f_{X,Y}(x_i, y_j) = \sum_{i=1}^{k} x_i \sum_{j=1}^{m} f_{X,Y}(x_i, y_j) = \sum_{i=1}^{k} x_i f_X(x_i) = \mathrm{E}(X).$$

We are allowed to switch the order of the two sums in the second double-sum, after which we can make the analogous computation,

$$\sum_{j=1}^{m}\sum_{i=1}^{k} y_j f_{X,Y}(x_i, y_j) = \sum_{j=1}^{m} y_i f_Y(y_i) = \mathrm{E}(Y).$$

Replacing the two double sums in the first expression gives the result.

Exercise Set 5-4:

2) Here’s a set of commands that will let you explore the parameter set (1,1):

```
dosm.beta.hist(1, 10000, shape1 = 1, shape2 =  1)
dosm.beta.hist(2, 10000, shape1 = 1, shape2 =  1)
dosm.beta.hist(3, 10000, shape1 = 1, shape2 =  1)
dosm.beta.hist(4, 10000, shape1 = 1, shape2 =  1)
dosm.beta.hist(5, 10000, shape1 = 1, shape2 =  1)
dosm.beta.hist(10, 10000, shape1 = 1, shape2 =  1)
dosm.beta.hist(50, 10000, shape1 = 1, shape2 =  1)
```

In this case, the normal distribution is an acceptable approximation for the distribution of means of samples of size 5, and it's a great approximation for samples of 10 and 50. The other values of the shape parameters suggest that for distributions that are more skewed or more U-shaped, larger samples are required before the normal distribution is a good approximation, but for the parameter sets seen here, the means of samples of size 50 are well-approximated by a normal distribution. Try to find more extreme parameter sets and see how large the samples need to be before their distribution looks normal.

3) Below is commented code that performs the simulations and makes the comparisons. With the parameters and sample size requested, the distribution of sample means is a good fit to the normal within about 2 standard deviations of the expectation, and the histogram looks kind of normal. Beyond two standard deviations, though, the Pareto sample mean distribution has much heavier tails than the normal—extreme observations are much more likely than normal theory predicts. For example, there are about 100 times as many observations beyond 5 standard deviations from the expectation as would be predicted by the normal distribution, and there are thousands of times as many observations beyond 6 standard deviations as the normal distribution predicts. Thus, with this distribution and $n = 1{,}000$, convergence in the center of the distribution is good, but convergence in the tails is poor. If the probability of an extreme event (such as, say, an earthquake of Richter magnitude >8) is important to know, then the central limit theorem can lead to spectacularly poor predictions. Convergence is worse with smaller shape parameters and smaller sample size.

```
#Sample size per simulation (n) and number of simulations.
n <- 1000
n.sim <- 100000

#Pareto parameters. Variance is finite, and so
#CLT applies, if a > 2. For large a, convergence to
#normal is better. With small a, convergence is slow,
#especially in the tails.
a <- 3
b <- 1

#Compute the expectation and variance of the distribution
#of the sample mean. a must be above 2 for these expressions
#to hold.
expec.par <- a*b/(a-1)
var.par <- a*b^2 / ((a-1)^2 * (a-2))
sd.mean <- sqrt(var.par / n)

#Simulate data, compute sample means.
sim <- matrix(rpareto(n*n.sim, a, b), nrow = n.sim)
```

```
means.sim <- rowMeans(sim)

#Draw a histogram of the sample means along with the approximate
#normal pdf that follows from the CLT.
hist(means.sim, prob = TRUE)
curve(dnorm(x, expec.par, sd.mean), add = TRUE)

compare.tail.to.normal(means.sim, 1/2, expec.par, sd.mean)
compare.tail.to.normal(means.sim, 1, expec.par, sd.mean)
compare.tail.to.normal(means.sim, 2, expec.par, sd.mean)
compare.tail.to.normal(means.sim, 3, expec.par, sd.mean)
compare.tail.to.normal(means.sim, 4, expec.par, sd.mean)
compare.tail.to.normal(means.sim, 5, expec.par, sd.mean)
compare.tail.to.normal(means.sim, 6, expec.par, sd.mean)
```

Exercise Set 5-5:

1) From equation 5.16, the correlation coefficient is

$$\rho_{X,Y} = \beta \frac{\sigma_X}{\sigma_Y} = \frac{\beta \sigma_X}{\sqrt{\beta^2 \sigma_X^2 + \sigma_\epsilon^2}}.$$

Squaring the correlation coefficient gives

$$\rho_{X,Y}^2 = \left(\frac{\beta \sigma_X}{\sqrt{\beta^2 \sigma_X^2 + \sigma_\epsilon^2}} \right)^2 = \frac{\beta^2 \sigma_X^2}{\beta^2 \sigma_X^2 + \sigma_\epsilon^2} = 1 - \frac{\sigma_\epsilon^2}{\beta^2 \sigma_X^2 + \sigma_\epsilon^2} = 1 - \frac{\mathrm{Var}(Y|X = x)}{\mathrm{Var}(Y)}.$$

If the relationship between X and Y is linear—as it is in the model here—then the square of the correlation coefficient of X and Y is equal to one minus the proportion of the variance in Y that remains after conditioning on X. This is why people refer to the square of the correlation coefficient as “the proportion of variance explained.” This phrase is justified when the relationship between X and Y has the properties of the model we developed in the previous section.

2) When one repeatedly simulates with the same parameters, the resulting data vary. The degree to which they vary depends on the values of the parameters—in particular, making `var.eps` larger both decreases the apparent strength of the relationship between `x` and `y` and increases the extent to which the results of different simulations vary. Varying `a` changes the numbers on the y-axis but little else. Varying `b` changes the strength and direction of the relationship between `x` and `y`—large absolute values give apparently stronger relationships, and changing the value from positive to negative changes the direction of the relationship. Changing `var.x` changes the spread of the observations on both the x- and y-axes.