Review of Quiz 1-5

- Cover representative questions with high false rates only (impossible to cover all questions in 80 min)
- All data files can be downloaded from the course website
- Make sure your computers are in excellent condition for final exam (bringing a spare computer is a good idea).
- Accommodations: please contact Julie Carty and let me know about your start time, end time and section number by April 1st.
- Volunteer for USAT (course evaluation) next class

Quiz 5

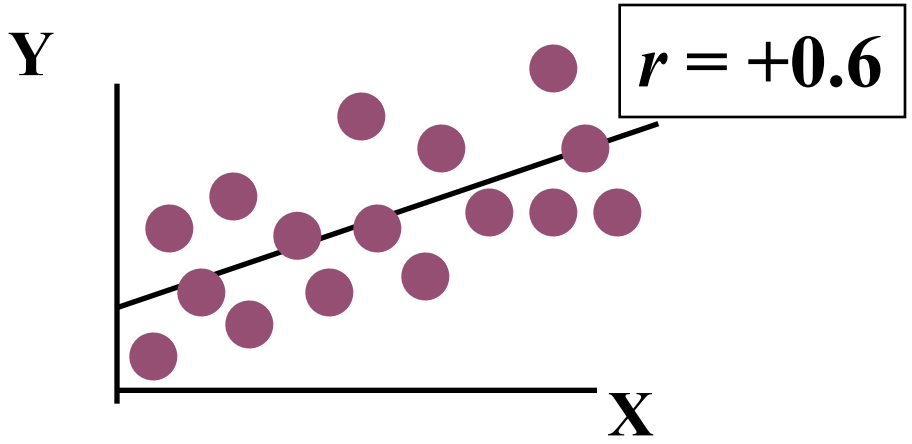
Q1 (Questions 1 - 3)

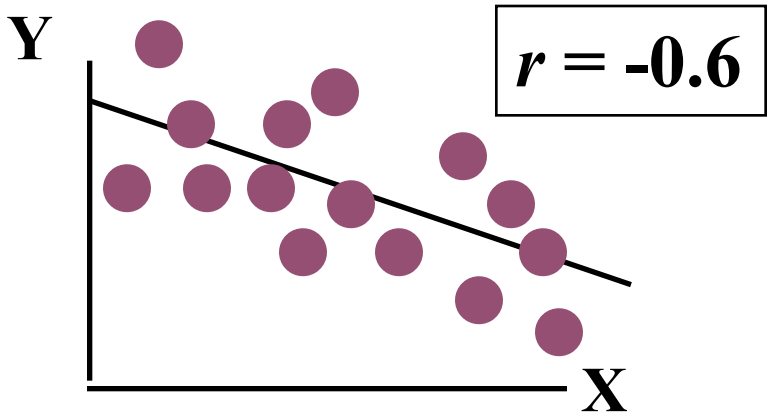
Random Question 1 Difficulty: 1

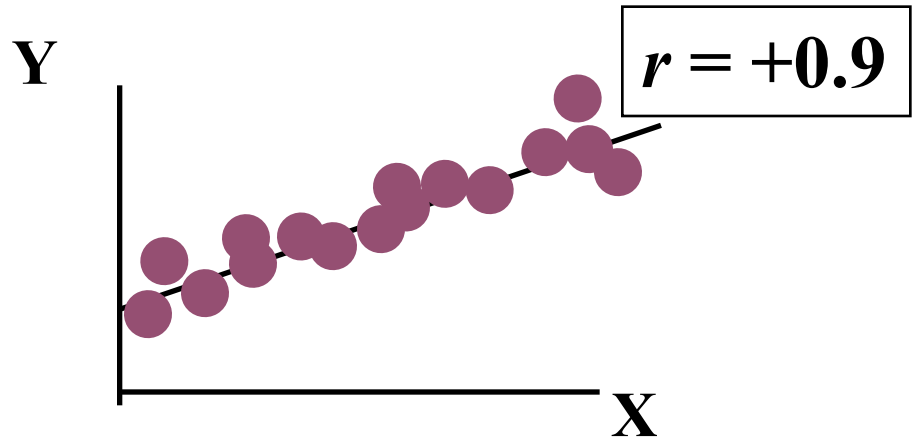
If two variables have a correlation coefficient equal to -0.95, then we know that when one variable decreases, the other variable increases.

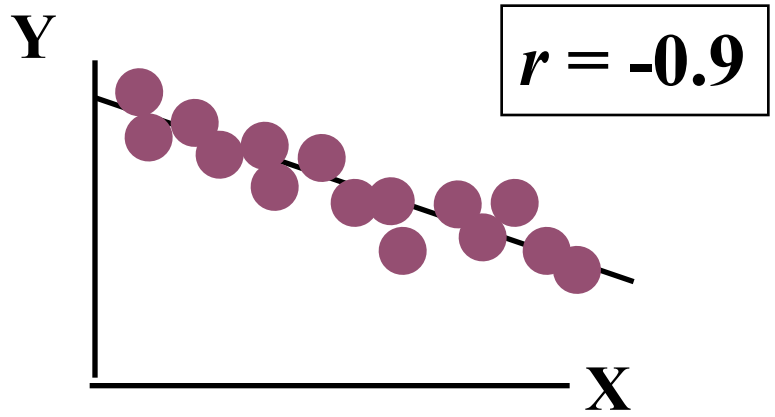
Average Grade: 0.84 / 1 (83.78 %)

➡	True	62	(83.78 %)	Standard Deviation n/a
	False	12	(16.22 %)	Point Biserial n/a
				Discrimination Index n/a









- The sign of r tells the directions (positive/negative)
- The absolute value of r tells the strength.

Random Question 3 Difficulty: 1

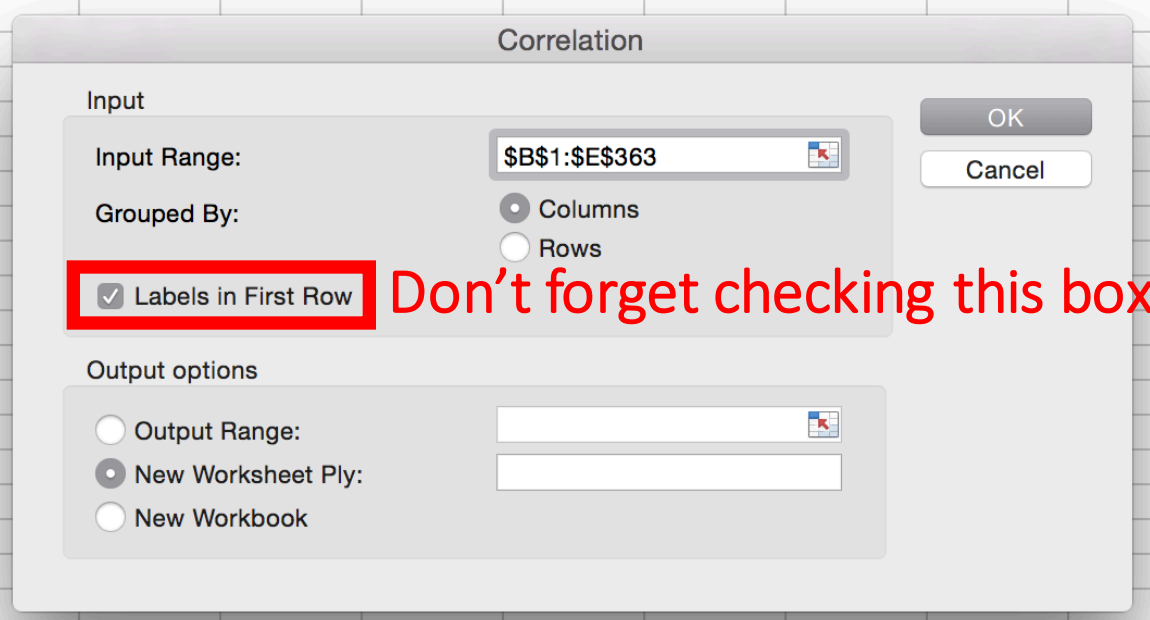
USinflation.xlsx This file contains inflation and unemployment rate data for the United States. CPI denotes inflation in all items contained in the consumer price index; Core denotes inflation in a measure of core consumer prices that excludes energy and other volatile prices from the measurement; and PCE measures inflation in personal consumption expenditures. The unemployment rate is denoted by Unemployment. Out of these three measures (CPI, Core, and PCE), the Unemployment rate is most highly correlated with CPI.

Average Grade: 0.81 / 1 (81.08 %)

➡	True	60	(81.08 %)	Standard Deviation	n/a
	False	14	(18.92 %)	Point Biserial	n/a
				Discrimination Index	n/a

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	DATE	CPI	CORE	PCE	Unemployment									
2	12/01/1983	1.61	2.01	1.58	8.3									
3	01/01/1984	1.82	2.13	1.55	8.0									
4	02/01/1984	1.99	2.08	1.76	7.8									
5	03/01/1984	2.07	2.16	1.85	7.8									
6	04/01/1984	1.93	2.19	1.81	7.7									
7	05/01/1984	1.84	2.23	1.74	7.4									
8	06/01/1984	1.84	2.26	1.66	7.2									
9	07/01/1984	1.83	2.21	1.57	7.5									
10	08/01/1984	1.83	2.24	1.52	7.5									
11	09/01/1984	1.82	2.23	1.43	7.3									
12	10/01/1984	1.81	2.18	1.46	7.4									
13	11/01/1984	1.77	2.05	1.46	7.2									
14	12/01/1984	1.72	2.08	1.55	7.3									
15	01/01/1985	1.5	1.91	1.59	7.3									
16	02/01/1985	1.54	2.02	1.5	7.2									
17	03/01/1985	1.62	2.01	1.53	7.2									
18	04/01/1985	1.53	1.93	1.44	7.3									
19	05/01/1985	1.53	1.92	1.49	7.2									
20	06/01/1985	1.56	1.87	1.53	7.4									
21	07/01/1985	1.48	1.78	1.48	7.4									
22	08/01/1985	1.43	1.78	1.49	7.1									
23	09/01/1985	1.39	1.69	1.51	7.1									
24	10/01/1985	1.38	1.72	1.47	7.1									
25	11/01/1985	1.5	1.88	1.53	7.0									
26	12/01/1985	1.62	1.83	1.56	7.0									
27	01/01/1986	1.69	1.9	1.54	6.7									
28	02/01/1986	1.37	1.78	1.34	7.2									
29	03/01/1986	0.93	1.73	1.07	7.2									
30	04/01/1986	0.68	1.77	0.91	7.1									
31	05/01/1986	0.72	1.68	0.88	7.2									

1R x 4C



A	B	C	D	E
	CPI	*CORE*	*PCE*	*Unemployment*
CPI	1			
CORE	0.64844132	1		
PCE	0.95876999	0.70185161	1	
Unemployment	-0.2744507	-0.072742	-0.161721703	1

Remember: the **absolute value** tells us about the strength!

