

# The Algebra of Random Variables

## A Exploration through Simulation

### Exploration 1: Linear Functions of a Random Variable

Procedure

- A. Simulate the flipping of a fair coin. Assign each ‘heads’ the value  $X = 0$ ; each ‘tails’ the value  $X = 1$ .
- B. Copy the table below and record your data.
- C. Repeat steps A–B until you have recorded 50 flips.

Flip Number	Flip Result $X$	$2X$	$-4X$	$X+2$	$X-1$	$-2X+1$	$\frac{X}{2} - 2$
1							
2							
⋮							
50							
Mean							
Std. Dev							
Variance							

- D. Complete the table, as shown above, by calculating values of the flip results times two, times four, etc.
- E. Calculate the mean, sample standard deviation, and sample variance for each column in the table.
- F. What are your observations about the summary statistics in the table? Compare your results to those of your classmates.
- G. Calculate the theoretical mean, standard deviation, and variance of the probability distribution of a coin flip, treating heads as 0 and tails as 1.
- H. Repeat step G for each of the linear functions on the flip result as shown in the table, i.e. if the linear function is  $2X + 1$ , then a heads is  $2(0) + 1 = 1$  and a tails is  $2(1) + 1 = 3$ .
- I. Are your actual results calculated in Step E. equal to the theoretical results calculated in Steps G–H? If different, how much difference is there and to what can you attribute the difference?
- J. Are the relationships between your summary statistics for  $X$ ,  $2X$ , etc. consistent with the relationships between the theoretical values? If so, how? If not, can you account for the difference?

The Algebra of Random Variables  
 Exploration 2: Linear Combinations of Random Variables  
 Part 1

Procedure

- A. Simulate the flipping of two fair coins. Assign each ‘heads’ the value 0; each ‘tails’ the value 1.
- B. Copy the table at right and record your data.
- C. Repeat steps A–B until you have recorded 50 pairs of flips.
- D. Complete the table, as shown at right, by finding the sum of  $X_1$  and  $X_2$ .
- E. Calculate the mean, sample standard deviation, and sample variance for each column in the table.
- F. What are your observations about the summary statistics in the table? Combining the data of you and your classmates, create a histogram of the means, standard deviations, and variances of the sums.
- G. Calculate the theoretical mean, standard deviation, and variance of the probability distributions of both individual flips and the sum of the flips.
- H. What conclusions can you draw about the means, standard deviations, and variances of random variables that are added together?

Trial Number	Flip 1 $X_1$	Flip 2 $X_2$	Sum $X_1 + X_2$
1			
2			
⋮			
50			
Mean			
Std. Dev.			
Variance			

Exploration 2: Linear Combinations of Random Variables  
 Part 2

Repeat Exploration 2, Part 1 finding the difference of  $X_1$  and  $X_2$  and drawing conclusions about the means, standard deviations, and variances of random variables that are subtracted from each other.

Trial Number	Flip 1 $X_1$	Flip 2 $X_2$	Difference $X_1 - X_2$
1			

The Algebra of Random Variables  
 Exploration 2: Linear Combinations of Random Variables  
 Part 3

Procedure

A. Simulate the flipping of a fair coin. Assign each ‘heads’ the value 0; each ‘tails’ the value 1. Record the result in column “Flip 1,  $X_1$ .”

B. Flip 2 depends on the outcome of Flip 1. If Flip 1 is a 1 (tails), then simulate Flip 2 as you did Flip 1, i.e. heads = 0, tails = 1. If Flip 1 is a 0 (heads), then Flip 2 is automatically tails, i.e. assign it the value 1. Record the result in column “Flip 2,  $X_2$ .”

Trial Number	Flip 1 $X_1$	Flip 2 $X_2$	Sum $X_1 + X_2$
1			
2			
⋮			
Mean			
Std. Dev.			
Variance			

C. Repeat steps A–B until you have recorded 50 pairs of flips.

D. Complete the table by finding the sum of  $X_1$  and  $X_2$ .

E. Calculate the mean, sample standard deviation, and sample variance for each column in the table.

F. What are your observations about the summary statistics in the table? Compare your results to those of your classmates.

G. Combining the data of you and your classmates, create a histogram of the means, standard deviations, and variances of the sums.

H. Calculate the theoretical mean, standard deviation, and variance of the probability distributions of both individual flips and the sum of the flips.

I. What conclusions can you now draw about the means, standard deviations, and variances of random variables that are added together?

J. How are your conclusions different than those from Parts 1 and 2 of Exploration 2? If different, to what might you attribute these differences?

The Algebra of Random Variables  
 Exploration 2: Linear Combinations of Random Variables  
 Part 4

Repeat Exploration 2, Part 3 with the following changes.

1. Assign the first coin a numerical value  $X_1$  as in Parts 1–3. The numerical value  $X_2$  assigned to Flip 2, is equal to that of Flip 1, i.e. if  $X_1 = 0$ , then  $X_2 = 0$ .

Trial Number	Flip 1 $X_1$	Flip 2 $X_2$	Difference $X_1 - X_2$
1			

2. Complete the table by calculating the difference of  $X_1$  and  $X_2$ .

The Algebra of Random Variables  
A Exploration through Simulation  
Teacher Notes

The algebra of random variables is an important, yet often overlooked or underdeveloped, concept in AP Statistics. When discussed, it is frequently presented as a series of formulas that students find unimpressive. Students may also consider such rules as another set of ‘magical’ theories with little practical meaning.

This activity seeks to not only verify the rules of the algebra of random variables, but to also reduce their level of mystery. Through simulation, students will hopefully gain a greater appreciation for these statistical notions.

In the commentary that follows, a variety of simulation techniques are discussed, including hands-on methods, use of random digit tables, and incorporation of technology. You may choose to use whatever methods with which your students are comfortable. If students have not had much exposure to simulation with technology, more concrete approaches, such as random digit tables or actual coin flipping, should be done first. Projected results and possible trouble spots are also addressed.

