Statistical Testing II & Multiple Random Variables

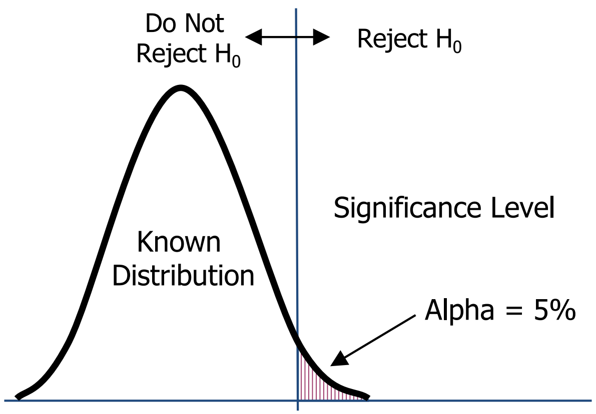
What we will cover in this lesson

- Inference Testing
  - Hypothesis Testing – a little more formal
  - Examples of Hypothesis Tests
  - Chi-Square Tests
- Multiple Random Variables
  - Covariance and Correlations
  - Linear Combinations
  - Sums of Random Variables

Hypothesis Testing – More Formally

Hypothesis Testing

- Three possible hypotheses or outcomes to a test
  1. Unknown distribution is the same as the known distribution (Always H_0)
  2. Unknown distribution is 'higher' than the known distribution
  3. Unknown distribution is 'lower' than the known distribution
- Types of tests that we perform

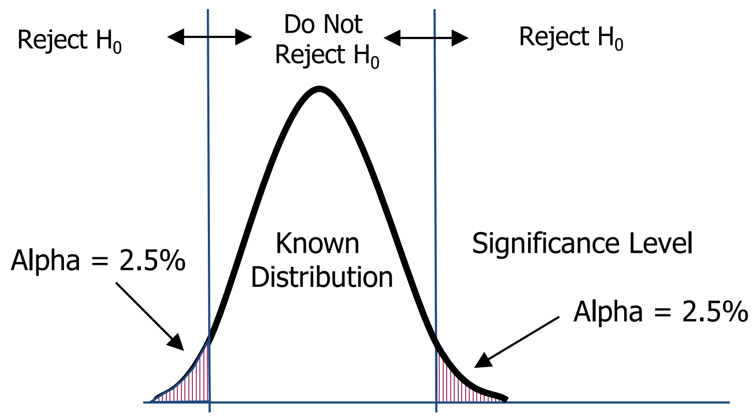


One Tailed Predictions

(test for direction)

H_1 is (2) Unknown distribution is 'higher' than the known

H_1 is (3) Unknown distribution is 'lower' than the known



Two Tailed Predictions

(test for any difference)

H_1 is either (2) or (3)

Alpha value (significance) is divided on each side of the distribution

Error Types

Suppose we are testing whether the MoonDoe's NextGen stores have improved sales. We know the current average sales per customer from traditional stores, μ, as well as the mean and standard deviation ($\overline{x}$, s) of a sample we took of n customers.

Null & Alternative hypotheses

H_0: $\overline{x}_{NextGen} \leq \mu_{Current}$ *(NextGen sales are not greater than current sales)*

H_1: $\overline{x}_{NextGen} > \mu_{Current}$ *(NextGen sales are greater than current sales)*

There are four possible outcomes from this test:

a. We reject H_0, and in fact H_0 was false (We think sales increased and they actually did)

b. We fail to reject H_0, and in fact H_0 was true (We think sales don't increase and they didn't)

Type I c. We reject H_0, but in fact H_0 was true (We think sales increased, but they didn't)

Type II d. We fail to reject H_0, but in fact H_0 was false (We think sales don't increase, but they did)

False Positive

		Actual Outcome	
		H_0 is True	H_0 is False
Hypothesis Test Result	Reject H_0	Type I Error	OK
	Fail to Reject H_0	OK	Type II Error

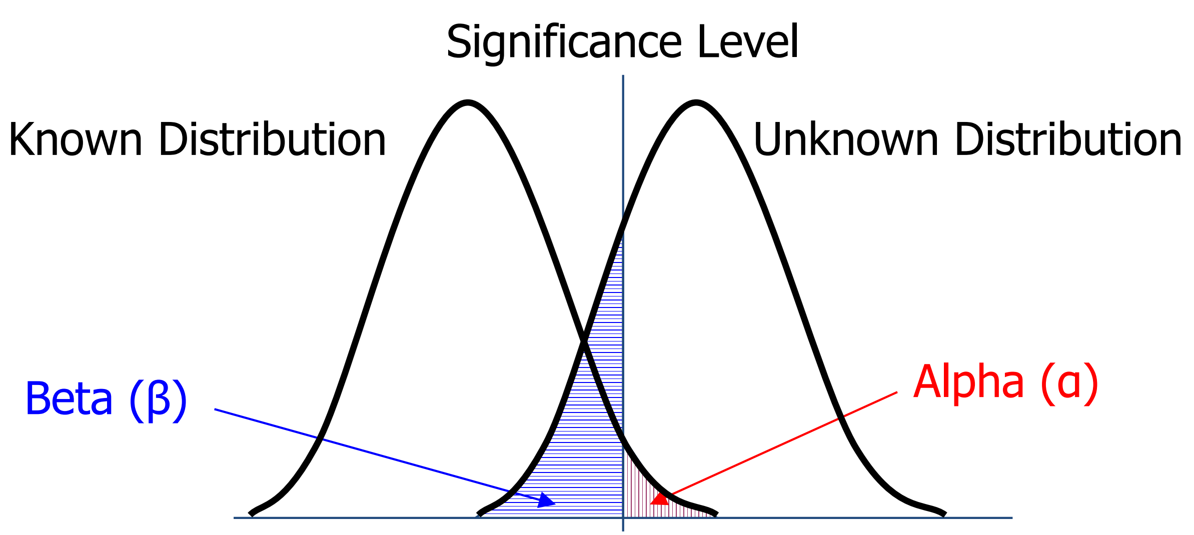
False Negative

When we reject the Null hypothesis, we are essentially saying the sample belongs to a different population, at significance level α.

