Studio 4 solutions, 18.05, Spring 2014
Jeremy Orloff and Jonathan Bloom

Problem 1. Covariance and correlation

Jack and Jill are dealers at a casino in Las Vegas.

• Every day during their break, Jack and Jill play roulette 10 times together, each betting one dollar on red each time.

• Jill then plays 5 more times alone, betting one dollar each time.

(In Las Vegas the roulette wheel has 18 red, 18 black and 2 green slots.)

(a) What are the expected winnings on a random day for each of them?

answer: For one play the probability of winning \$1 is 18/38 and the probability of losing \$1 is 20/38. Let W_i be the winnings on the ith bet.

$$E(W_i) = 18/38 - 20/38 = -2/38 = -1/19$$

Let W_{jill} be Jill's total winnings on a random day. Since $W_{\text{jill}} = W_1 + W_2 + \ldots + W_{15}$ we have

$$\boxed{E(W_{\text{jill}}) = \sum_{i=1}^{15} E(W_i) = -15/19 = -0.789}.$$

Likewise $W_{\text{jack}} = W_1 + W_2 + \ldots + W_{10}$ and $\boxed{E(W_{\text{jack}}) = -10/19 = -0.526.}$

(b) What is the variance of their winnings on a random day?

answer: *All the numerical calculations for this part are in studio4-sol.r.*

The variance for one bet is $\text{Var}(W_i) = E((W_i - 1/19)^2) = 0.9972$.

Since the bets W_i are independent we have

$$\text{Var}(W_{\text{jill}}) = 15 * \text{Var}(W_i) = 14.958 \qquad \text{and} \qquad \text{Var}(W_{\text{jack}}) = 10 * \text{Var}(W_i) = 9.972$$

(c) What is the covariance and correlation between their winnings.

answer: To answer this question we use the linearity of covariance. They key is that when

$i \neq j$ we know W_i and W_j are independent, so $\text{Cov}(W_i, W_j) = 0$. Therefore

$$\begin{aligned}
\text{Cov}(W_{\text{jill}}.W_{\text{jack}}) &= \text{Cov}(\sum_{i=1}^{15} W_i, \sum_{j=1}^{10} W_j) \\
&= \sum_{i,j} \text{Cov}(W_i, W_j) \\
&= \sum_{i=1}^{10} \text{Cov}(W_i, W_i) \qquad \text{(ignore all the terms that are 0)} \\
&= \sum_{i=1}^{10} \text{Var}(W_i) \\
&= 10 \cdot \text{Var}(W_i) \\
&= 9.972
\end{aligned}$$

We now have all the pieces to compute the correlation

$$\begin{aligned}
\text{Cor}(W_{\text{jill}}, W_{\text{jack}}) &= \frac{\text{Cov}(W_{\text{jill}}, W_{\text{jack}})}{\sqrt{\text{Var}(W_{\text{jill}})\text{Var}(W_{\text{jack}})}} \\
&= 0.816496 \qquad \text{(computed in studio4-sol.r)}
\end{aligned}$$

(d) Suppose Jill and Jack play roulette the same way for 100 days. Let JillTotal and JackTotal be their total winnings. What are the expected values, variances, covariance and correlation of these totals?

answer: Because JillTotal and JackTotal are the sum of 100 independent copies of W_{jill} and W_{jack} respectively, their expected value, variance and covariance are 100 times the corresponding values for W_{jill} and W_{jack}:

$E(\text{JillTotal}) = 100{*}E(W_{\text{jill}}) = -78.94737, \qquad E(\text{JackTotal}) = 100{*}E(W_{\text{jack}}) = -52.63158$

$\text{Var}(\text{JillTotal}) = 100 * \text{Var}(W_{\text{jill}}) = 1495.8, \qquad \text{Var}(\text{JackTotal}) = 100 * \text{Var}(W_{\text{jack}}) = 997.2$

$\text{Cov}(\text{JillTotal, JackTotal}) = 997.2.$

Becauses covariance and variances are all scaled by the same amount the correlation remains the same no matter what the number of days:

$$\text{Cor}(\text{JillTotal, JackTotal}) = \text{Cor}(W_{\text{jill}}, W_{\text{jack}}) = 0.816.$$

(e) Comment on how each of the quantities depends on the number of days.

answer: In general expected value, variance and covariance are the number of days times the corresponding values for W_{jill} and W_{jack}. The correlation is the same as that between W_{jill} and W_{jack}.