Table of Contents:

Exploration 1: Linear Functions of Random Variables ..... Page 1  
Exploration 2: Linear Combinations of Random Variables  
    Part 1: Sum of Independent Random Variables ..... Page 4  
    Part 2: Difference of Independent Random Variables..... Page 7  
    Part 3: Sum of Dependent Random Variables ..... Page 9  
    Part 4: Difference of Dependent (and Perfectly  
        Correlated) Random Variables ..... Page 13  
    Some Final Commentary ..... Page 16

Exploration 1: Linear Functions of a Random Variable

**The objective of this exploration is to verify, via simulation, the following rule:**

If  $x$  is a random variable with mean  $m_x$  and variance  $s_x^2$ , and  $a$  and  $b$  are numerical constants, the random variable  $y$  defined by  
$$y = a + bx$$
is called a **linear function of the random variable  $x$** .  
The mean of  $y = a + bx$  is  
$$m_y = a + bm_x$$
The variance of  $y$  is  
$$s_y^2 = b^2 s_x^2$$
from which it follows that the standard deviation of  $y$  is  
$$s_y = |b|s_x$$
(Peck, Olsen, Devore, 2001)

### Steps A–C: Simulation of the coin flips

- Hands-on: Students would use actual coins.
- Random Digit Table: Some scheme that partitions possible digits (or groups of digits) into two equally likely groups. Simple single-digit schemes might include 0–4 = heads and 5–9 = tails, or even = heads, odd = tails. Allow students to create their own schemes.
- Graphing calculators: Through either built-in functions or programs, students generate a series of 0's and 1's. The TI-83 code is `RandInt(0,1,50)`.
- Statistical Software: Many statistical software packages allow for the generation of random data. Consult your software manual to determine how. A few sample commands are shown below.

Microsoft Excel:	<code>Int(2*Rand())</code>	Note:	this generates a single cell of data.
MINITAB:	<code>Random 50 C1;</code> <code>Integer 0 1.</code>	Note:	this command generates 50 data points in column C1
JMP-INTRO:	<code>Random Integer(2) – 1</code>	Note:	this generates data in every row of the selected column.

The sheer speed at which software can generate large data sets is its major advantage.

### Steps D–E: Calculations of the linear functions and summary statistics

Having students enter their data into calculator lists, spreadsheets, or statistical software will accomplish this. Most calculators, spreadsheets, and software packages allow lists to be functions of other lists.

The following is a portion of a MINITAB analysis of a simulation. The variance column was not included in the analysis and was calculated by squaring the standard deviation.

Variable	N	Mean	StDev	Variance
X	50	0.5600	0.5014	0.2514
2X	50	1.120	1.0028	1.0056
-4X	50	-2.240	2.0056	4.0224
X+2	50	2.5600	0.5014	0.2514
X-1	50	-0.4400	0.5014	0.2514
-2X+1	50	-0.120	1.0028	1.0056
0.5X-2	50	-1.7200	0.2507	0.0629

### Step F: Observations and analysis

Students should observe that multiplying  $X$  by a constant results in the mean also being multiplied by that constant, the standard deviation multiplied by the absolute value of the constant, and the variance multiplied by the square of the constant. Students also quickly see that adding a constant changes the mean by that constant, but the standard deviation and variance remain the same. A combination of two operations produces similar effects.

**Steps G–H: Theoretical statistics**

The probability distribution of a coin flip where heads = 0, tails = 1 is given at right. This can be calculated with the formulas  $m = \sum x \cdot P(x)$  and  $s^2 = \sum (x - m)^2 \cdot P(x)$ , via lists in a graphing calculator, or by computer software.

Prob. Dist. for $X$	
Outcome ( $x$ )	Probability $P(x)$
0	0.5
1	0.5
Mean	0.50
Std. Dev.	0.50
Variance	0.25

The probability distributions for  $-4X$  and  $\frac{X}{2} - 2$  are also provided to illustrate theoretical results. Students should observe that the theoretical results are similar to those from the simulation.

Prob. Dist. for $-4X$	
Outcome ( $x$ )	Probability $P(x)$
0	0.5
-4	0.5
Mean	-2.00
Std. Dev.	2.00
Variance	4.00

**Steps I–J: Analysis and Summary**

Many students may not obtain the theoretical mean, standard deviation, and variance for  $X$  from their respective simulations. Any difference can be attributed to random variation and that difference will carry over to the linear functions of  $X$ . Regardless of this, the means of the linear functions of  $X$  must follow the law for linear functions. That is, their simulated mean for  $2X$  must be twice the mean of  $X$ , the mean for  $-2X + 1$  must be one more than twice the opposite of the mean of  $X$ , etc. The properties for standard deviation and variance must also follow.

Prob. Dist. for $\frac{X}{2} - 2$	
Outcome ( $x$ )	Probability $P(x)$
-2	0.5
-1.5	0.5
Mean	-1.75
Std. Dev.	0.25
Variance	0.0625

The Algebra of Random Variables  
 Exploration 2: Linear Combinations of Random Variables  
 Part 1: Sum of Independent Random Variables

A linear combination of random variables is defined as follows:

If  $x_1, x_2, \dots, x_n$  are random variables and  $a_1, a_2, \dots, a_n$  are numerical constants, the random variable defined as

$$y = a_1x_1 + a_2x_2 + \dots + a_nx_n$$

is a **linear combination of the  $x_i$ 's**.  
 (Peck, Olsen, Devore, 2001)

The objective of this exploration is to verify, via simulation, the following rule:

If  $x_1, x_2, \dots, x_n$  are random variables with means  $m_1, m_2, \dots, m_n$  and variances  $s_1^2, s_2^2, \dots, s_n^2$ , respectively, and

$$y = a_1x_1 + a_2x_2 + \dots + a_nx_n$$

then

1.  $m_y = m_{a_1x_1 + a_2x_2 + \dots + a_nx_n} = a_1m_{x_1} + a_2m_{x_2} + \dots + a_nm_{x_n}$   
 This result is true regardless of whether the  $x_i$ 's are independent.
2. When  $x_1, x_2, \dots, x_n$  are independent random variables  
 $s_y^2 = s_{a_1x_1 + a_2x_2 + \dots + a_nx_n}^2 = a_1^2s_1^2 + a_2^2s_2^2 + \dots + a_n^2s_n^2$   
 $s_y = s_{a_1x_1 + a_2x_2 + \dots + a_nx_n} = \sqrt{a_1^2s_1^2 + a_2^2s_2^2 + \dots + a_n^2s_n^2}$