Error Types

		Actual Outcome	
		H_0 is True	H_0 is False
Test Result	Reject H_0	Type I	OK
	Fail to Reject H_0	OK	Type II

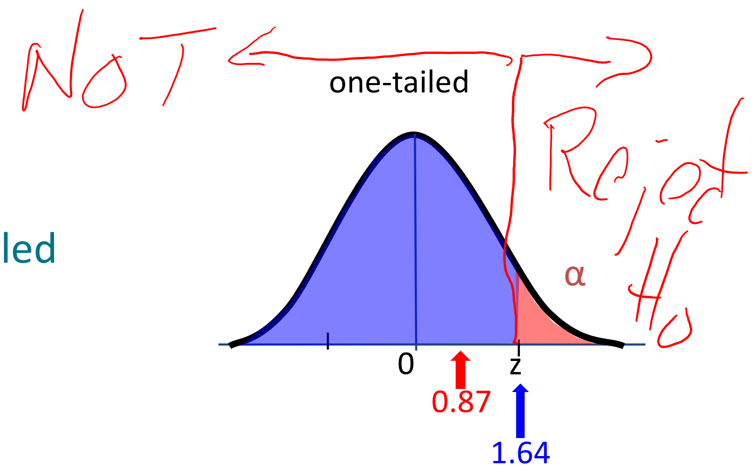
- Each type of mistake maps to a certain probability:
  - Type I: Reject the Null hypothesis when in fact it is True (Alpha) False Positive
  - Type II: Fail to reject the Null hypothesis when in fact it is False (Beta) False Negative
- These probabilities are not independent
  - Reducing the probability of Type I errors (i.e., using a smaller and smaller α value) actually increases the probability of committing Type II errors!!
  - Decide which error type is more important to you.
  - We typically focus on Type I errors when setting significance level (α=0.05, 0.01, etc.)



Hypothesis Testing Examples

Hypothesis Testing: NextGen Stores

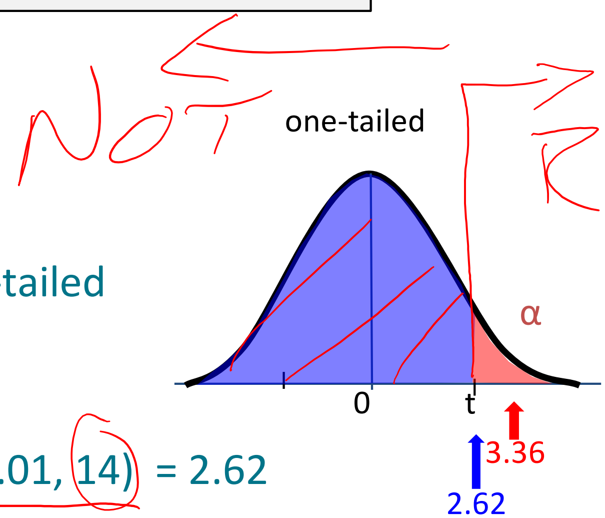
Test whether the average food sales per customer for the NextGen store **is greater than** that of regular cafés. Sampled 50 customers, $\bar{x}_{NextGen}$ =\$13.10 with s=\$6.90 while historically it is \$12.25 per customer visit. We want to be 95% confident of our result.



1. Select the test statistic of interest
   - $\bar{x}_{NextGen}$ and μ using Normal Distribution (n>30)
2. Determine whether this is a one or two tailed test
   - We are testing if NextGen sales are larger, so it is one-tailed
3. Pick your significance level and critical value
   - Set alpha = 5% or 0.05
   - The corresponding z value = NORM.S.INV(1-α)=1.64
4. Formulate your Null & Alternative hypotheses
   - H_0: $\bar{x}_{NextGen} \leq \mu_{Current}$ *(NextGen food sales are not greater than current sales)*
   - H_1: $\bar{x}_{NextGen} > \mu_{Current}$ *(NextGen food sales are greater than current sales)*
5. Calculate the test statistic
   - $z = (\bar{x}_{NextGen} - \mu_{Current})/(s/\sqrt{n}) = (13.10-12.25)/(6.90/\sqrt{50}) = 0.87$
6. Compare the test statistic to the critical value
   - Since 0.87 ≤ 1.64 we cannot reject H_0 that the sales will not be greater.
   - More precisely, since the p-value =1 - norm.s.dist(.87,1) = 0.192, we can say that we would only reject the Null hypothesis at the 81% confidence level. Not so confident!

Hypothesis Testing: NextGen Stores

You discover that customers at NextGen stores tend to drink more coffee but not buy more food than at regular cafes. Test whether the average coffee sales per customer for the NextGen store **is greater than** that of regular cafés. You could only sample 15 customers, $\overline{x}_{NextGen}$ =\$4.80 with s=\$1.90 while historically it is \$3.15 per customer visit. We want to be 99% confident of our result.



1. Select the test statistic of interest
   - $\overline{x}_{NextGen}$ and μ using t-Distribution (n<30)
2. Determine whether this is a one or two tailed test
   - We are testing if NextGen coffee sales are larger, so it is one-tailed
3. Pick your significance level and critical value
   - Set alpha = 1% or 0.01
   - The corresponding t value = T.INV.2T(2*0.01, 14) = T.INV(1-0.01, 14) = 2.62
4. Formulate your Null & Alternative hypotheses
   - H_0: $\overline{x}_{NextGen} \leq \mu_{Current}$ *(NextGen coffee sales are not greater than current sales)*
   - H_1: $\overline{x}_{NextGen} > \mu_{Current}$ *(NextGen coffee sales are greater than current sales)*
5. Calculate the test statistic
   - $t = (\overline{x}_{NextGen} - \mu_{Current})/(s/\sqrt{n}) = (4.80-3.15)/(1.90/\sqrt{15}) = 3.36$
6. Compare the test statistic to the critical value
   - Since 3.36 > 2.62 we reject H_0 that the coffee sales will not be greater.
   - More precisely, since the p-value =1 - t.dist(3.36, 14, 1) = 0.002, we can say that we reject the Null hypothesis at 99.8% confidence! This is typically reported as p-value≤0.01.

Chi-Square Tests

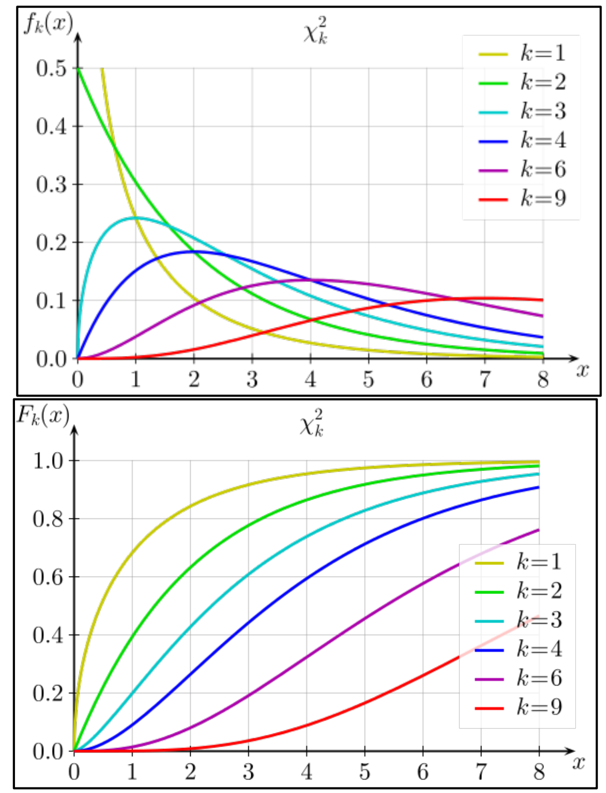
Example: Promotional Cups

MoonDoe is launching a new set of colored cups for a promotion in order to boost sales: Blue, Red, Green, and Yellow. The marketing team expects all of the colors to be equally favored by the customers. In order to test this, you surveyed 100 customers and the results are: Blue 40, Red 22, Green 24, and Yellow 14. Was their hypothesis correct at the 1% level of significance?