Q2 (Questions 4 - 5)

Random Question 1 Difficulty: 1

recessions.xlsx This file contains data on the 10 year Government of Canada bond yield (tengoc), the 91 day Treasury Bill rate (91daytb), the US-Canada exchange rate (er), the unemployment rate (unrate), and a dummy variable that indicates whether there is a recession (recess=1) or not (recess=0). A regression of the unemployment rate on a constant and the 91 day Treasury Bill rate and the exchange rate over the period from January 1980 to October 2014, inclusive, leads one to conclude that an increase in the exchange rate lowers the unemployment rate.

Average Grade: 0.93 / 1 (93.24 %)

	True	5	(6.76 %)	Standard Deviation	n/a
➡	False	69	(93.24 %)	Point Biserial	n/a
				Discrimination Index	n/a

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i}$$

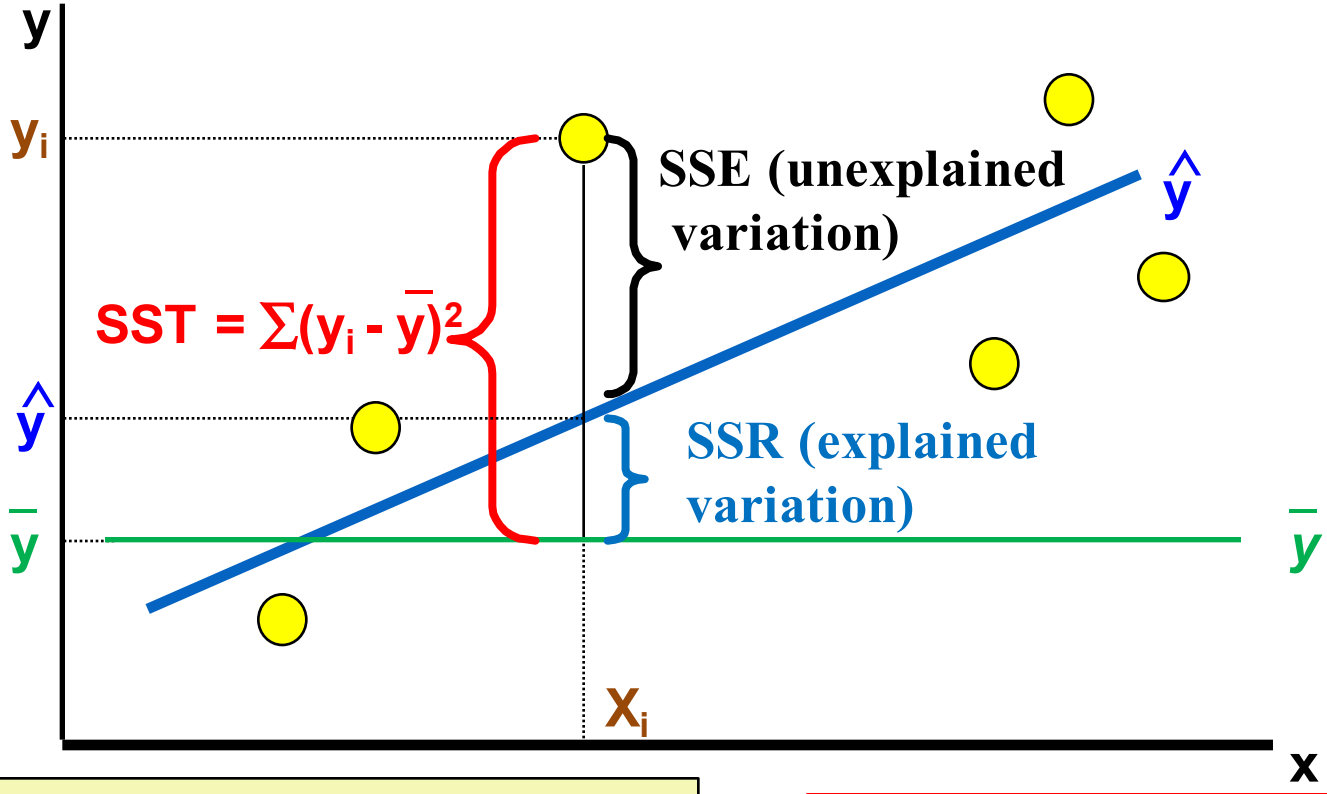
	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.39520239							
5	R Square	0.156184929							
6	Adjusted R Square	0.15211835							
7	Standard Error	1.573028789							
8	Observations	418							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	190.0698735	95.0349367	38.40696133	4.96947E-16			
13	Residual	415	1026.884122	2.47441957					
14	Total	417	1216.953995						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	5.349359175	0.610209676	8.76642797	4.78572E-17	4.149872013	6.54884634	4.14987201	6.54884634
18	91daytb	0.130954028	0.017516905	7.47586569	4.59726E-13	0.096521106	0.16538695	0.09652111	0.16538695
19	er	1.852352594	0.482208535	3.84139321	0.00014146	0.904476856	2.80022833	0.90447686	2.80022833
20									

Don’t forget checking P-value (it must be < alpha)

recessions.xlsx This file contains data on the 10 year Government of Canada bond yield (tengoc), the 91 day Treasury Bill rate (91daytb), the US-Canada exchange rate (er), the unemployment rate (unrate), and a dummy variable that indicates whether there is a recession (recess=1) or not (recess=0). A regression of the unemployment rate on a constant and the 91 day Treasury Bill rate over the period from January 1980 to October 2014, inclusive, gives an R-squared value of _____ (to four decimal points).

A	B	C	D	E	F	G	H	I	J
SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.355219788								
R Square	0.126181098								
Adjusted R S	0.124080572								
Standard Err	1.598825744								
Observations	418								
ANOVA									
	df	*SS*	*MS*	*F*	*Significance F*				
Regression	1	153.5565915	153.556591	60.07118489	7.08555E-14				
Residual	416	1063.397404	2.55624376						
Total	417	1216.953995							
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	
Intercept	7.639355966	0.132430562	57.6857476	1.3922E-200	7.379039474	7.89967246	7.37903947	7.89967246	
91daytb	0.137364666	0.017723192	7.7505603	7.08555E-14	0.102526491	0.17220284	0.10252649	0.17220284	

R-sqare: Percentage of Explained Variation



$$\text{SSE} = \text{Sum of Squares Error} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

$$\text{SSR} = \text{Sum of Squares Regression} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$$

$$\text{SST} = \text{Sum of Squares Total} = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$$

SST=SSR+SSE

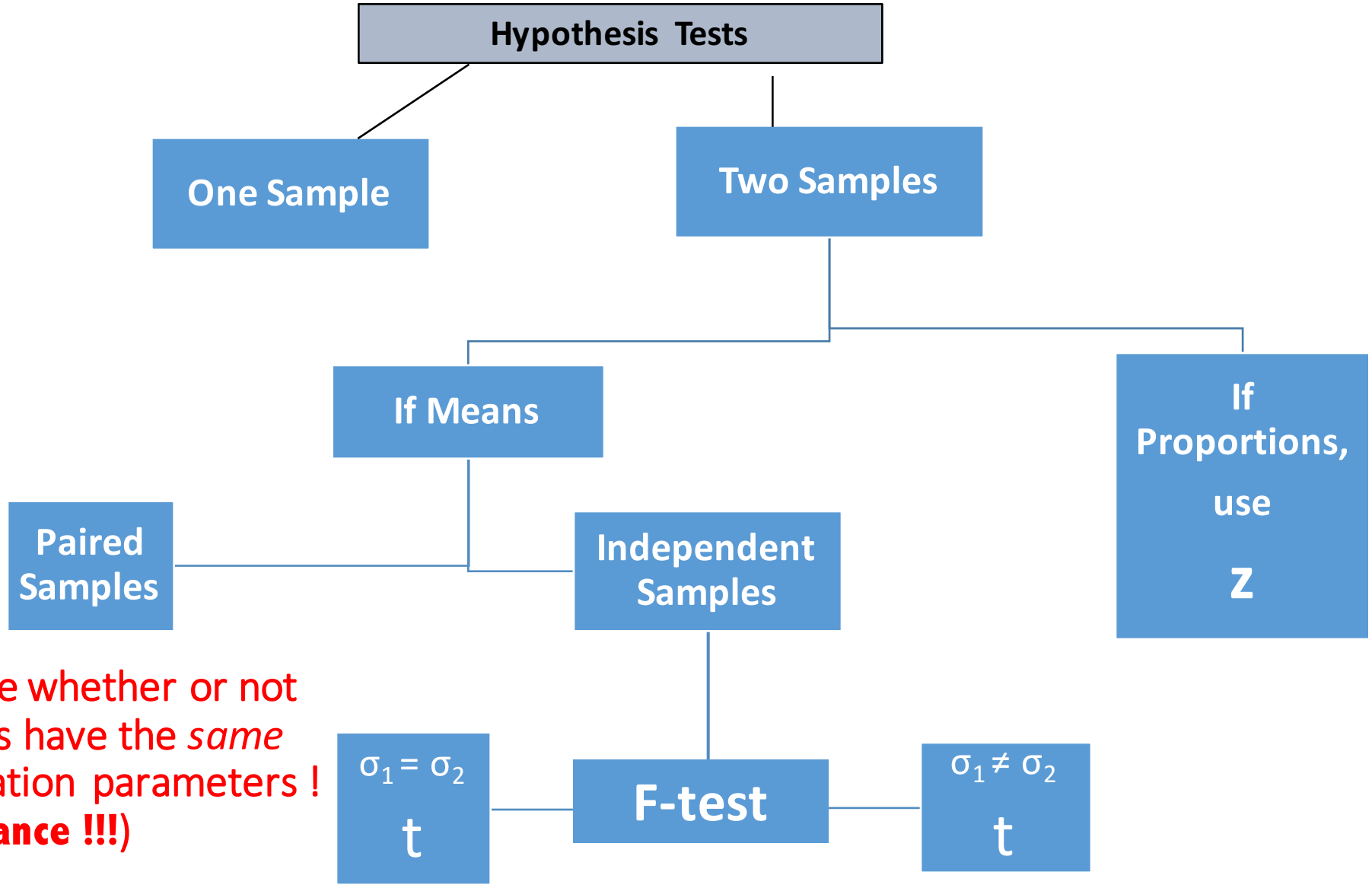
$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SST - SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}} \le 1$$