This result is true **only** when the  $x_i$ 's are independent.  
 (Peck, Olsen, Devore, 2001)

For purposes of this activity, we will simplify the rule to be:

If  $X_1$  and  $X_2$  are random variables with means  $m_{x_1}$  and  $m_{x_2}$ , and variances  $s_{x_1}^2$  and  $s_{x_2}^2$ , then the random variable  $Y = X_1 \pm X_2$  has mean  $m_y = m_{x_1} \pm m_{x_2}$  and variance  $s_y^2 = s_{x_1}^2 + s_{x_2}^2$ . The rule for the mean always holds, but the rule for the variance only holds if  $X_1$  and  $X_2$  are independent.

**Objective: Students will observe that the above rule holds for the mean (which it *always* does) and also holds for the variance because the random variables are independent.**

Steps A–C: Simulation of the coin flips

The process here is the same as in Exploration 1 except that two sets of flips will be generated.

Steps D–E: Calculations of the linear combinations and summary statistics

The process here is again the same as in Exploration 1, except that the columns of data for  $X_1$  and  $X_2$  will be summed.

The following is a portion of a MINTAB analysis of a simulation by a single student. The variance column was not included in the analysis and was calculated by squaring the standard deviation.

Variable	N	Mean	StDev	Variance
X1	50	0.5400	0.5035	0.2535
X2	50	0.6800	0.4712	0.2220
X1+X2	50	1.220	0.708	0.5013

### Step F: Observations and distributions of class data

Students should observe that the mean of the sum  $X_1 + X_2$  is the sum of the means of  $X_1$  and  $X_2$ . This will always be the case because they are algebraically equivalent.

Students should also observe that the standard deviation of the sum is not the sum of the standard deviations—this is correct. Note that the variance of the sum is not equal to sum of the variances exactly due to sampling variability.

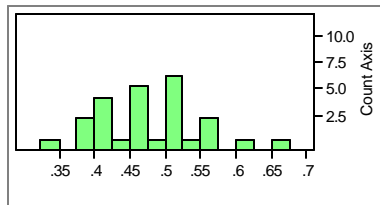
One way to avoid this is to increase the number of pairs of flips. The number of pairs was chosen to be 50 because it can be reasonably be done by hand with coins or a random digit table in a brief period of time. A simulation with a graphing calculator or statistics software can generate larger samples quickly.

The comparison with classmates is very important here, as some sums will be close and some will not. The construction of the histograms in the next step is crucial.

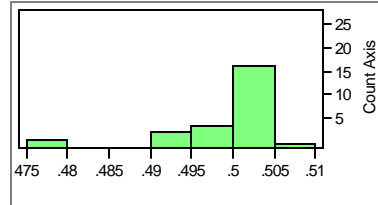
The JMP-INTRO graphs below are simulated results for 30 students.

#### Distributions

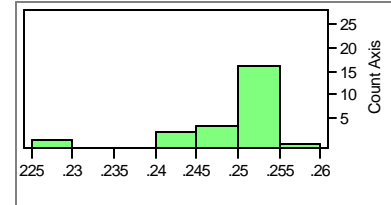
##### Flip 1 Mean



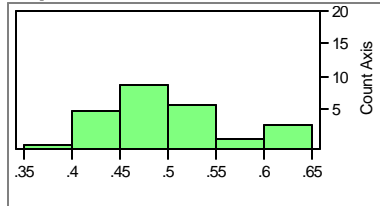
##### Flip 1 SD



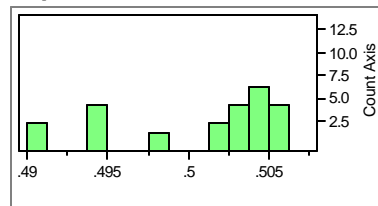
##### Flip 1 Var



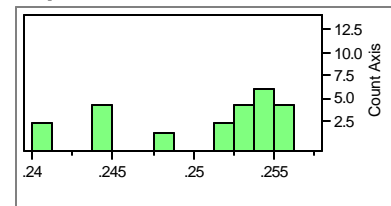
##### Flip 2 Mean



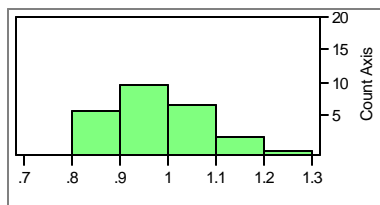
##### Flip 2 SD



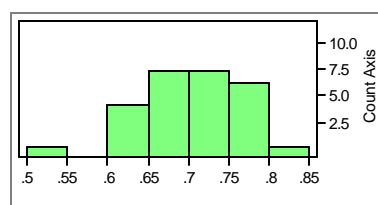
##### Flip 2 Var



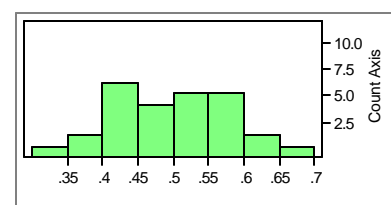
##### Sum Mean



##### Sum SD



##### Sum Var



The centers of the Mean histograms for Flip 1, Flip 2, and the Sum are about 0.5, 0.5 and 1.0 respectively. The mean of the sum equals the sum of the means no matter what data are generated. This will always be the case because they are algebraically equivalent.

The centers of the Standard Deviation histograms are about 0.5, 0.5, and 0.7. The standard deviation of the sum does not appear to be the sum of the standard deviations. This is expected.