- We want to compare two distributions and answer the question, “Does my actual observed distribution ‘fit’ the proposed or expected one?”
- We will use the chi-square (χ^2) distribution
- Assuming your distribution is divided into c categories, then:

$$\chi^2 = \sum_{i=1}^{c}\left(\frac{\left(\#Observed_i - \#Expected_i\right)^2}{\#Expected_i}\right) \qquad df = c - 1$$

Function	Microsoft Excel	LibreOffice->Calc
Returns the p-value for Chi-Square Test	=CHISQ.TEST(obs_values, exp_values)	=CHISQ.TEST(obs_values; exp_values)
Returns the X^2 statistic	=CHISQ.INV(1-α, df)	=CHISQ.INV(1-α, df)
Returns the left-tailed cumulative distribution	=CHISQ.DIST(X^2, df, 1)	=CHISQ.DIST(X^2, df, 1)



Source: By Geek3 - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=9884225

Example: Promotional Cups

MoonDoe is launching a new set of colored cups for a promotion in order to boost sales: Blue, Red, Green, and Yellow. The marketing team thinks that all of the colors will be equally favored by the customers. In order to test this, you surveyed 100 customers and the results are: Blue 40, Red 22, Green 24, and Yellow 14. You want to be 99% confident of your result.

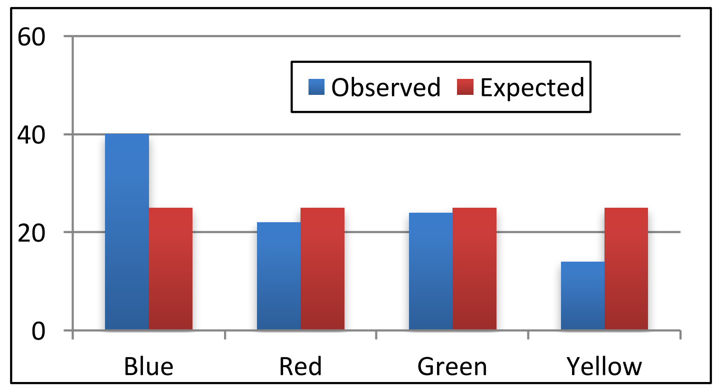
1. Select the test statistic of interest
   - Comparing distributions – so χ^2 statistic
2. Select number of categories
   - 4 colors so 4 categories –degrees of freedom = c-1 = 3
3. Pick your significance level and critical value
   - Set alpha = 1% or 0.01
   - The corresponding χ^2 value = CHISQ.INV(0.99,3) = 11.34
4. Formulate your Null & Alternative hypotheses
   - H_0: All colors are equally favored
   - H_1: Colors are not equally favored
5. Calculate the test statistic
   - χ^2 = 14.24

$$\chi^2 = \sum_{i=1}^{c}\left(\frac{(\#Observed_i - \#Expected_i)^2}{\#Expected_i}\right)$$

6. Compare the test statistic to the critical value
   - Since 14.24 > 11.34 we **can reject** H_0 that the colored cups are equally favored!
   - The p-value= 1 - CHISQ.DIST(14.24,3,1) = 1- 0.997 = 0.003

	Observed	Expected	(o-e)²/e
Blue	40	25	9.00
Red	22	25	0.36
Green	24	25	0.04
Yellow	14	25	4.84