Important Topics

- R-square, F-test (overall significance), t-test (individual coefficient)
- SST,SSR,SSE
- Interpretation of regression coefficients (slope, intercept) and finding the confidence interval
- Correlation coefficient and its interpretation

Quiz 4

Testing hypotheses for Two Samples (Route map)

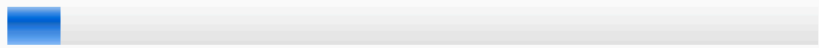
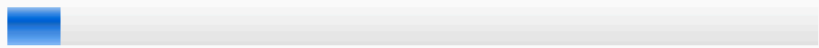
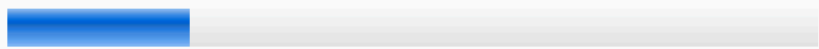
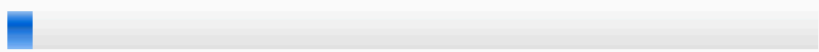


Use F-test to see whether or not the two samples have the *same* **variance** (population parameters ! **Not sample variance !!!**)

two population indep samples (Questions 1 - 2)

Random Question 1 Difficulty: 1

training.xlsx This file contains sales data on the number of cans of tuna sold in two cities during an advertising campaign. Assuming the variances of the sales in Toronto and Vancouver are the same, the test of whether the average difference between Toronto and Vancouver sales is equal to zero against the alternative that the average sales in Toronto are greater than those in Vancouver leads us to:

not reject the null hypothesis since the calculated test statistic is less than the critical five percent value for an upper-tailed test		2	(6.45 %)
not reject the null hypothesis at the seven percent level of significance		2	(6.45 %)
not reject the null hypothesis at the five percent level of significance		7	(22.58 %)
not reject the null hypothesis at the six percent level of significance		1	(3.23 %)
reject the null hypothesis at the five percent level of significance		19	(61.29 %)

Average Grade: 0.61 / 1 (61.29 %)

Standard Deviation n/a

Point Biserial n/a

Discrimination Index n/a

Check the Route map first

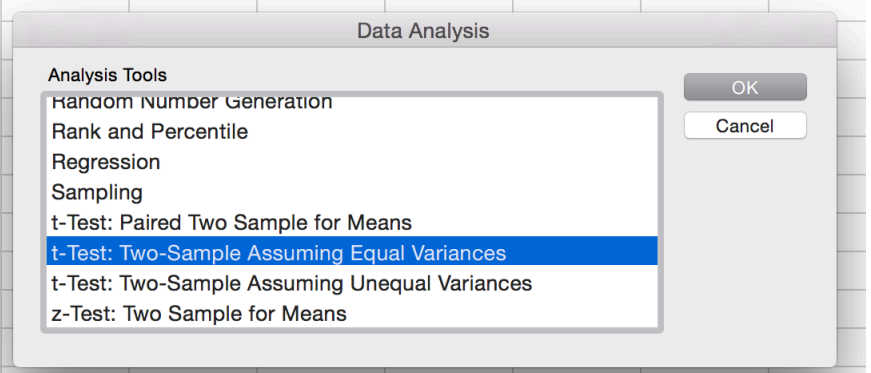
Hypothesis Tests

One Sample

If

Paired Samples

Session	Toronto	Vancouver
1	94	45
2	84	40
3	91	46
4	76	37
5	91	40
6	92	44
7	82	42
8	82	42
9	92	47
10	91	35
11	92	44
12	86	36
13	83	41
14	86	39
15	97	48
16	94	46