The centers of the Variance histograms are at about 0.25, 0.25, and 0.5. This illustrates that the variance of the sum is the sum of the variances and that these values are the theoretical variances.

The following MINITAB summary statistics match the observations from the graphs.  $N = 30$  is the number of students, each of whom did 50 of Flip 1 and Flip 2.

Variable	N	Mean
Flip 1 Mean	30	0.4787
Flip 2 Mean	30	0.4887
Sum Mean	30	0.9673
Flip 1 SD	30	0.49945
Flip 2 SD	30	0.50074
Sum SD	30	0.7020
Flip 1 Var	30	0.24950
Flip 2 Var	30	0.25076
Sum Var	30	0.4966

Prob. Dist. for $X_1, X_2$	
Outcome ( $x$ )	Probability $P(x)$
0	0.5
1	0.5
Mean	0.50
Std. Dev.	0.50
Variance	0.25

### Step G: Theoretical statistics

The probability distribution of a coin flip where heads = 0, tails = 1 is given at right. This will be the same for both individual coins. The second table is the probability distribution for the sum.

Students should observe that the theoretical results parallel those from the simulation.

### Step H: Conclusions

After the above analysis, students should conclude that when independent random variables are added together, the means and variances add, but the standard deviations do not.

Prob. Dist. for $X_1 + X_2$	
Outcome ( $x$ )	Probability $P(x)$
0	0.25
1	0.50
2	0.25
Mean	1.00
Std. Dev.	0.71
Variance	0.50

Exploration 2: Linear Combinations of Random Variables  
Part 2: Difference of Independent Random Variables

**Objective:** Students will observe that the rule stated in Part 1 holds for the mean (which it *always* does) and also holds for the variance because the random variables are independent.

Essentially, Part 2 differs from Part 1 only in that we are now subtracting the random variables instead of adding them. What is counterintuitive to students is that when subtracting independent random variables, the variance of the difference is the *sum* of the variances.

The process of the activity is the same as in Part 1 and the concerns over small size ( $n = 50$ ) are also the same. Sample output and theoretical values are provided as before.

**Steps D–E:** Calculations of the linear combinations and summary statistics

The following is a portion of a JMP-INTRO analysis of a simulation.

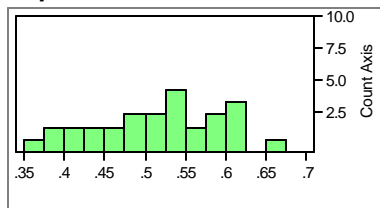
<b>X1</b>		<b>X2</b>		<b>X1-X2</b>	
<b>Moments</b>		<b>Moments</b>		<b>Moments</b>	
Mean	0.52	Mean	0.5	Mean	0.02
Std Dev	0.504672	Std Dev	0.5050763	Std Dev	0.6543419
N	50	N	50	N	50
Variance	0.2546939	Variance	0.255102	Variance	0.4281633

## Step F: Observations and distributions of class data

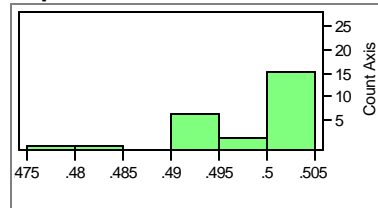
The JMP-INTRO graphs below are simulated results for 30 students, each collecting 50 pairs of data.

### Distributions

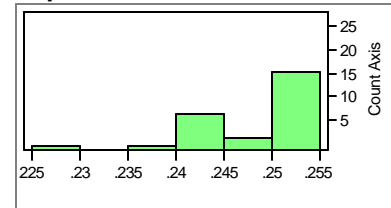
**Flip 1 Mean**



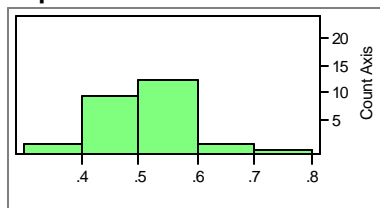
**Flip 1 SD**



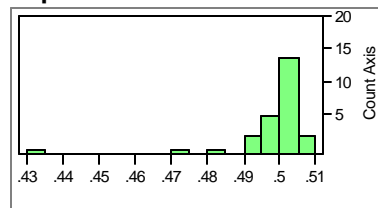
**Flip 1 Var**



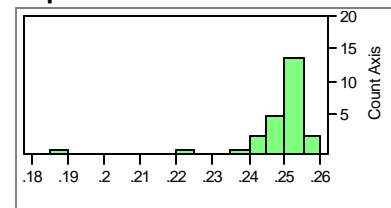
**Flip 2 Mean**



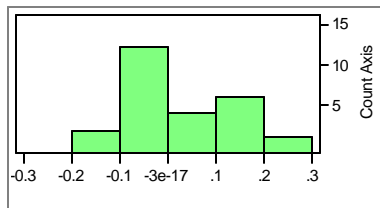
**Flip 2 SD**



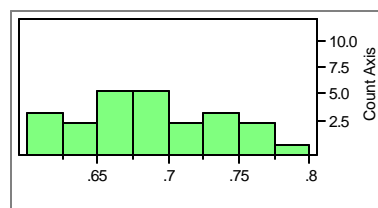
**Flip 2 Var**



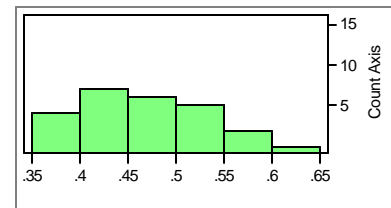
**Diff Mean**



**Diff SD**



**Diff Var**



The centers of the Mean histograms for Flip 1, Flip 2, and the Difference are about 0.5, 0.5 and 0.0 respectively. The mean of the difference equals the difference of the means no matter what data are generated. This will always be the case because they are algebraically equivalent.

The centers of the Standard Deviation histograms are about 0.5, 0.5, and 0.7. The standard deviation of the difference does not appear to be the sum or difference of the standard deviations.

The centers of the Variance histograms are at about 0.25, 0.25, and 0.5. This tends to confirm that the variance of the difference is the *sum* of the variances.