X^2=14.24

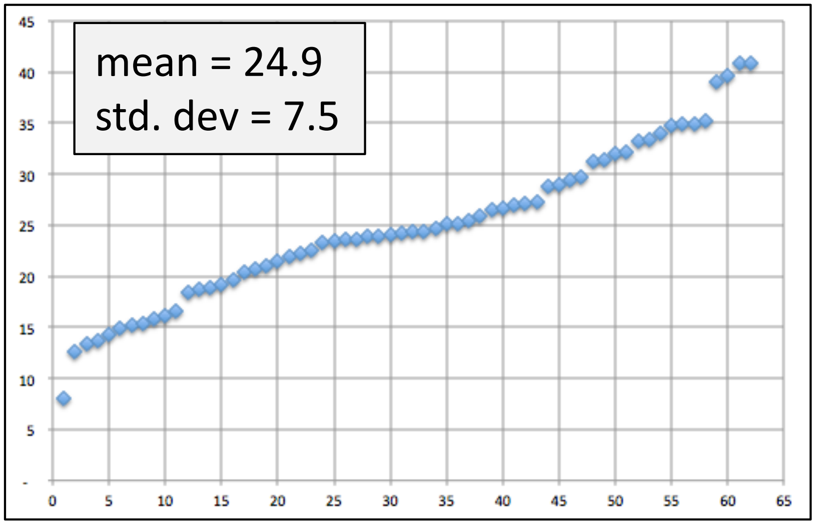


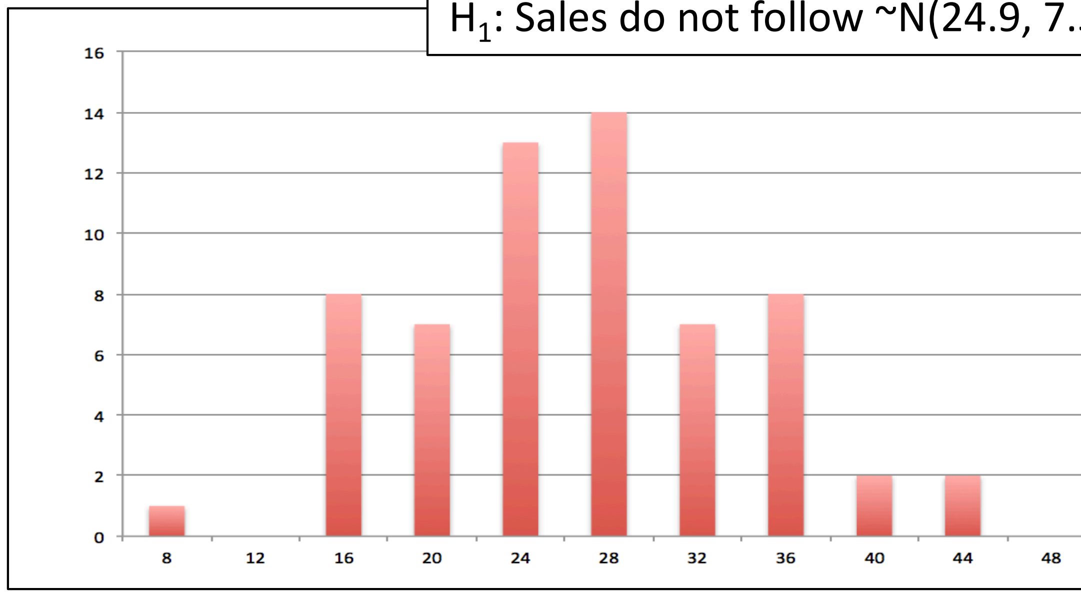
Chi-Square Test Example II

Are my Sales Normally Distributed?

You are a MoonDoe store manager trying to forecast demand and set inventory replenishment policies. You have sales for the last 62 days and would like to see if you can assume that it is distributed Normally at 95% confidence.

H_0: Sales follow ~N(24.9, 7.5)
H_1: Sales do not follow ~N(24.9, 7.5)





We first need to create categories, lets say buckets of 4, so c=11 and k=c-1=10

$$\chi^2 = \sum_{i=1}^{c} \left(\frac{(\#Observed_i - \#Expected_i)^2}{\#Expected_i} \right)$$

$\chi^2 = 7.92$

Range<	Observed	Expected
8	1	0.7
12	0	1.9
16	8	4.6
20	7	8.6
24	13	12.1
28	14	12.9
32	7	10.4
36	8	6.4
40	2	3.0
44	2	1.0
>44	0	0.3

How would I determine what my expected number to be in each bucket?

If ~N(24.9, 7.5) then we expect to have:
=(NORM.DIST(12, 24.9, 7.5,1)
-NORM.DIST(8, 24.9, 7.5,1)) * (# obs)

Critical Value =CHISQ.INV(.95,10) = 18.3

Since 7.92 < 18.3 we CANNOT reject H_0 that the sales ~N(24.9, 7.5)!

Are my Sales Normally Distributed?

You are a MoonDoe store manager trying to forecast demand and set inventory replenishment policies. You have sales for the last 62 days and would like to see if you can assume that it is distributed Normally at 95% confidence.

H_0: Sales follow ~N(24.9, 7.5)
H_1: Sales do not follow ~N(24.9, 7.5)

Suppose we use buckets of 5 (c = 9) and so k=8

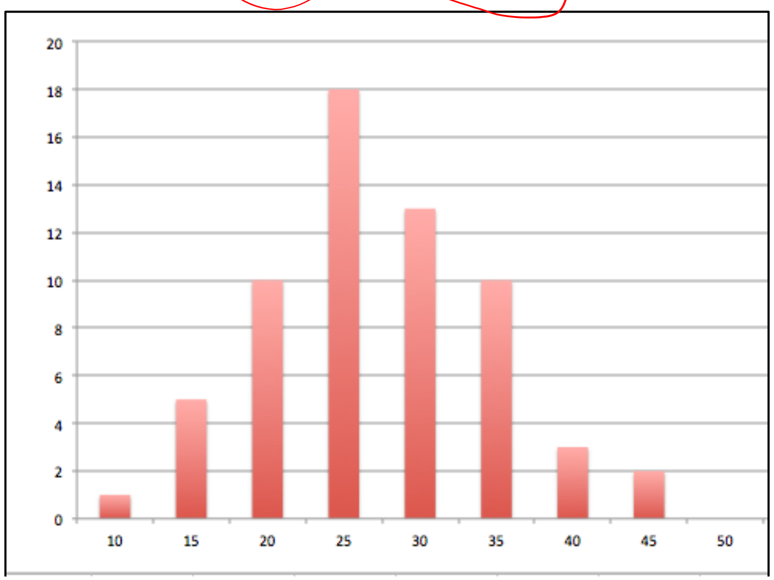
$$\chi^2 = \sum_{i=1}^{c} \left(\frac{(\#Observed_i - \#Expected_i)^2}{\#Expected_i} \right)$$

Critical Value =CHISQ.INV(.95,8) = 15.5

Range<	Observed	Expected
10	1	1.4
15	5	4.3
20	10	10.1
25	18	15.4
30	13	15.3
35	10	9.9
40	3	4.2
45	2	1.1
50	0	0.2

X^2 =2.21

Since 2.21 < 15.5 we CANNOT reject H_0 that the sales ~N(24.9, 7.5)!

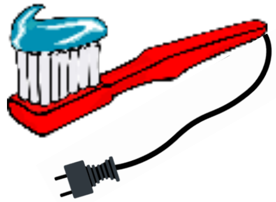


Notes:

- Smaller χ^2 values means it is closer to assumed distribution (χ^2=0 means it is identical)
- We can also use the =CHISQ.TEST() function that returns the p-value
- Buckets should always be equal sized with a rule of thumb of >5 unit per category
- Can be somewhat arbitrary – based on the segmentation
- Other tests available in statistics packages