F-test

$\sigma_1 = \sigma_2$

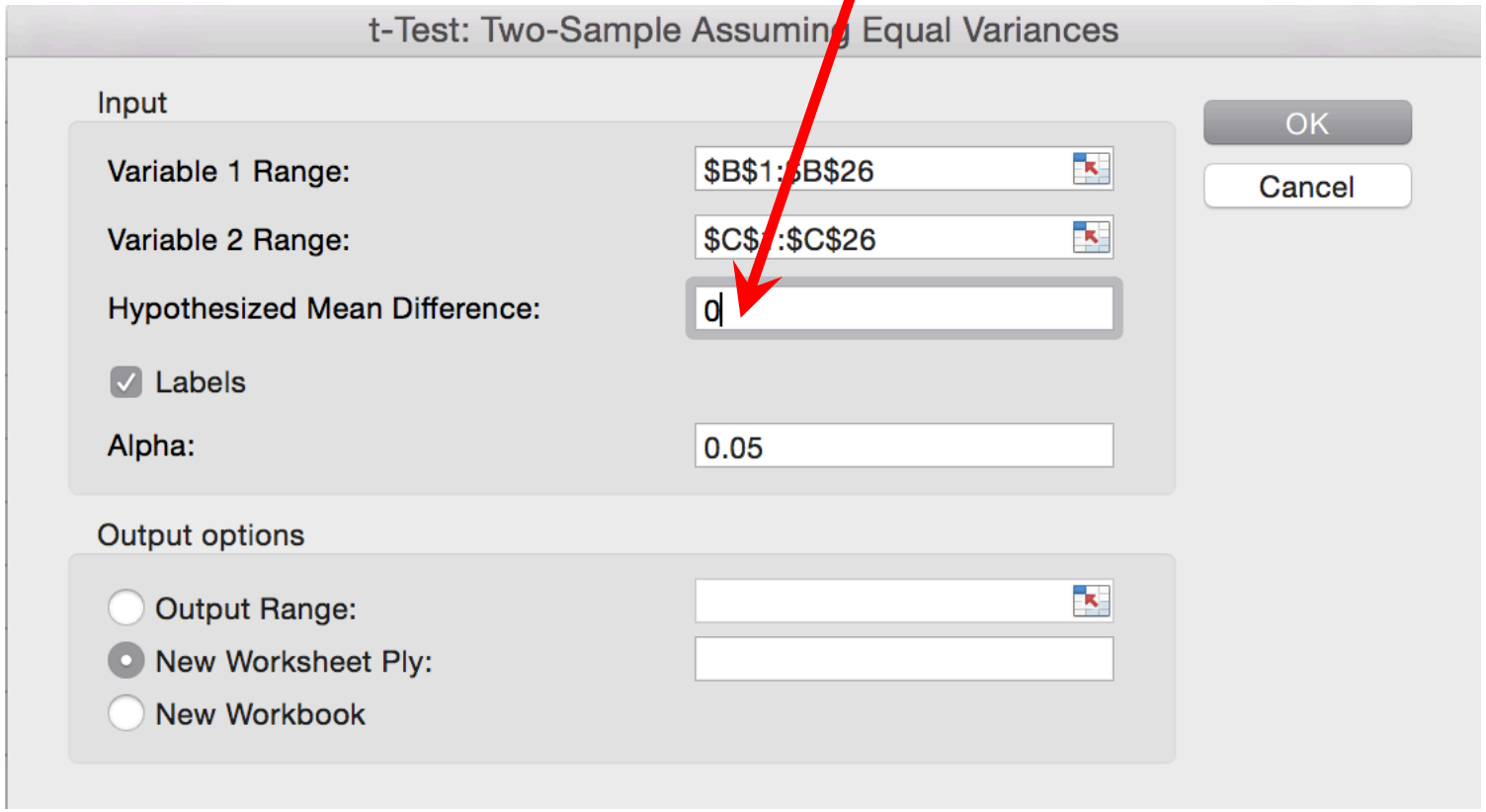
t

$\sigma_1 \neq \sigma_2$

t

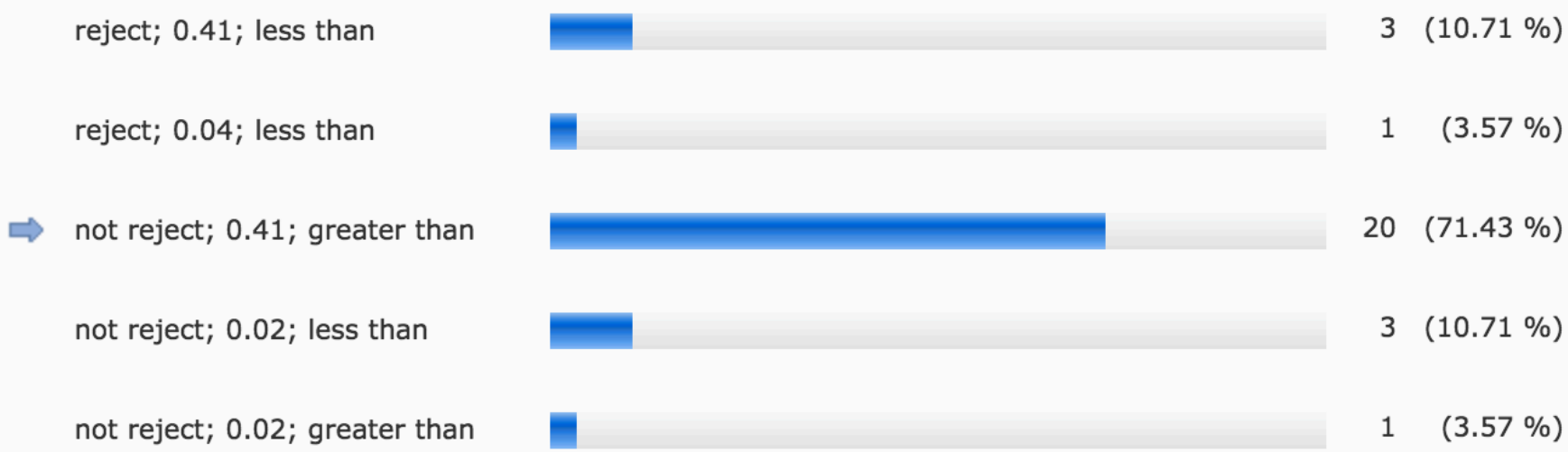
training.xlsx This file contains sales data on the number of cans of tuna sold in two cities during an advertising campaign. Assuming the variances of the sales in Toronto and Vancouver are the same, the test of whether the average difference between Toronto and Vancouver sales is equal to zero against the alternative that the average sales in Toronto are greater than those in Vancouver leads us to:

t-Test: Two-Sample Assuming Equal Variances		
	Toronto	*Vancouver*
Mean	86.2	40.64
Variance	43.16666667	21.49
Observations	25	25
Pooled Variance	34.82833333	
Hypothesized Mean Difference	0	
df	48	
t Stat	7.29432608	
P(T<=t) one-tail	3.45558E-31	
t Critical one-tail	1.677224196	
P(T<=t) two-tail	6.91117E-31	
t Critical two-tail	2.010634758	



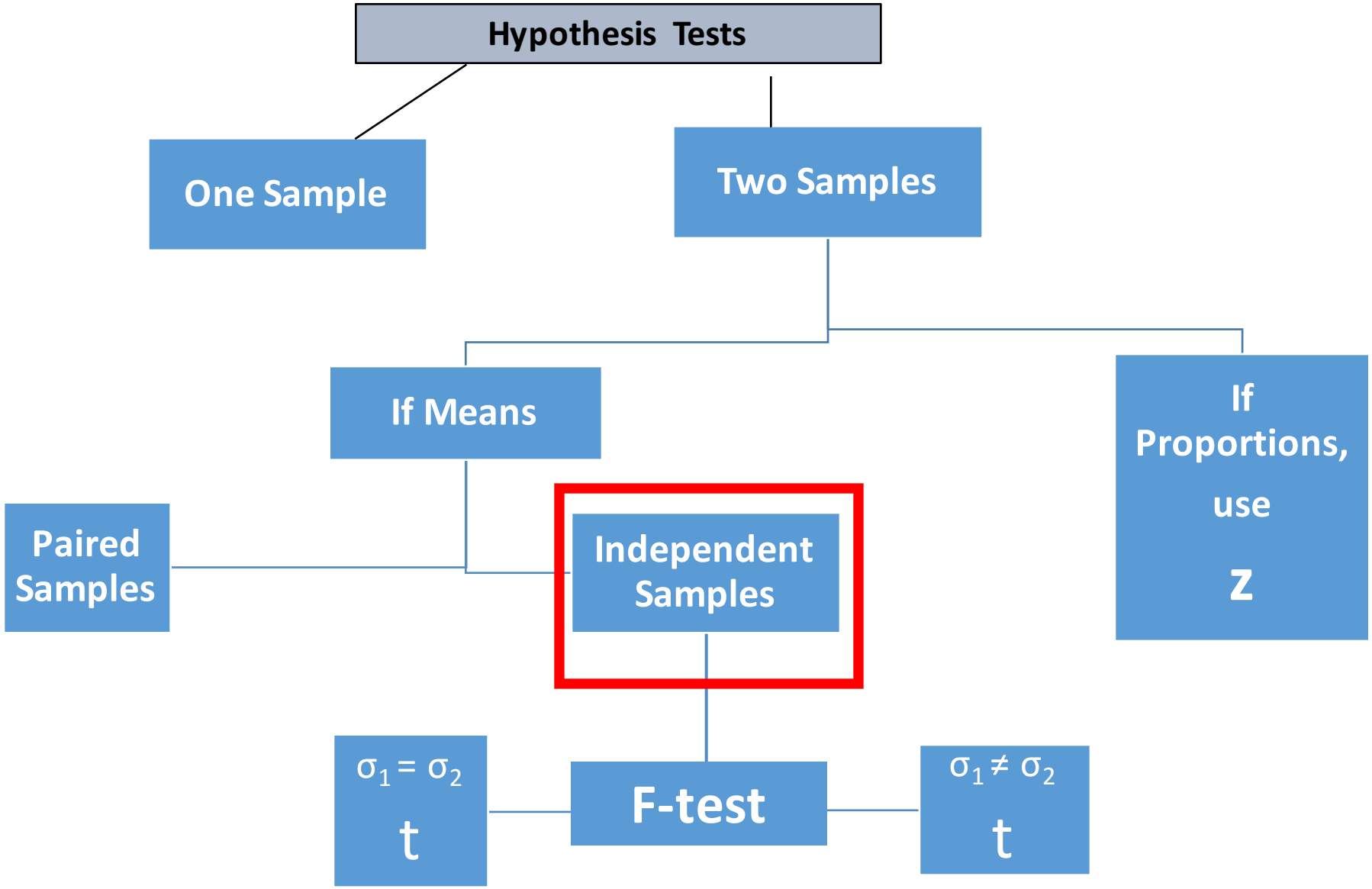
Random Question 2 Difficulty: 1

Apparel.xlsx This file contains customer satisfaction scores for several firms. Consider the data for Levi Strauss and Hanes. Using a five percent level of significance, a test of whether the average Levi Strauss score is equal to the average Hanes score leads one to _____ (reject/not reject) the null since the one-tailed p-value is (to two decimal points) _____, which is _____ (greater than/less than) the alpha.



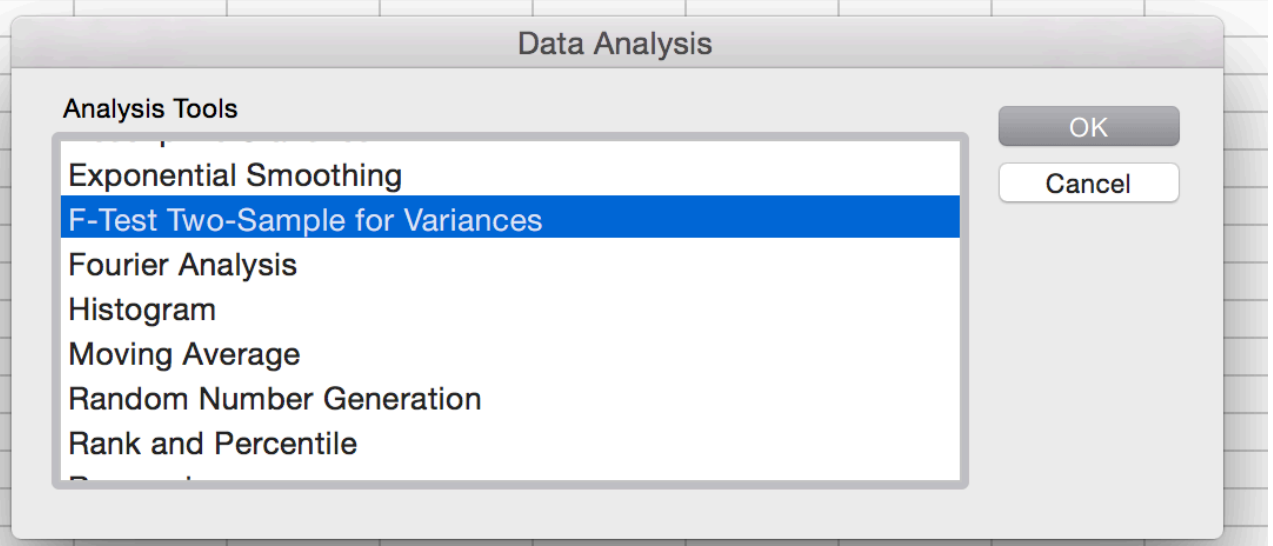
Average Grade: 0.71 / 1 (71.43 %)
Standard Deviation n/a
Point Biserial n/a
Discrimination Index n/a

Step 1: Check the Route map first



Step 2: F-test

year	VF	Levi Strauss	Hanes	Liz Claiborne
1994	83	84	83	84
1995	80	83	81	81
1996	80	80	75	81
1997	81	81	81	77
1998	79	75	77	78
1999	78	76	78	76
2000	82	79	78	79
2001	84	80	76	79
2002	82	78	78	80
2003	84	80	80	78
2004	79	80	79	79
2005	82	79	79	78
2006	82	79	82	81
2007	84	80	82	79
2008	83	78	80	79
2009	81	83	82	82
2010	85	81	81	79
2011	83	81	82	79