The following MINITAB summary statistics match the observations from the graphs.

Variable	N	Mean	Median
Coin 1 Mean	30	0.4787	0.4600
Coin 2 Mean	30	0.4887	0.4800
Diff Mean	30	0.9673	0.9600
Coin 1 SD	30	0.49945	0.50244
Coin 2 SD	30	0.50074	0.50346
Diff SD	30	0.7020	0.7071
Coin 1 Var	30	0.24950	0.25245
Coin 2 Var	30	0.25076	0.25347

Prob. Dist. for $X_1, X_2$	
Outcome ( $x$ )	Probability $P(x)$
0	0.5
1	0.5
Mean	0.50
Std. Dev.	0.50
Variance	0.25

Diff Var	30	0.4966	0.5000
----------	----	--------	--------

### Step G: Theoretical statistics

The probability distribution of a coin flip where heads = 0, tails = 1 is given at right. This will be the same for both individual coins. The second table is the probability distribution for the difference. Students should observe that the theoretical results parallel those from the simulation.

### Step H: Conclusions

After the above analysis, students should conclude that when independent random variables are subtracted from each other, the means subtract, the variances add, but the standard deviations do neither.

Prob. Dist. for $X_1 - X_2$	
Outcome ( $x$ )	Probability $P(x)$
-1	0.25
0	0.50
1	0.25
Mean	0.00
Std. Dev.	0.71
Variance	0.50

## Exploration 2: Linear Combinations of Random Variables Part 3: Sum of Dependent Random Variables

**Objective: Students will observe that the rule stated in Part 1 holds for the mean (which it *always* does), but does not hold for the variance because the random variables are dependent.**

In Part 3, the results of the coin flips,  $X_1$  and  $X_2$  are not independent as they were in Parts 1 and 2. The value of  $X_2$  is dependent on the value of  $X_1$ . The objective here is to demonstrate that while the mean of the sum of dependent random variables is the sum of the means, the variance of the sum is *not* the sum of the variances.

### Steps A–D: Simulation of the coin flips

- Hands-on: Students would use actual coins. To simulate the dependence, students would do their Flips 1 as usual. If Flip 1 is ‘heads’ ( $X_1 = 0$ ), they need not flip again and may simply record a value of 1 for  $X_2$ .
- Random Digit Table: Schemes for Flip 1 can be done as in Parts 1 and 2. Students should realize if the first flip corresponds to ‘heads’, the reading of a second digit is not necessary for that step.
- Graphing calculators: Simply generating two lists of 0’s and 1’s using a random integer functions will not work. Creating one for the first coin flip is necessary, but the second flip now depends on the first. One way around this is to generate the second list, then go through it by hand changing values in the list to 1 that correspond to a 0 in the first list. A second method would be to write a program that generates the numbers itself, or makes the substitutions mentioned in the previous sentence.

TI-83 Code to generate second list  
(assume Flips 1 are in L1)

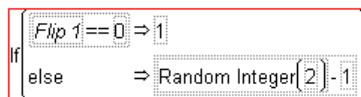
```
PROGRAM:DEPEND
For(A,1,50)
If L1(A)=0
Then
1→L2(A)
Else
randInt(0,1)→L2(A)
End
End
```

TI-83 Code to switch 0’s to 1’s in  
second list (assume Flips 1 are in  
L1, Flips 2 are in L2)

```
PROGRAM:SWITCH
For(A,1,50)
If L1(A)=0
1→L2(A)
End
```

- Statistical Software: Some software packages (like JMP-INTRO or Excel) allow one to use conditional statements when entering a formula. Others (like MINITAB) will require the creation of a Macro to deal with the condition.

JMP-INTRO Formula:



*Column Menu, Select  
Formula: “If” is in the  
Conditional Function  
menu, “Random Integer” is  
in the Random Function  
menu.*

MINITAB Macro:

```
gmacro
depend
name c1='X1'
name c2='X2'
name c3='X1+X2'
Random 50 c1 c2;
Integer 0 1.
do k2=1:50
if c1(k2)=0
let c2(k2)=1
endif
```

```
enddo  
let c3 = c1 + c2  
endmacro
```

### Step E: Calculations of the linear combinations and summary statistics

The following is a portion of a MINITAB analysis of a simulation. The variance column was not included in the analysis and was calculated by squaring the standard deviation.

Variable	N	Mean	StDev	Variance
X1	50	0.4800	0.5047	0.2547
X2	50	0.8000	0.4041	0.1633
X1+X2	50	1.2800	0.4536	0.2058

### Step F: Observations and analysis

Students should observe that the mean of the sum  $X_1 + X_2$  is the sum of the means of  $X_1$  and  $X_2$ . They should also observe that the standard deviation of the sum is not the sum of the standard deviations, nor are the variances. This is in contradiction of Part 1, and this is because the variables are no longer independent. As we will show in Step H, for the theoretical values of the random variables, the variance of the sum is no longer the sum of the variances. (Shhhh. Don't tell them this! Yet.)

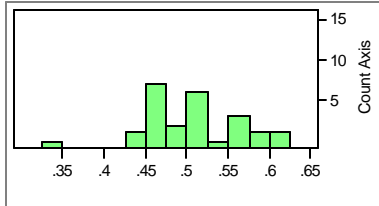
Again, students should compare their results with classmates.

### Step G: Distributions of class data

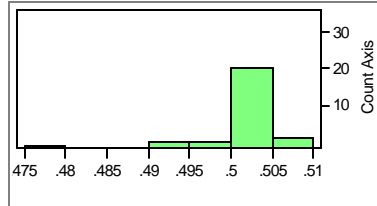
The JMP-INTRO graphs below are simulated results for 30 students.