Multiple Random Variables

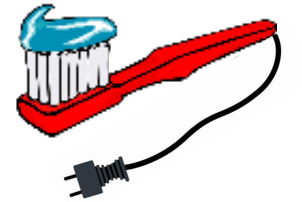
Zippy Bright Floss

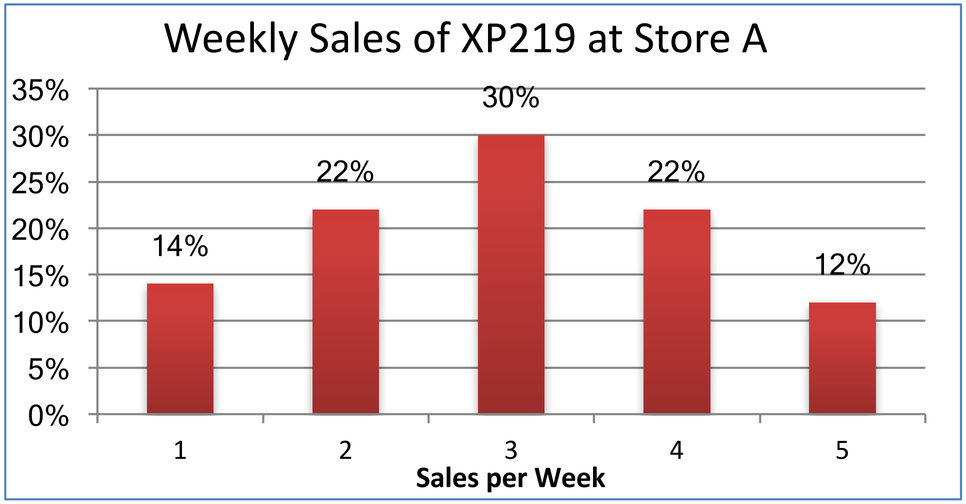


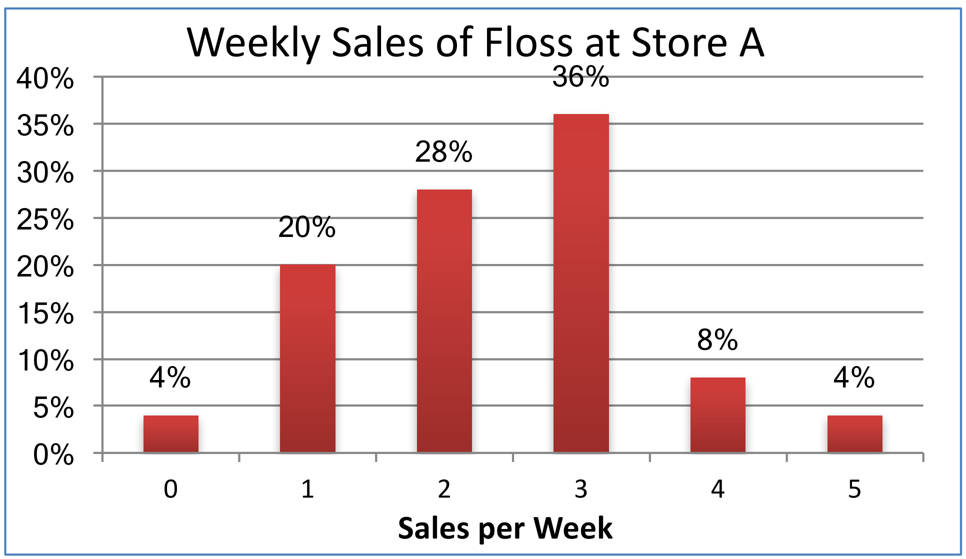
- In addition to electric toothbrushes, Zippy Bright manufactures other dental products such as dental floss. Marketing is interested in better understanding the sales of floss. They want to see if there is any connection to sales of the XP219. The table below lists a year's worth of weekly sales at a single store for both products. What can we say?

Week #	XP219	Floss	Week #	XP219	Floss	Week #	XP219	Floss	Week #	XP219	Floss	Week #	XP219	Floss
1	1	1	11	3	3	21	2	2	31	1	1	41	2	2
2	5	5	12	2	1	22	4	3	32	2	2	42	3	3
3	3	3	13	3	3	23	4	3	33	3	2	43	3	2
4	2	2	14	4	4	24	3	2	34	4	3	44	3	2
5	3	3	15	2	2	25	4	3	35	5	3	45	4	3
6	3	2	16	1	0	26	1	0	36	5	3	46	1	1
7	3	3	17	3	3	27	2	2	37	1	1	47	2	1
8	2	1	18	4	3	28	3	2	38	5	5	48	3	2
9	5	4	19	4	3	29	4	3	39	5	4	49	4	4
10	2	1	20	3	2	30	4	3	40	1	1	50	2	1