Problem 2. Simulation; Central Limit Theorem

(a) Write R code to simulate JackTotal (Jack's total winnings) for 100 days. Run 5000 trials and plot a density histogram of the results. (The code in studio4.r may be helpful.)

answer: See studio4-sol.r

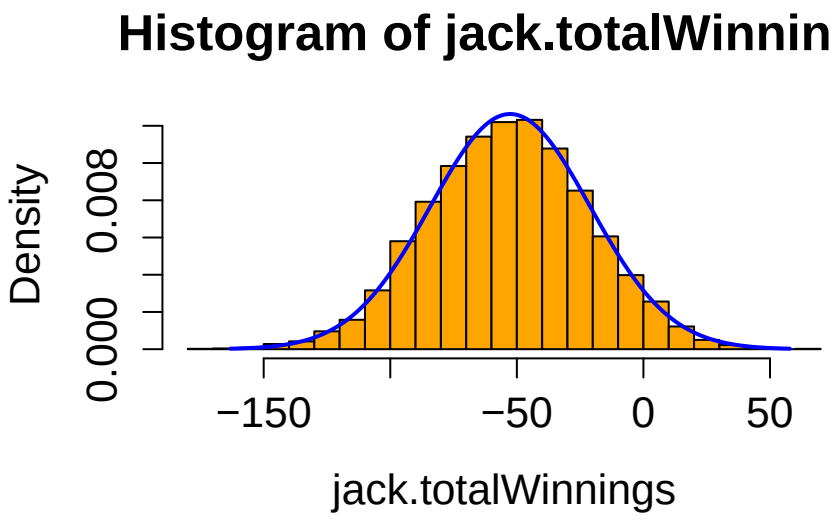
(b) Why does the central limit theorem apply to approximating JackTotal.

answer: The central limit theorem says that the sum of a large number of independent identically distributed (i.i.d.) random variables is approximately normal. That's exactly what we have when we sum the winnings for 100 days. In fact, since one day's winnings for Jack is itself the sum of 10 i.i.d. random variables, the total for 100 days is the sum of 1000 i.i.d. random variables.

(c) What μ and σ should be used in $\mathrm{N}(\mu, \sigma^2)$ to approximate JackTotal? Add a graph of the pdf of $\mathrm{N}(\mu, \sigma^2)$ to your histogram in part (a).

answer: $\mu = E(\text{winnings over 100 days}) = 100 \cdot E(W_{\text{jack}}) = \boxed{-52.6}$.

$\sigma^2 = \text{Var}(\text{winnings over 100 days}) = 100 \cdot \text{Var}(W_{\text{jack}}) = 9972$. So $\boxed{\sigma = 31.6}$.



Density histogram and normal pdf for problem 2

(d) Using the central limit theorem give a range with about 95% probability of containing JackTotal.

answer: For a normal distribution we know that 95% of the probability is within 2 standard deviations of the mean. Since the central limit theorem says JackTotal is approximately normal the 95% interval is approximately

$$\mu \pm 2\sigma = -52.6 \pm 63.2 = [-115.8,\ 10.5]$$

(e) What percentage of simulated JackTotal values fall in the range found in part (d).

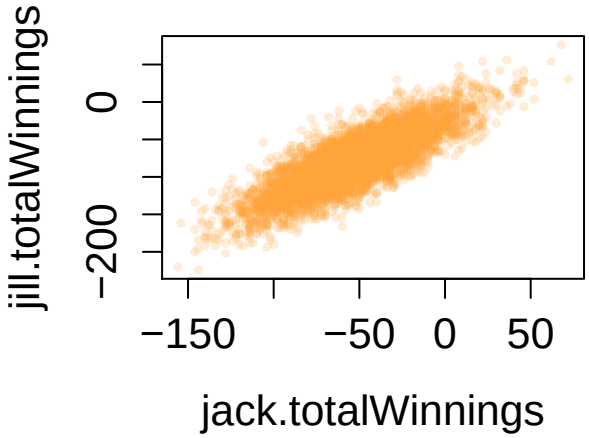
answer: In studio4-sol.r we found that the fraction in the interval in (d) was 0.953. Since the code does not set the random seed, running the code again would give a slightly different answer.

Problem 3. Simulation; covariance and correlation

Extend the simulation from problem 2 to include JillTotal, Jill's total winnings over 100 days.

(a) Plot JillTotal vs. JackTotal for the simulation data.

answer: *See studio4-sol.r for the code doing this.*



Plot of simulation of JillTotal vs. JackTotal for problem 3

(b) Compute the (sample) covariance and correlation of JackTotal and JillTotal from the simulation.

answer: *Computation done in studio4-sol.r*

In the simulation we computed a (simulated) correlation of 0.8134085 based on 5000 trials. This is very close to our theoretical value computed in problem 1.

Again, since the code doesn't set the random seed, running it again will give slighlty different results.

18.05 Introduction to Probability and Statistics
Spring 2014