#### Distributions

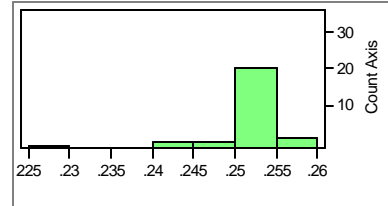
**Flip 1 Mean**



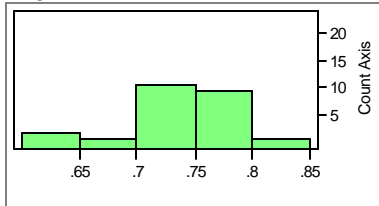
**Flip 1 SD**



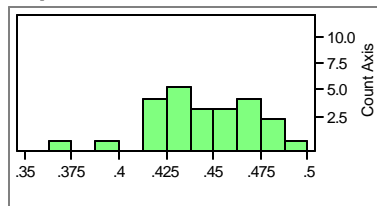
**Flip 1 Var**



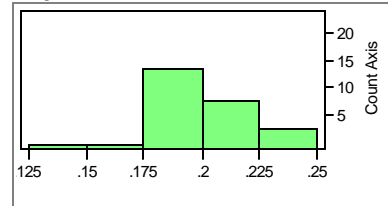
**Flip 2 Mean**



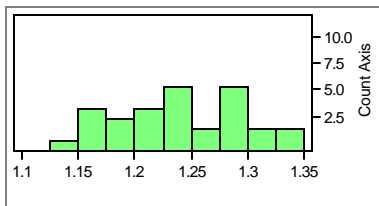
**Flip 2 SD**



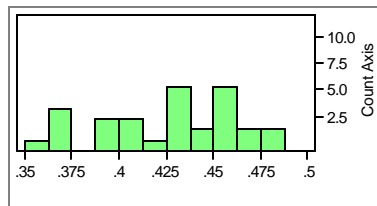
**Flip 2 Var**



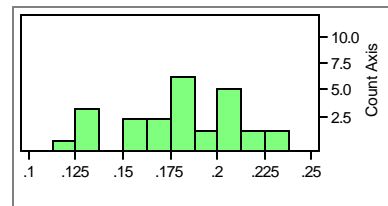
**Sum Mean**



**Sum SD**



**Sum Var**



The centers of the Mean histograms for Flip 1, Flip 2, and the Sum are about 0.5, 0.75 and 1.25 respectively. This illustrates that the mean of the sum is the sum of the means, even when the variables are *not independent*. Again, this will always be the case because they are algebraically equivalent.

The centers of the Standard Deviation histograms are about 0.5, 0.45, and 0.43. The standard deviation of the sum does not appear to be the sum of the standard deviations.

The centers of the Variance histograms are at about 0.253, 0.20, and 0.18. This tends to confirm that when the variables are not independent that the variance of the sum is *not* the sum of the variances.

The following MINITAB summary statistics match the observations from the graphs.

Variable	N	Mean	Median
Coin 1 Mean	30	0.5027	0.5000
Coin 2 Mean	30	0.73333	0.74000
Sum Mean	30	1.2360	1.2400
Coin 1 SD	30	0.50162	0.50346
Coin 2 SD	30	0.44291	0.44309
Sum SD	30	0.42380	0.43142
Coin 1 Var	30	0.25165	0.25347
Coin 2 Var	30	0.19690	0.19633
Sum Var	30	0.18087	0.18612

Prob. Dist. for $X_1$	
Outcome ( $x$ )	Probability $P(x)$
0	0.5
1	0.5
Mean	0.50
Std. Dev.	0.50
Variance	0.25

### Step H: Theoretical statistics

Calculation of the theoretical statistics of the distributions may require construction of the sample space or a tree diagram.

The probability distribution of a coin flip where heads = 0, tails = 1 is given at right in the top table. The second table is the probability distribution for the second value, and the third table is the probability distribution for the sum. Students should observe that the theoretical results are similar to those from the simulation.

Prob. Dist. for $X_2$	
Outcome ( $x$ )	Probability $P(x)$
0	0.25
1	0.75
Mean	0.750
Std. Dev.	0.433
Variance	0.187

### Steps I–J: Conclusions

Prob. Dist. for $X_1 + X_2$	
Outcome ( $x$ )	Probability $P(x)$
1	0.75
2	0.25
Mean	1.250
Std. Dev.	0.433

Variance	0.187
----------	-------

After the above analysis, students should conclude that when *dependent* random variables are added together, the mean of the sum is the sum of the means, but the standard deviation and variance of the sum is not the sum of individual standard deviations and variances.

Exploration 2: Linear Combinations of Random Variables  
Part 4: Difference of Dependent (and Perfectly Correlated) Random Variables

**Objective: Students will observe that the rule stated in Part 1 holds for the mean (which it *always* does), but does not hold for the variance because the random variables are dependent and, in this case, perfectly correlated.**

In Part 4, the results of the coin flips,  $X_1$  and  $X_2$  are not independent. As a matter of fact, there is no random element at all for the value of  $X_2$ , and it is perfectly correlated with  $X_1$ .

**Steps A–D: Simulation of the coin flips**

That of Flip 1 is done as in Parts 1-3. Flip 2's value is equal to Flip 1's.

**Step E: Calculations of the linear combinations and summary statistics**

The following is a portion of a JMP-INTRO analysis of a simulation.

<b>X1 Moments</b>		<b>X2 Moments</b>		<b>X1-X2 Moments</b>	
Mean	0.58	Mean	0.58	Mean	0
Std Dev	0.4985694	Std Dev	0.4985694	Std Dev	0
N	50	N	50	N	50
Variance	0.2485714	Variance	0.2485714	Variance	0

**Step F: Observations and analysis**

Students should observe that the mean of the difference is zero! But so is the standard deviation. The only way that the standard deviation can be 0 is if all of the data are equal. We have a perfect correlation between  $X_1$  and  $X_2$ , and they are dependent on each other. It is very clear that the mean of the difference is equal to the difference in means, but the variance of the difference is not the sum of the variances as before.

Again, students should compare their results with classmates.

A cautionary note: In this instance the standard deviation of the difference is equal to the difference of the standard deviations. The same is true for variance. This is not a coincidence, but it is not a general rule either. It occurs in this special case because  $X_1$  and  $X_2$  are equal.