Zippy Bright Floss – Summary Statistics



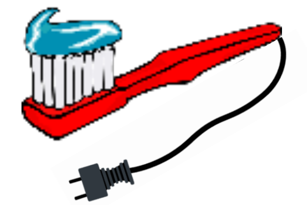


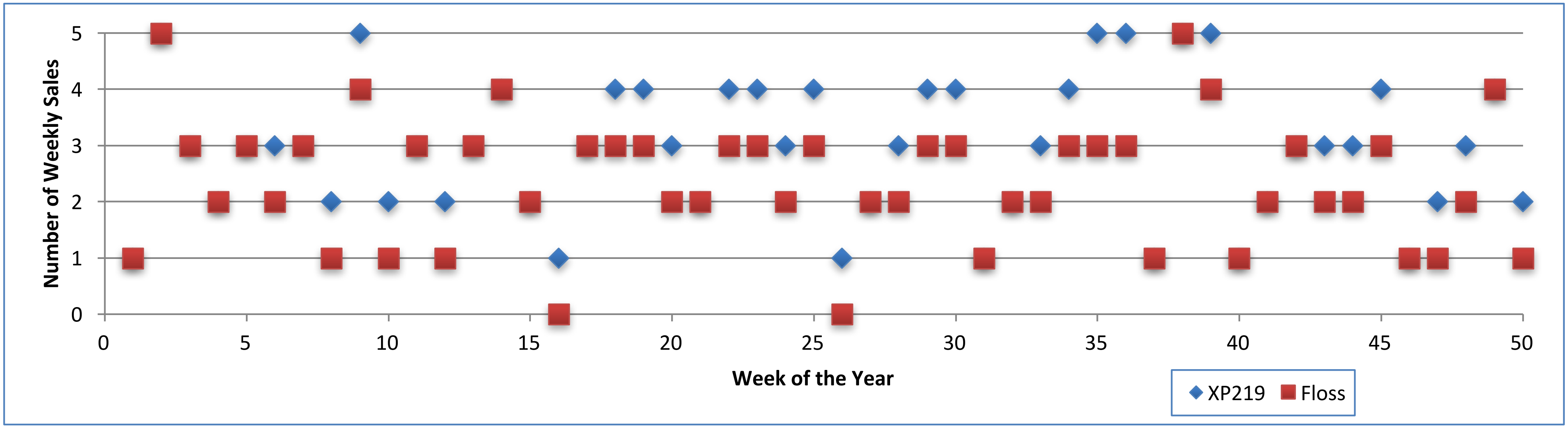


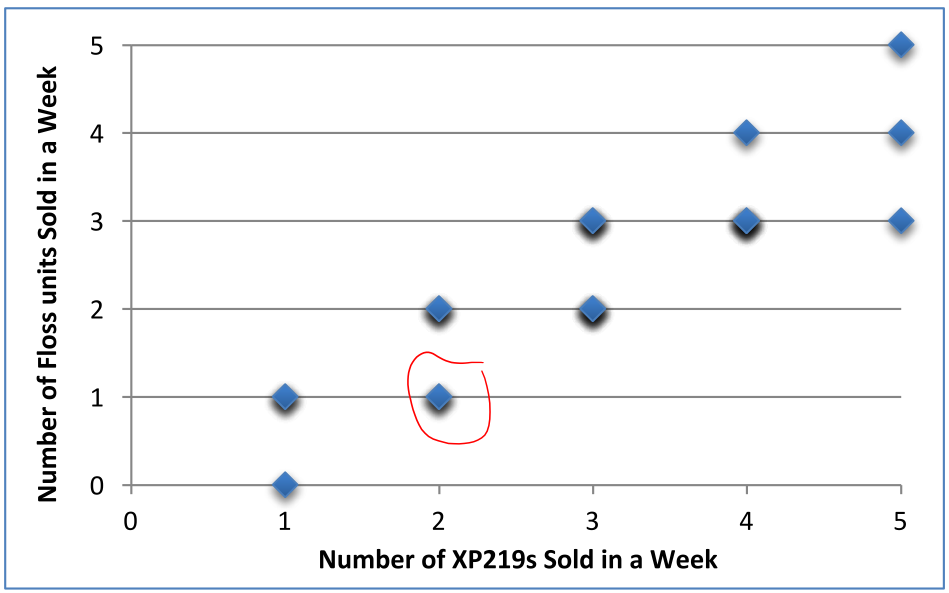
	XP219	Floss
Minimum	1.00	-
Median	3.00	2.00
Mode	3.00	3.00
Mean	2.96	2.36
Max	5.00	5.00
Range	4.00	5.00
Std Dev (population)	1.216	1.127
Variance (population)	1.478	1.270
Coefficient of Variation	0.41	0.48

Is there any relationship between sales of the XP219 and Dental Floss?

Zippy Bright: XP219 & Floss







XP219	Floss	Probability
1	0	4%
1	1	10%
2	1	10%
2	2	12%
3	2	16%
3	3	14%
4	3	18%
4	4	4%
5	3	4%
5	4	4%
5	5	4%

How can I measure the relationship (if any) between these two random variables?

Covariance and Correlation

Covariance and Correlation

Covariance – a measure of how two variables change together.

$$\mathrm{Cov}(X,Y) = E[\ (X - \mu_X\ )\ (Y - \mu_Y\ )\ ]$$

$$= \sum_{i=1}^{n} P(X = x_i, Y = y_i)[(x_i - \mu_X)(y_i - \mu_Y)]$$

X=XP219	Y=Floss	Probability	$(x_i-\mu_X)$	$(y_i-\mu_Y)$	$P()(x_i-\mu_X)(y_i-\mu_Y)$
1	0	4%	-1.96	-2.36	0.185
1	1	10%	-1.96	-1.36	0.267
2	1	10%	-0.96	-1.36	0.131
2	2	12%	-0.96	-0.36	0.041
3	2	16%	0.04	-0.36	-0.002
3	3	14%	0.04	0.64	0.004
4	3	18%	1.04	0.64	0.120
4	4	4%	1.04	1.64	0.068
5	3	4%	2.04	0.64	0.052
5	4	4%	2.04	1.64	0.134
5	5	4%	2.04	2.64	0.215
μ_X=2.96	μ_Y = 2.36				Σ=1.214

COV(X,Y) = 1.214

Alternatively, if we have n observations, I can find the covariance using:

$$\mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{n}$$

But how should I interpret this? How can I compare this?

Covariance and Correlation

Correlation Coefficient – a measure between -1 and +1 that indicates the degree and direction of the relationship between two random numbers. .

$$\text{CORR(X,Y)} = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$$

$$\text{CORR(X,Y)} = \frac{\sum_{i=1}^{n} \text{P}\left(\text{X=x}_\text{i}, \text{Y=y}_\text{i}\right)\left[(\text{x}_\text{i} - \mu_X)(\text{y}_\text{i} - \mu_Y)\right]}{\sigma_X \sigma_Y}$$

COV(X,Y) = 1.214

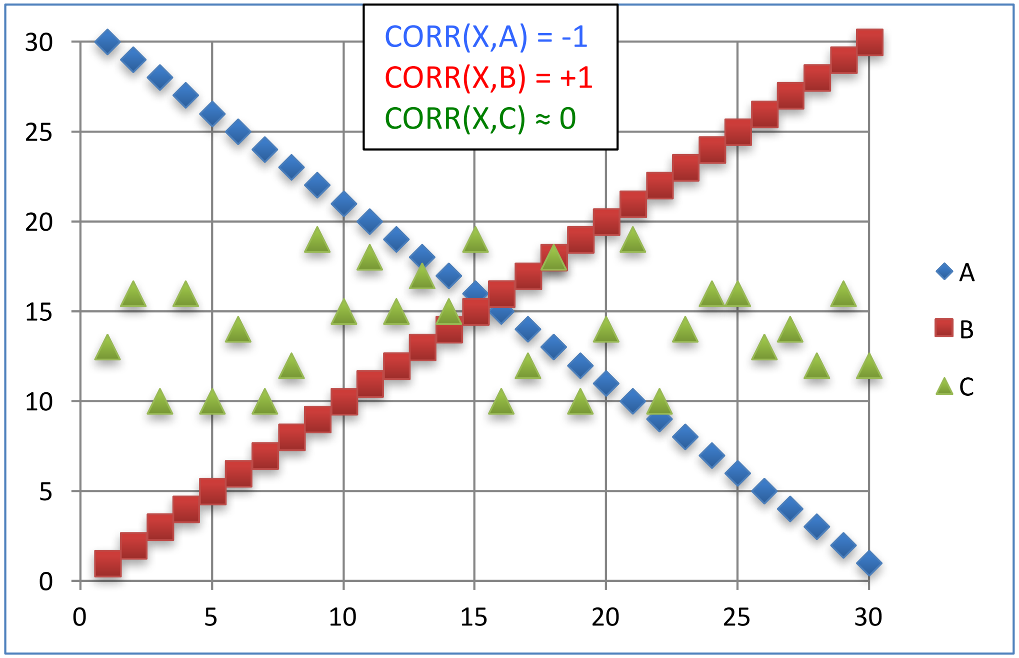
$\sigma_X = 1.23$

$\sigma_Y = 1.14$

CORR(X,Y) =

= 1.214/(1.23)(1.14)