$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1;$$

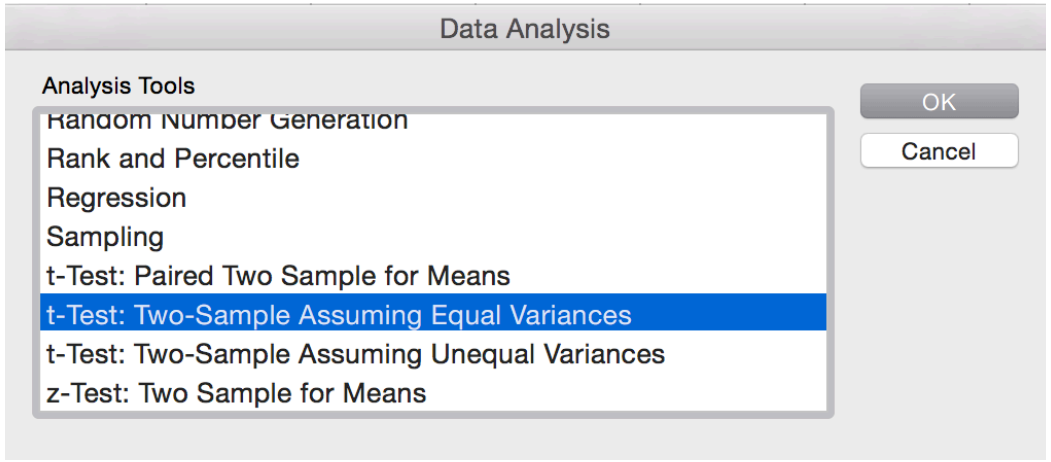
$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

F-Test Two-Sample for Variances

	Levi Strauss	*Hanes*
Mean	79.83333333	79.66666667
Variance	5.205882353	5.294117647
Observations	18	18
df	17	17
F	0.983333333	
P(F<=f) one-tail	0.486381544	
F Critical one-tail	0.440161596	

Not done yet !

Step 3: t-test for the mean !

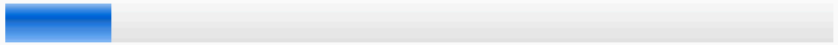
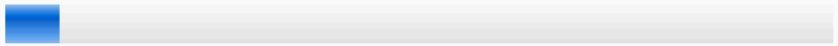


H0: mean of Levi Strauss = mean of Hanes
Ha: mean of Levi Strauss > mean of Hanes

t-Test: Two-Sample Assuming Equal Variances		
	Levi Strauss	*Hanes*
Mean	79.83333333	79.66666667
Variance	5.205882353	5.294117647
Observations	18	18
Pooled Variance	5.25	
Hypothesized Mean Difference	0	
df	34	
t Stat	0.21821789	
P(T<=t) one-tail	0.41428192	
t Critical one-tail	1.690924255	
P(T<=t) two-tail	0.828563841	
t Critical two-tail	2.032244509	

6

Apparel.xlsx This file contains customer satisfaction scores for several firms. Consider the data for Levi Strauss and Liz Claiborne. A test of whether the average Levi Strauss score is 1 point higher than the average Liz Claiborne score leads one to:

reject the null at the five percent level of significance since the one tailed pvalue is about 0.02		6	(19.35 %)
not reject the null at the five percent level of significance since the one tailed pvalue is about 0.84		4	(12.9 %)
reject the null at the five percent level of significance since the one tailed pvalue is almost zero		4	(12.9 %)
not reject the null at the five percent level of significance since the one tailed pvalue is about 0.21		15	(48.39 %)
not reject the null at the five percent level of significance since the one tailed pvalue is nearly 1		2	(6.45 %)

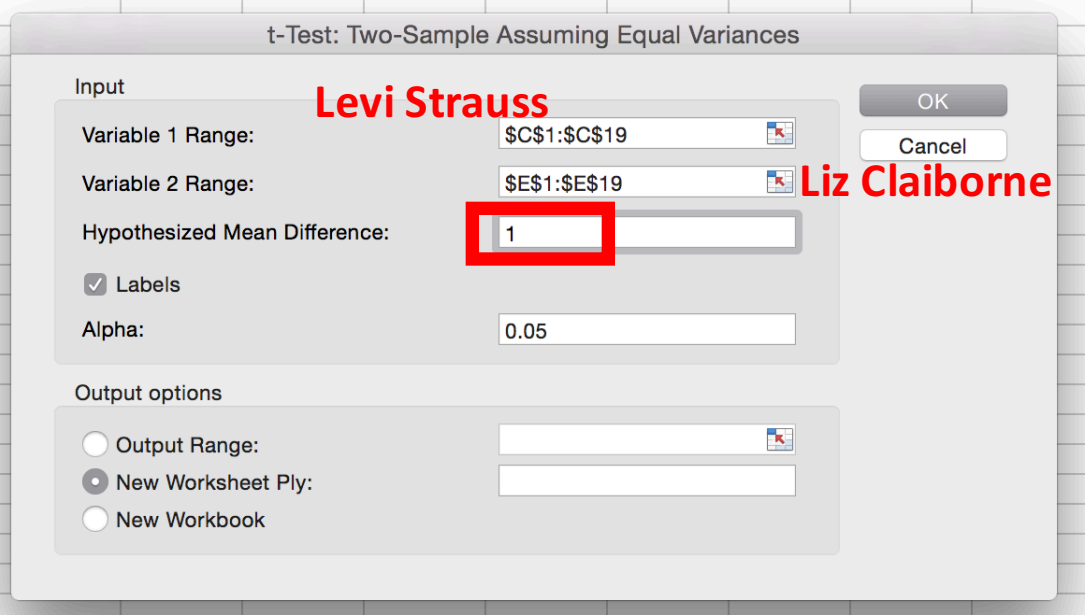
Average Grade: 0.48 / 1 (48.39 %)
Standard Deviation n/a
Point Biserial n/a
Discrimination Index n/a

H0: mean of **Levi Strauss** - mean of **Liz Claiborne** >1
Ha: mean of Levi Strauss - mean of Liz Claiborne <=1

Step 1: Check the Route map first
Step 2: F-test
Step 3: t-test assuming equal variances

year	VF	Levi Strauss	Hanes	Liz Claiborne
1994	83	84	83	84
1995	80	83	81	81
1996	80	80	75	81
1997	81	81	81	77
1998	79	75	77	78
1999	78	76	78	76
2000	82	79	78	79
2001	84	80	76	79
2002	82	78	78	80
2003	84	80	80	78
2004	79	80	79	79
2005	82	79	79	78
2006	82	79	82	81
2007	84	80	82	79
2008	83	78	80	79
2009	81	83	82	82
2010	85	81	81	79
2011	83	81	82	79

19R x 1C



t-Test: Two-Sample Assuming Equal Variances		
	Levi Strauss	*Liz Claiborne*
Mean	79.83333333	79.38888889
Variance	5.205882353	3.545751634
Observations	18	18
Pooled Variance	4.375816993	
Hypothesized Mean Difference	1	
df	34	
t Stat	-0.796744684	
P(T<=t) one-tail	0.215564015	
t Critical one-tail	1.690924255	
P(T<=t) two-tail	0.43112803	
t Critical two-tail	2.032244509	

4

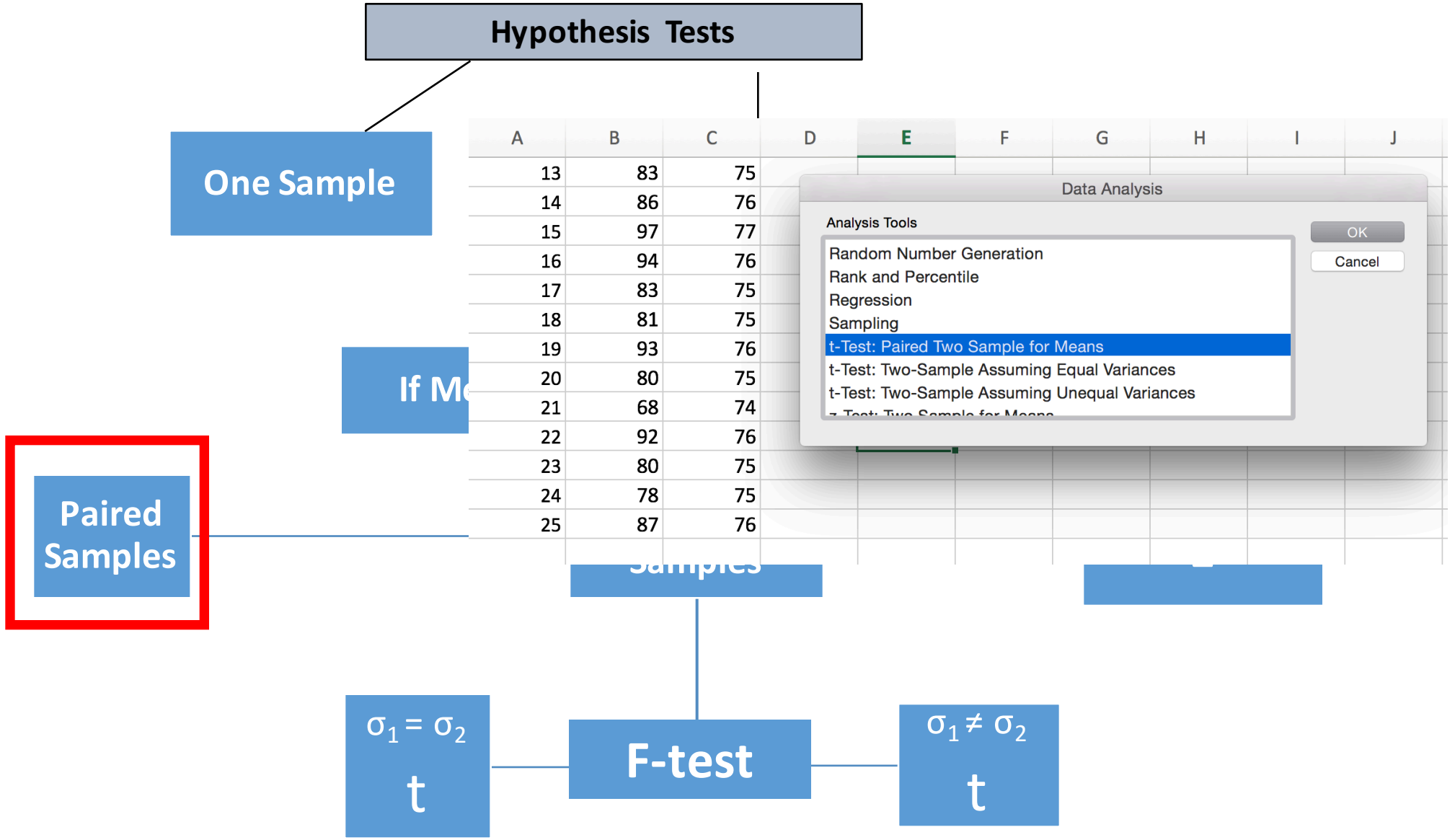
two population matched samples (Questions 3 - 3)