**Step G: Distributions of class data**

Sample distributions of Flips 1 and 2 should be identical. The distributions of the difference should be a single column at 0.

### Step H: Theoretical statistics

The probability distribution of a coin flip where heads = 0, tails = 1 is given at right in the top table. The second table is the probability distribution for the difference.

Prob. Dist. for $X_1, X_2$	
Outcome ( $x$ )	Probability $P(x)$
0	0.5
1	0.5
Mean	0.50
Std. Dev.	0.50
Variance	0.25
Prob. Dist. for $X_1 - X_2$	
Outcome ( $x$ )	Probability $P(x)$
0	1
Mean	0
Std. Dev.	0
Variance	0

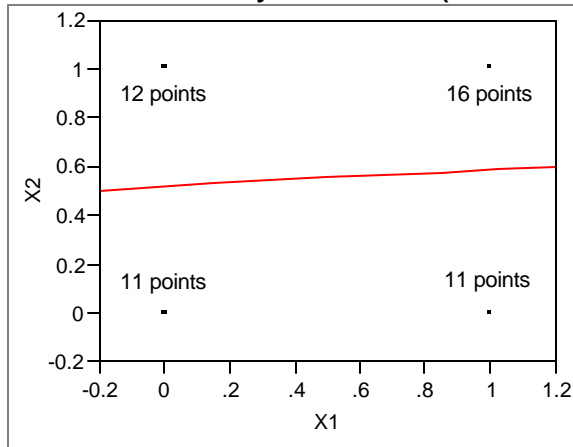
### Steps I–J: Conclusions

After the above analysis, students should definitely conclude that when random variables are completely *dependent* on each other, as in this case, the mean of the difference is the difference in means, but the rules for standard deviation and variance of independent random variables do not hold.

### Additional Analysis

The notions of independence and dependence are also observed in a bivariate analysis of  $X_2$  vs.  $X_1$ . The graphs and analyses below are from JMP-INTRO and show both simulate and theoretical results from Parts 1, 3, and 4. Part 2 has the same distribution of points as Part 1, and results from Part 2 would be equivalent to Part 1. The frequencies of points have been added to the graphs to better visualize their distribution.

**Bivariate Fit of X2 By X1 for Part 1 (Simulated)**



#### Linear Fit

$$X_2 = 0.5217391 + 0.0708535 X_1$$

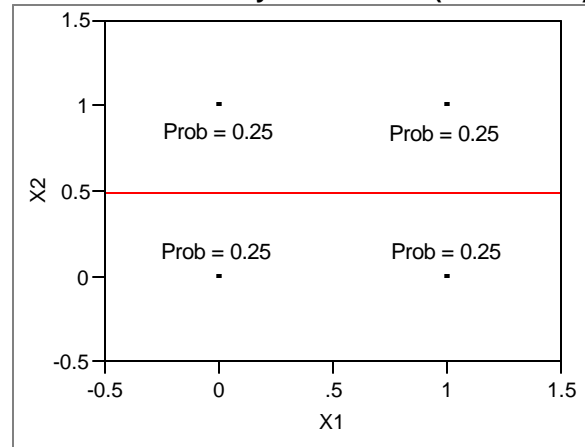
#### Summary of Fit

RSquare 0.005061

#### Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	0.5217391	0.105371	4.95	<.0001
X1	0.0708535	0.143391	0.49	0.6235

**Bivariate Fit of X2 By X1 for Part 1 (Theoretical)**



#### Linear Fit

$$X_2 = 0.5 + 0 X_1$$

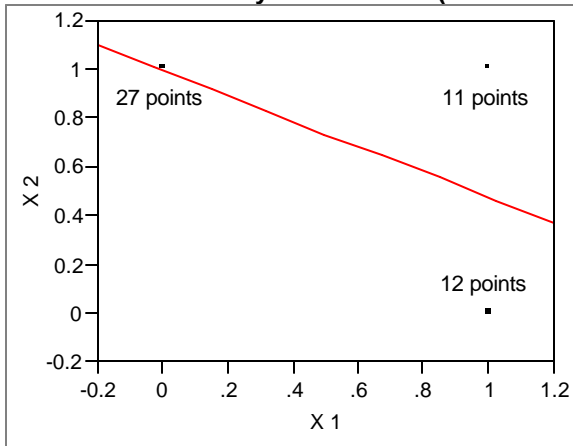
#### Summary of Fit

RSquare 0

#### Parameter Estimates

Term	Estimate
Intercept	0.5
X1	0

**Bivariate Fit of X2 By X1 for Part 3 (Simulated)**



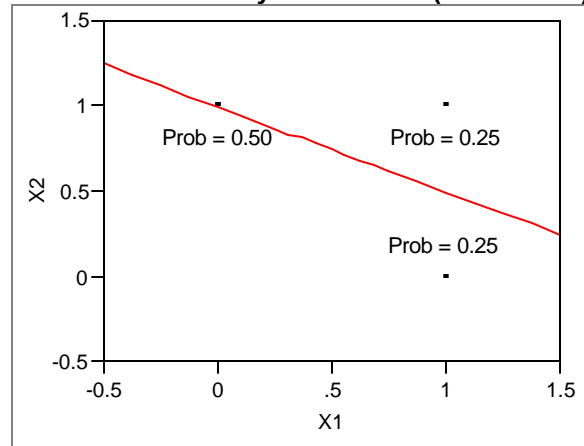
**Linear Fit**  
 $X2 = 1 - 0.5217391 X1$

**Summary of Fit**  
 RSquare 0.370709

**Parameter Estimates**

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	1	0.066546	15.03	<.0001
X 1	-0.521739	0.098116	-5.32	<.0001