= 0.868



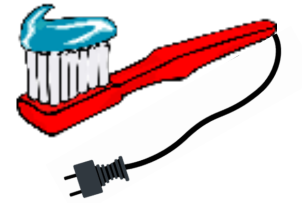
Covariance and Correlation

- Positively Correlated Random Variables:
  - If **higher** than average values of X are apt to occur with **higher** than average values of Y, then COV(X, Y) > 0 and CORR(X, Y) > 0.
- Negatively Correlated Random Variables:
  - If **higher** than average values of X are apt to occur with **lower** than average values of Y, then COV(X, Y) < 0 and CORR(X, Y) < 0.
- Correlation is not the same as causality!
  - We are finding a mathematical relationship – not a causal one!
  - We do not know which way (if any) the relationship works!
- Related to Independence of the two Random Variables
  - If X and Y are independent random variables, then COV(X,Y) = 0 and CORR(X,Y) = 0

Function	Microsoft Excel	Google Sheets	LibreOffice->Calc
Covariance	=COVAR(array,array)	=COVAR(array,array)	=COVAR(array;array)
Correlation	=CORREL(array,array)	=CORREL(array, array)	=CORREL(array;array)

Linear Function of Random Variables

Zippy Bright – Retail Fees



- Suppose that in order to keep the XP219 on the shelves and prominently displayed, Zippy Bright has agreed to pay the retailer a "shelving fee." The retailer charges this fee as both a fixed cost per week ($75) and a per unit sold cost ($1.50). The fees are paid to the retailer in order to keep the products in high traffic locations in the store, such as "end cap" displays.
- Zippy Bright is trying to get a handle on these fees, specifically:
    - What are the expected value, the variance, and the standard deviation for weekly retailer fees?

Linear Functions of Random Variables

- Suppose that **Y = aX + b**
  - where
    - X, Y are random variables
    - a, b are constants

Then;

- $E[Y] = \mu_Y = a\mu_X + b$
- $VAR[Y] = \sigma^2_Y = a^2\sigma^2_X$
- $\sigma_Y = |a|\sigma_X$

Note that the *a* value "scales" or changes the distribution while the *b* value simply "shifts" the distribution.
Scaling impacts the variability while shifting does not!

For X = # of Electric Toothbrushes sold/week, a= 1.50 \$/item, and b = 75 \$/week:

Y = Retail Fees = 1.50 X + 75

$\mu_Y = a\mu_X + b = (1.50)(2.96) + 75 = \mathbf{79.44}$ \$/week

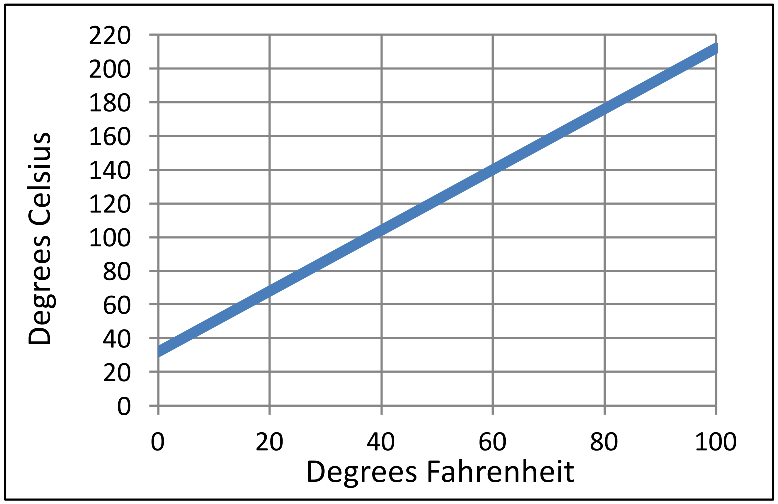
$\sigma^2_Y = a^2\sigma^2_X = (1.5)^2(1.216)^2 = \mathbf{3.327\ \$^2}/\text{weeks}^2$

$\sigma_Y = |a|\sigma_X = (1.50)(1.216) = \mathbf{1.824}$ \$/week

X=# sold	Probability	$y_i=ax_i+b$	$y_i \times P[Y=y_i]$	$(y_i-\mu_y)$	$(y_i-\mu_y)^2$	$P[Y=y_i]\times(y_i-\mu_y)^2$
1	14%	76.5	10.71	(2.940)	8.644	1.210
2	22%	78.0	17.16	(1.440)	2.074	0.456
3	30%	79.5	23.85	0.060	0.004	0.001
4	22%	81.0	17.82	1.560	2.434	0.535
5	12%	82.5	9.90	3.060	9.364	1.124
			$\boldsymbol{\mu_Y = 79.44}$			$\boldsymbol{\sigma^2_Y = 3.326}$
						$\boldsymbol{\sigma_Y = 1.824}$

Linearity and Correlation

- Linearly related random variables have perfect correlation
  - If Y=aX+b, with a>0 then Corr(X,Y)=1
  - If Y=aX+b, with a<0 then Corr(X,Y)= -1
- Example: Converting temperatures
  - X= temperature in Fahrenheit,
  - Y= temperature in Celsius
  - So that, X = (9/5)Y + 32



- Checking with Zippy Bright retailer fees

X=#sold	Y=Fees	Probability	$(x_i-\mu_X)$	$(y_i-\mu_Y)$	$P()(x_i-\mu_X)(y_i-\mu_Y)$
1	76.5	4%	(1.960)	(2.940)	0.807
2	78.0	10%	(0.960)	(1.440)	0.304
3	79.5	10%	0.040	0.060	0.001
4	81.0	12%	1.040	1.560	0.357
5	82.5	16%	2.040	3.060	0.749

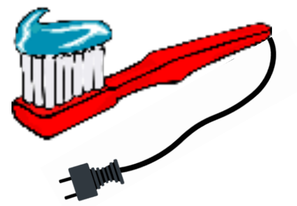
μ_X=2.96	μ_Y = 79.44
σ_X=1.216	σ_Y = 1.824

COV=Σ=2.218

CORR=2.218/(1.216)(1.824) = 1.00

Multiple Random Variables: Sums

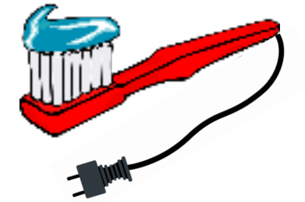
Zippy Bright – Weekly Revenue



- Zippy Bright is interested in determining the expected revenue and its variance for the sales of both the XP219 toothbrushes and the dental floss. Zippy Bright receives $67 for each XP219 sold and $13.50 for each dental floss package. The summary statistics for each of the items are shown in the table below.
- What is Zippy Bright's expected weekly revenue and its variance?