Random Question 1 Difficulty: 1

shiftwork.xlsx This file contains the number of iPhones sold by the same sales agent working during the day shift or the night shift over the past 50 business days. Using an upper-tailed test, you want to see whether their sales performance during the day shift is 8 units higher than during the night shift. When you perform the test, you find that the test statistic is equal (to two decimal points) to _____, and this yields a pvalue (to three decimal points) equal to _____, so you (reject/do not reject) _____ the null hypothesis at the five percent level of significance.

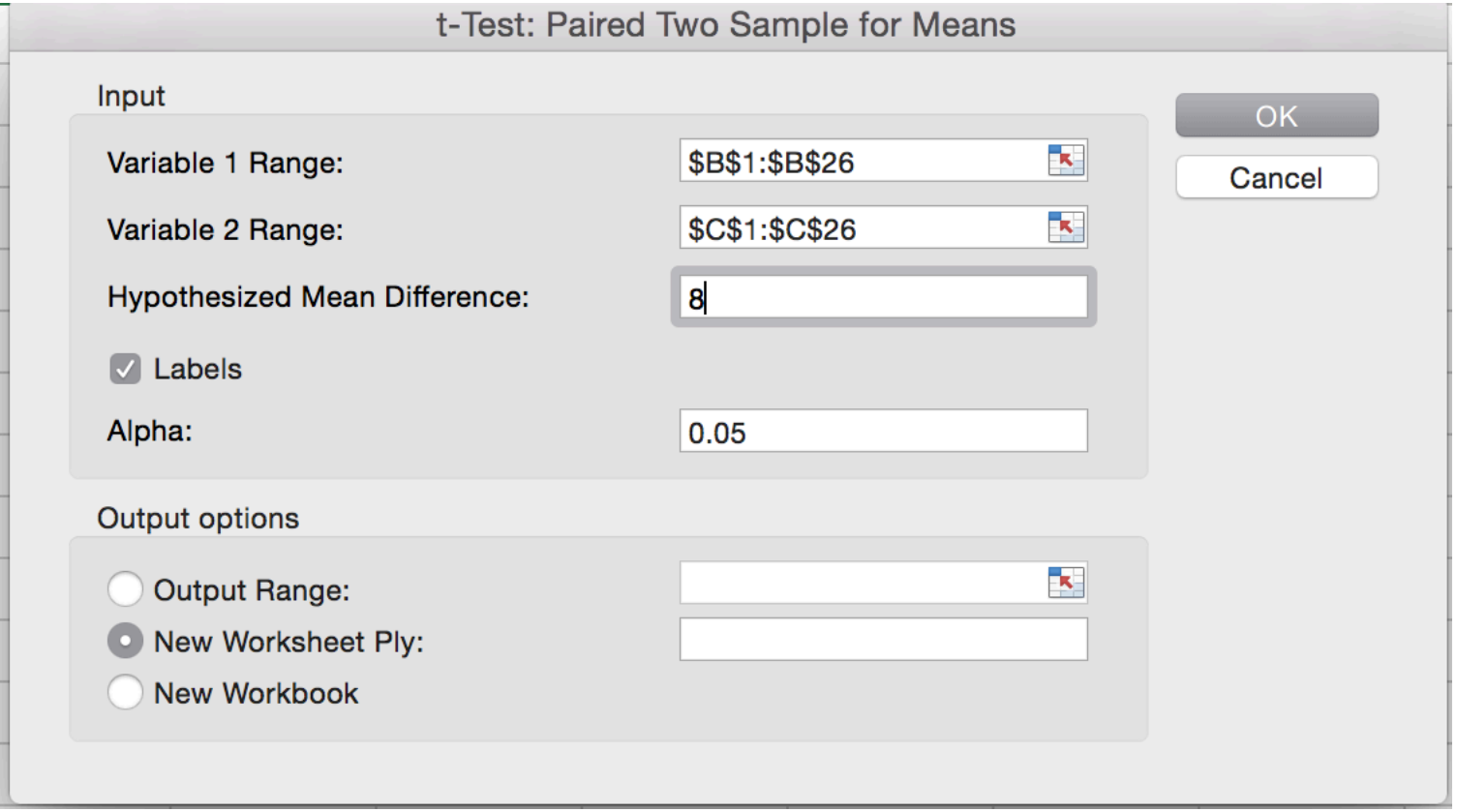
➡	2.08; 0.024; reject	8	(40 %)	
	1.97; 0.032; reject	4	(20 %)	Average Grade: 0.4 / 1 (40 %)
	2.44; 0.067; do not reject	3	(15 %)	Standard Deviation n/a
	2.08; 0.024; do not reject	1	(5 %)	Point Biserial n/a
	1.97; 0.032; do not reject	4	(20 %)	Discrimination Index n/a

Check the Route map first



H0: mean performance of **day**- mean performance of **night** >8
Ha: mean performance of **day**- mean performance of **night** <=8

Session	Day	Night
1	94	76
2	84	75
3	91	76
4	76	75
5	91	76
6	92	76
7	82	75
8	82	75
9	92	76
10	91	76
11	92	76
12	86	76

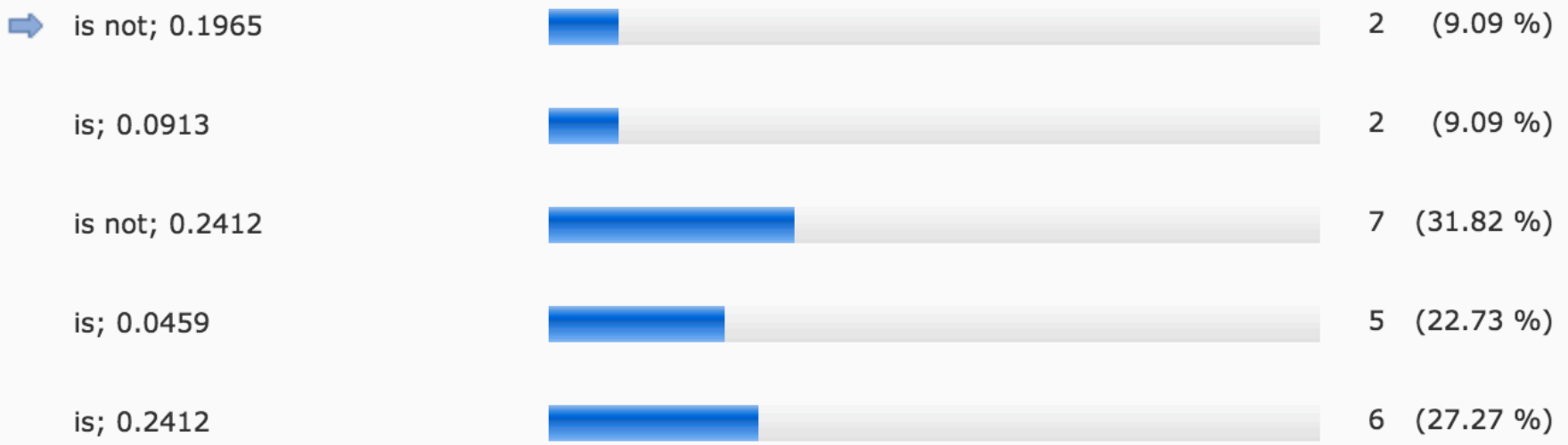


t-Test: Paired Two Sample for Means		
	Day	*Night*
Mean	86.2	75.56
Variance	48.16666667	0.42333333
Observations	25	25
Pearson Correlation	0.924574583	
Hypothesized Mean Difference	8	
df	24	
t Stat	2.080869982	
P(T<=t) one-tail	0.024144082	
t Critical one-tail	1.71088208	
P(T<=t) two-tail	0.048288164	
t Critical two-tail	2.063898562	

two sample proportions (Questions 4 - 4)

Random Question 1 Difficulty: 1

Three hundred people were asked a question. 50 were selected and asked if they wanted a new cellphone; 40% said they did. The remaining people were asked if they wanted a new cellphone, and 50% said yes. Using a 5% level of significance, then null hypothesis that the proportion from the first group who do not want a new cellphone is equal to the proportion from the second group who do not want a new cellphone (is/is not) rejected since the pvalue (rounded to 4 decimal points) is equal to _____.