**Bivariate Fit of X2 By X1 for Part 3 (Theoretical)**



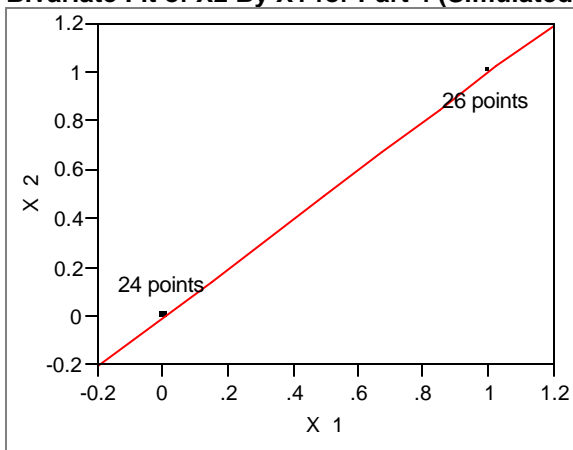
**Linear Fit**  
 $X2 = 1 - 0.5 X1$

**Summary of Fit**  
 RSquare 0.333333

**Parameter Estimates**

Term	Estimate
Intercept	1
X1	-0.5

**Bivariate Fit of X2 By X1 for Part 4 (Simulated)**



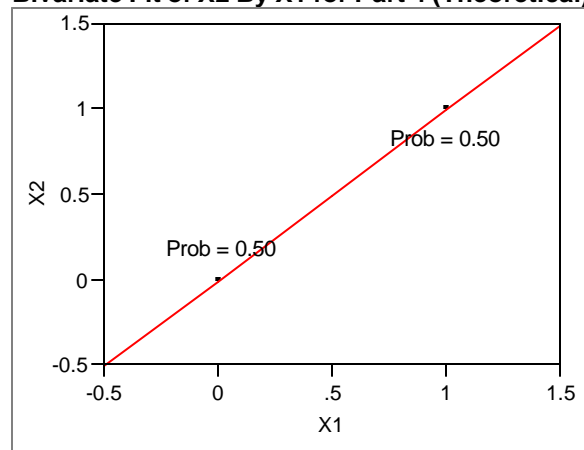
**Linear Fit**  
 $X2 = 0 + 1 X1$

**Summary of Fit**  
 RSquare 1

**Parameter Estimates**

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	0	0	.	.
X 1	1	0	.	.

**Bivariate Fit of X2 By X1 for Part 4 (Theoretical)**



**Linear Fit**  
 $X2 = 0 + 1 X1$

**Summary of Fit**  
 RSquare 1

**Parameter Estimates**

Term	Estimate
Intercept	0
X1	1

In the simulated data for Part 1, the  $p$ -value from a test of significance for slope is 0.6235. This is very high and suggests that there is not a significant relationship between  $X_1$  and  $X_2$ . An examination of the frequencies of the points does not reveal a trend. The theoretical distribution shows no relationship between  $X_1$  and  $X_2$ .

The simulated data for Part 3 have a  $p$ -value less than 1 in 10,000. This strongly suggests a linear relationship between  $X_1$  and  $X_2$ . Dependence is apparent from the theoretical coefficient of determination,  $r^2 = \frac{1}{3}$ . (There is a linear relationship between relationship between  $X_1$  and  $X_2$  because  $r^2 \neq 0$ .)

Looking at the results from the Part 4 simulation, we see that  $X_1$  and  $X_2$  are perfectly correlated, as they were defined in the simulation. The theoretical results are equivalent.

### Some final commentary:

The mean of the sum/difference of random variables will always equal the sum/difference of the individual means regardless of whether the random variables are independent.

The variance of the sum and the variance of the difference of random variables will equal the sum of the individual variances if the random variables are independent.

The rule we explored is for means and variances, not standard deviations. The standard deviation of the sum/difference of random variables can be calculated by taking the square root of the sum of the variances *only* if the random variables are independent.

Technical point: The variance of the sum/difference of two random variables  $X_1$  and  $X_2$  can be calculated whether the variables are independent or not. The formula for the variance of the sum/difference is

$$\text{Var}(X_1 \pm X_2) = \text{Var}(X_1) + \text{Var}(X_2) \pm 2\text{Cov}(X_1, X_2),$$

where the covariance between  $X_1$  and  $X_2$  is  $\text{Cov}(X_1, X_2) = s_{x_1} s_{x_2} \text{Corr}(X_1, X_2)$  and  $\text{Corr}(X_1, X_2)$  is the correlation between  $X_1$  and  $X_2$ .

If  $X_1$  and  $X_2$  are independent, then  $\text{Cov}(X_1, X_2) = 0$ . (The converse is not always true.)

This is beyond the scope of AP Statistics and teachers are encouraged to explore this subject further as they desire.

The observations made in this exploration have direct connections to inference. When comparing means of two populations, we usually have a choice between two tests/confidence intervals: (1) Difference between two means using independent samples and (2) difference between two means using paired samples. We use the paired test when pairs of data are correlated in some meaningful way. In both cases, the statistics of interest are the difference of the means or the mean of the difference, which are algebraically equivalent.

For the difference between two means using independent samples, the sampling distribution of  $\bar{x}_1 - \bar{x}_2$  has mean  $\mathbf{m}_1 - \mathbf{m}_2$  and variance  $\frac{\mathbf{S}_1^2}{n_1} + \frac{\mathbf{S}_2^2}{n_2}$ . The variance of the sampling distribution of  $\bar{x}_1 - \bar{x}_2$  happens to be the sum of the individual variances of the sampling distributions of  $\bar{x}_1$  and  $\bar{x}_2$ , just like the rule we have been exploring says! Notice also that the mean seems to fit our rule.

For paired samples, we have no such combination of the variances. After finding the differences for all the paired data, the sampling distribution of  $\bar{x}_d$  still has mean  $\mathbf{m}_1 - \mathbf{m}_2$  (in this case equivalent to  $\mathbf{m}_d$ ) as with independent samples. But the variance of the sampling distribution  $\frac{\mathbf{S}_d^2}{n}$  is *not* equal to the sum of the variances of  $\bar{x}_1$  and  $\bar{x}_2$ .

Hopefully this will help students understand better when to use which test for differences of means, two independent samples or paired samples.