	XP219	**Floss**
Minimum	1.00	-
Median	3.00	2.00
Mode	3.00	3.00
Mean	2.96	2.36
Max	5.00	5.00
Range	4.00	5.00
Std Dev (population)	1.216	1.127
Variance (population)	1.478	1.270
Coefficient of Variation	0.41	0.48
Covariance	1.214	
Correlation	0.868	

Sum of Two Random Variables



Suppose that **$Z = aX + bY$**

- where
  - X, Y, Z are random variables
  - a, b are constants

Then;

- $E[Z] = aE[X] + bE[Y] = a\mu_X + b\mu_Y$
- $VAR[Z] = a^2VAR[X] + b^2VAR[Y] + 2abCOV[X,Y]$
  $= a^2\sigma^2_X + b^2\sigma^2_Y + 2abCOV[X,Y]$
  $= a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma_X\sigma_Y\, CORR[X,Y]$
- If X and Y are independent, then
  $VAR[Z] = a^2VAR[X] + b^2VAR[Y] = a^2\sigma^2_X + b^2\sigma^2_Y$

For Zippy Bright's weekly revenue, we have three Random Variables:

X = # of Electric Toothbrushes sold/week, E[X] = 2.96, VAR[X]=1.478

Y= # Dental Floss units sold/week, E[Y] = 2.36, VAR[Y]=1.270

Z = Revenue for Zippy Bright $/week, E[Z] = ????, VAR[Z]=????

The revenue, Z = aX + bY, where

a = Per item revenue of electric toothbrushes = 67 $/item

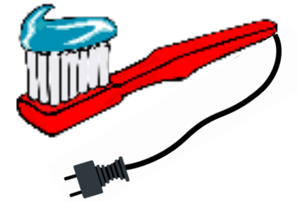
b = Per item revenue of dental floss units = 13.50 $/item

Expected weekly revenue = E[Z] = aE[X] + bE[Y] = (67)(2.96) + (13.50)(2.36) = **230.18 $/week**

Variance of wkly rev =VAR[Z] =a^2VAR[X] + b^2VAR[Y]=67^2(1.478) + 13.50^2(1.270)=**6866 $\$^2/wk^2$**

Standard deviation of weekly revenue = SQRT(VAR[Z]) = SQRT(6866) = **82.86 $/week**

Zippy Bright – Monthly Revenue



- Suppose that we assume that weekly revenue is i.i.d. (independent and identically distributed) with a mean μ_{WR}=230, a variance σ^2_{WR}= 6866, and a standard deviation σ_{WR}=\$82.86. What is the expected value, variance, and standard deviation of monthly and annual revenues?

Recall that if X and Y are independent, then:

- E[X+Y] = E[X] + E[Y]
- VAR[X+Y] = VAR[X] + VAR[Y]
- StdDev[X+Y] = √VAR[X+Y]

For a month (n=4 weeks), then

- E[MonthlyRev] = $\mu_{WR}+\mu_{WR}+\mu_{WR}+\mu_{WR} = n\mu_{WR}$ = 4(230) = **920 \$/month**
- VAR[MonthlyRev] = $\sigma^2_{WR}+\sigma^2_{WR}+\sigma^2_{WR}+\sigma^2_{WR}$+= $n\sigma^2_{WR}$ = 4(6866) = **27,464**
- StdDev[MonthlyRev] = $\sqrt{(n\sigma^2_{WR})} = \sqrt{n}\sigma_{WR}$ = √4(82.86) = **165.72 \$/month**

For a year (n=48 weeks), then

- E[AnnualRev] = 48(230) = **11,040 \$/year**
- StdDev[AnnualRev] = √48(82.86) = **574.07 \$/year**

MIT Center for Transportation & Logistics

Sums of Normal Random Variables

- Let X and Y be independent normal random variables. That is $X \sim N(\mu_X,\sigma_X)$ and $Y \sim N(\mu_Y,\sigma_Y)$. Suppose I have a new random variable, W, where $W = aX + bY$.
  - What is E(W)? $E[W] = \mu_W = a\mu_X + b\mu_Y$
  - What is Var(W)? $VAR[W] = \sigma^2_W = a^2\sigma^2_X + b^2\sigma^2_Y + 2abCOV(X,Y)$
    $= a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma_X\sigma_Y CORR(X,Y)$
  - What is SD(W)? $StdDev[W] = \sigma_W = \sqrt{VAR[W]}$
  - What is the distribution of W?

The weighted sum of independent Normally distributed random variables is itself a Normally distributed random variable!

Key Points from Lesson

Key Points (1/2)

- Hypothesis Testing
  - Method for making a choice between two mutually exclusive and collectively exhaustive alternatives:
    - Null Hypothesis (H_0)
    - Alternative Hypothesis (H_1)
  - There are 2 types of mistakes that we can make:
    - Type I: Reject the Null hypothesis when in fact it is True (Alpha)
    - Type II: Do Not Reject the Null hypothesis when in fact it is False (Beta)
  - Always report the p-value
    - The smallest level of alpha (level of significance) where we would reject H_0
- Chi Square Tests
  - Widely used test for "goodness of fit" to a distribution
  - Compare p-value to desired level of significance.

$$\chi^2 = \sum \left(\frac{(Observed - Expected)^2}{Expected} \right) \qquad df = c - 1$$

Key Points (2/2)

$$\text{Cov(X,Y)} = \frac{\sum_{i=1}^{n}(\text{x}_\text{i} - \mu_X)(\text{y}_\text{i} - \mu_Y)}{n}$$

$$\text{Cov(X,Y)} = \sum_{i=1}^{n}\text{P}(\text{X} = \text{x}_\text{i}, \text{Y} = \text{y}_\text{i})[(\text{x}_\text{i} - \mu_X)(\text{y}_\text{i} - \mu_Y)]$$

$$\text{CORR(X,Y)} = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$$

- Covariance and Correlation
  - $-1 \leq \text{CORR} \leq +1$
  - Indicates direction and magnitude of relation
- Random Variables
  - Linear Function: Y = aX + b
    - $E[Y] = \mu_Y = a\mu_X + b$
    - $VAR[Y] = \sigma^2_Y = a^2\sigma^2_X$
    - $\sigma_Y = |a|\sigma_X$
  - Sum of Random Variables: Z = aX + bY
    - $E[Z] = aE[X] + bE[Y] = a\mu_X + b\mu_Y$
    - $VAR[Z] = a^2VAR[X] + b^2VAR[Y] + 2abCOV[X,Y]$
      $= a^2\sigma^2_X + b^2\sigma^2_Y + 2abCOV[X,Y]$
      $= a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma_X\sigma_Y\, CORR[X,Y]$
    - If X and Y are independent, then COV[X,Y] = CORR[X,Y] = 0 so that:
      $VAR[Z] = a^2VAR[X] + b^2VAR[Y] = a^2\sigma^2_X + b^2\sigma^2_Y$

Questions, Comments, Suggestions? Use the Discussion Forum!

"Dexter testing the hypothesis that someone will take him to Disney World . . . it was rejected."