The hypotheses are:

H0: p1 = p2

H1: p1 ≠ p2

Where **H0 implies** $p_1 = p_2$ ($\Leftrightarrow$ $p_1 - p_2 = 0$), we can pool the estimated values to get a better pooled estimate:

$$\hat{p}_{pool} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

$$i.e.\ (\hat{p}_1 - \hat{p}_2) \sim \mathrm{N}\left(\mu = 0,\ \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

then we can calculate $Z^* = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$$n_1 = 50, \hat{p}_1 = 0.4, n_2 = 250, \hat{p}_2 = 0.5$$

$$\hat{p}_{pool} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{50 \times 0.4 + 250 \times 0.5}{300} = 145/300 = 0.4833$$

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{0.4833 \times (1 - 0.4833) \times \left(\frac{1}{50} + \frac{1}{250}\right)} = 0.07742$$

$$Z^\star = \frac{\hat{p}_1 - \hat{p}_2}{\sigma_{\hat{p}_1 - \hat{p}_2}} = \frac{-0.1}{0.07742} = -1.2917$$

P-value =2 * NORM.S.DIST (-1.2917, 1) = 0.1965

Hypothesis tests for two populations

Test for		Ho	Test Statistics	Distribution	Conditions
Difference of two means (independent samples) (μ_1-μ_2)	$\sigma_1 = \sigma_2$	μ_1-μ_2 $\leq$ / $\geq$ / $=$ μ_0	$t^* = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{S_{pool}^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ $S_{pool} = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$	t with $df = n_1 + n_2 - 2$	Normal distribution or $n > 30$, σ_1 , σ_2 unknown
	$\sigma_1 \neq \sigma_2$	μ_1-μ_2 $\leq$ / $\geq$ / $=$ μ_0	$t^* = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$	t with df = the smaller of $n_1 - 1$ or $n_2 - 1$	
Mean difference (paired samples) (μ_d)		μ_d $\leq$ / $\geq$ / $=$ μ_0	$t^* = \dfrac{\bar{d} - \mu_d}{s_d / \sqrt{n_d}}$	t with $df = n - 1$	Normal dist. or $n > 30$, σ unknown
Difference of two proportions (p_1-p_2)		$p_1 - p_2$ $\leq$ / $\geq$ / $=$ 0	$Z^* = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_{pool}(1-\hat{p}_{pool})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ $\hat{p}_{pool} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \dfrac{X_1 + X_2}{n_1 + n_2}$	Z	$n\hat{p} > 5$ $n\hat{q} > 5$ for each group
		$p_1 - p_2$ $\leq$ / $\geq$ / $=$ D	$z^* = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$		

Important Topics

- Hypothesis testing about mean, proportion with two population
- F-test, t-test with equal (or unequal) variances, paired test.
- Testing the nonzero differences in the mean (or proportion). For example, p_1-p_2>=D=0.08.

Quiz 3

Q1 (Questions 1 - 2)

Random Question 1 Difficulty: 1

It is widely believed that 98 percent of humans do not have red hair. Suppose we consider the population of Kingston of 125,000 people. If you randomly select 100 people living in Kingston, the probability, to four decimal points, of finding no more than 1.5 percent with red hair is:

	0.0321	1	(50 %)	
➡	0.3605	1	(50 %)	Average Grade: 0.5 / 1 (50 %)
	0.2375	0	(0 %)	Standard Deviation n/a
	0.1432	0	(0 %)	Point Biserial n/a
	0.1294	0	(0 %)	Discrimination Index n/a

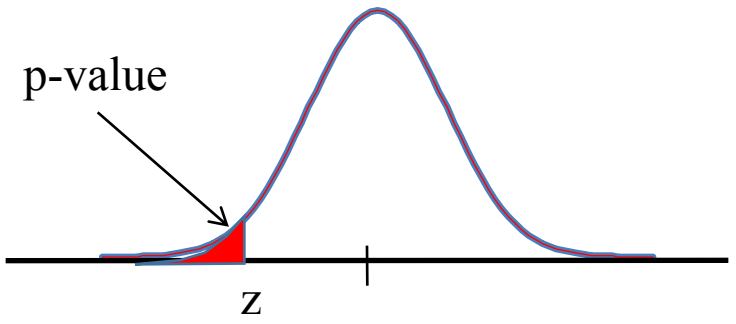
Sample Proportions

$\hat{p} = x / n$ **proportion of successes for a sample of size n.**

$$\hat{p} \sim approx\ N(\mu_{\hat{p}}, \sigma_{\hat{p}})$$

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{p(1\text{-}p)/n}$$

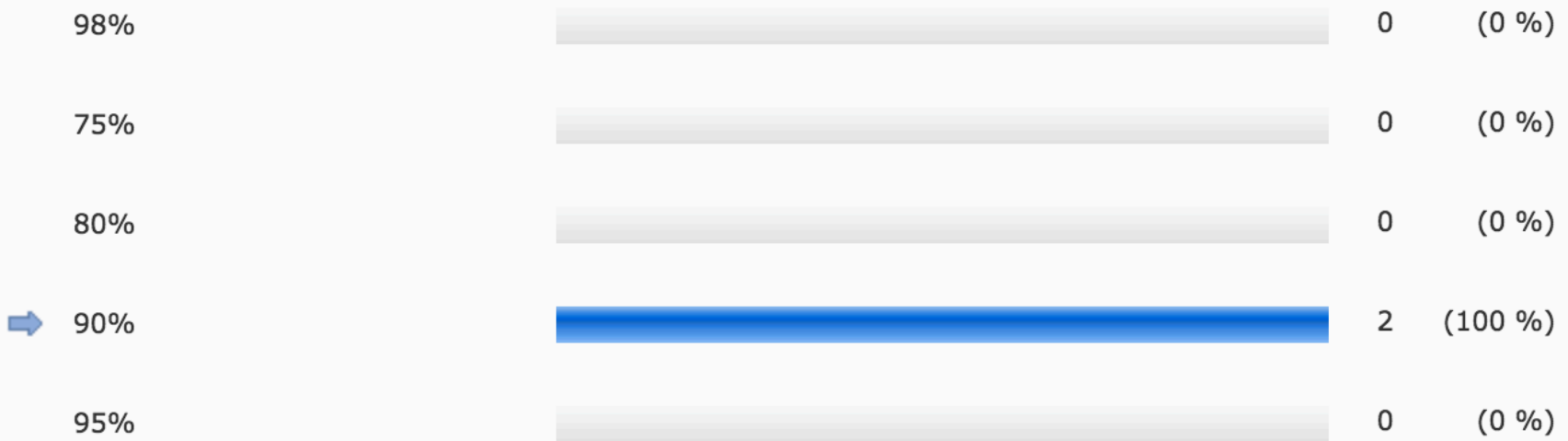
$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$



Z : standard normal

Random Question 3 Difficulty: 1

Unfilling cereal boxes can create problems for firms, so they periodically check to ensure the boxes contain the correct amount of cereal. Opening the boxes destroys the product, so firms try to limit the number of boxes in the sample to minimize their costs. Suppose you know that the weight of cereal boxes is not normally distributed, but they range from a low of 150 grams to a high of 220 grams. To estimate the population mean weight of the cereal boxes with _____% confidence, and be within 4 grams of the actual weight, you would need to have a sample that contains 52 cereal boxes.



Average Grade: 1 / 1 (100 %)
Standard Deviation n/a
Point Biserial n/a
Discrimination Index n/a

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{E}{\sigma_x / \sqrt{n}} \Rightarrow n = \left(\frac{z_{\alpha/2}\sigma_x}{E} \right)^2$$

Note: If we do not know σ, we can estimate it using:

- **a pilot study, previous research, or σ ≈ Range / 4.**

$$E = 4 \qquad \sigma_x \approx (200 - 150)/4 = 12.5$$

$$n \leqslant 52 \qquad z_{\alpha/2} \leqslant 1.6483$$

Power= probability that a statistical test will correctly reject a false null hypothesis
The power of a test is defined as 1– β.

Sample decision (H_0)	**Same (H_0 true)**	**Different (H_0 false)**
Same (do not reject)	Right (rats). **Confidence = 1-α**	Type II error. **P(Type II)=β**
Different (reject)	Type I error. **P(Type I)= α**	**Right! Power=1-β**

The power of a test depends on how far the true parameter value is from that assumed by the null hypothesis. The distance between the null hypothesis value and the truth is called the **effect size**.

Recall the Type II error problem in Lecture 14

You are given the following null and alternative hypotheses:

$$H_0: \mu = 200$$

$$H_1: \mu \neq 200$$

$$\alpha = 0.10$$

Calculate the probability of committing a Type II error when the population mean is actually 197, the sample size is 36, and the population standard deviation is known to be 24.

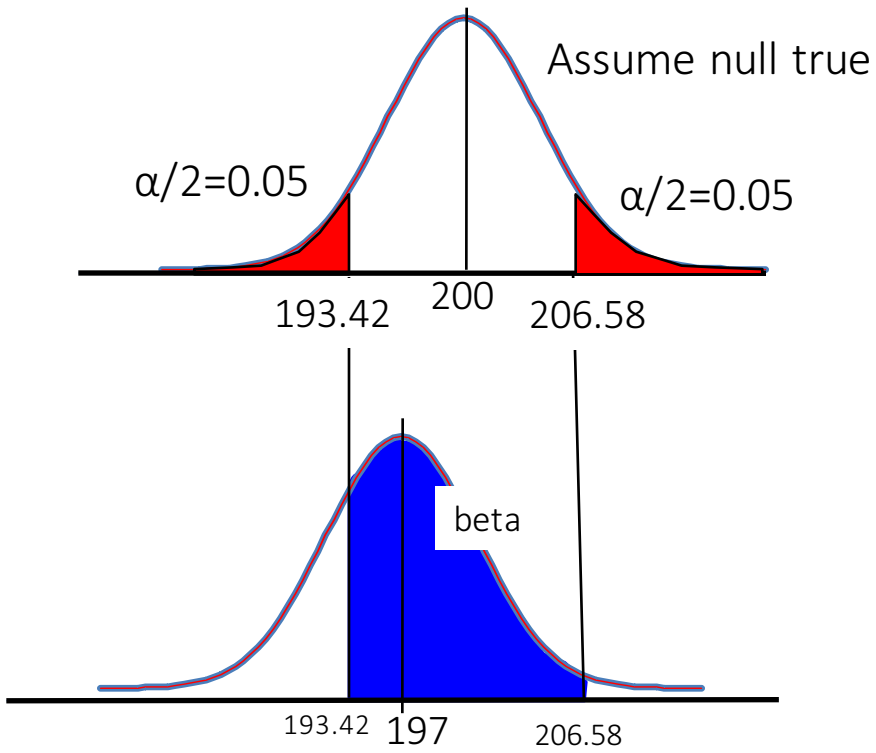
$H_0: \mu = 200$

$H_1: \mu \neq 200$

$\alpha = 0.10$

since **σ is known**

use $z_{0.05} = \pm 1.645$



Calculate the probability of committing a Type II error when the population mean is 197, the sample size is 36, and the population standard deviation is known to be 24.

$\beta = P(193.42 < \bar{x} < 206.58,$ assuming that the true mean is 197)

≅ NORM.DIST(206.58,197,(24/6),1)-NORM.DIST(193.42,197,(24/6),1)

=0.8063

Know These Z-Score Values

Some Z-score values appear far more often than others.

The single most commonly used value is **1.96.**

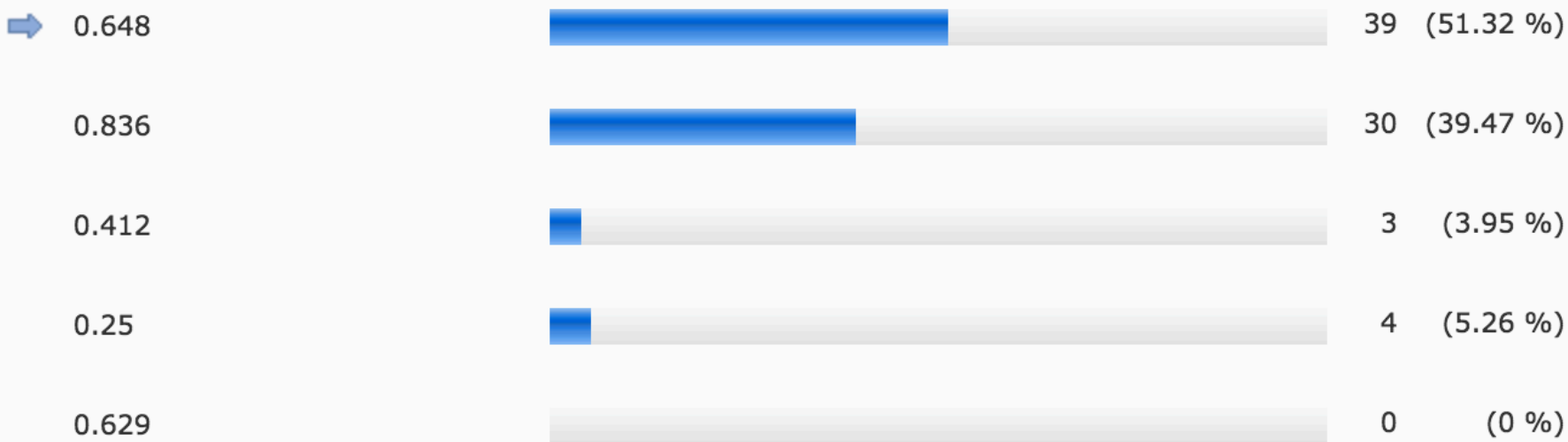
Tails	Probability	Z
1	0.10	1.28
1	0.05	1.645
1	0.01	2.33
2	0.10	1.645
2	**0.05**	**1.96**
2	0.01	2.576

Important Topics

- Confidence Interval for mean and proportion.
- T-distribution
- Calculate the P-value for a given test statistic
- Type I and Type II error, power of test.
- Determine the minimal sample size required for finding confidence interval for mean and proportion.

Quiz 2

Canada Dry has a 20% share in the ginger ale market. Suppose 15 ginger ale drinkers are randomly selected from the population. The probability that fewer than four choose Canada Dry is

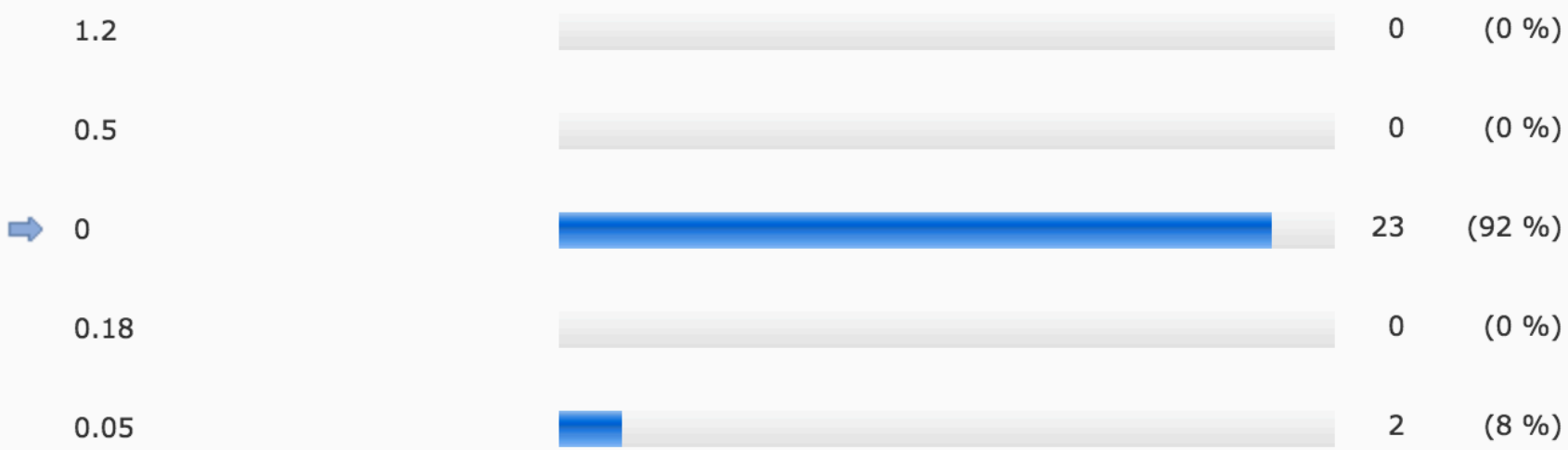


Average Grade: 0.51 / 1 (51.32 %)
Standard Deviation n/a
Point Biserial n/a
Discrimination Index n/a

continuous rv (Questions 2 - 2)

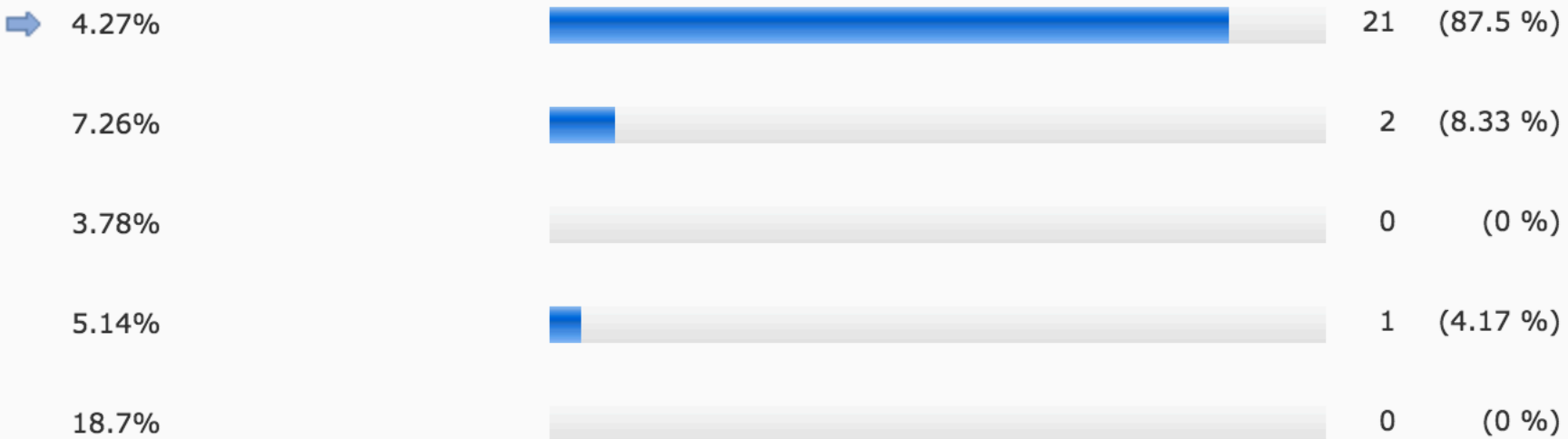
Random Question 1 Difficulty: 1

The number of packages of lifesavers produced during an 8 hour shift is uniformly distributed with a minimum of 0 and a maximum of 40,000. The probability that exactly 42,000 will be produced in an eight hour shift is



Average Grade: 0.92 / 1 (92 %)
Standard Deviation n/a
Point Biserial n/a
Discrimination Index n/a

The Commerce office has an average random arrival rate of 6 students every 10 minutes. What is the probability of having exactly 30 students during one hour?

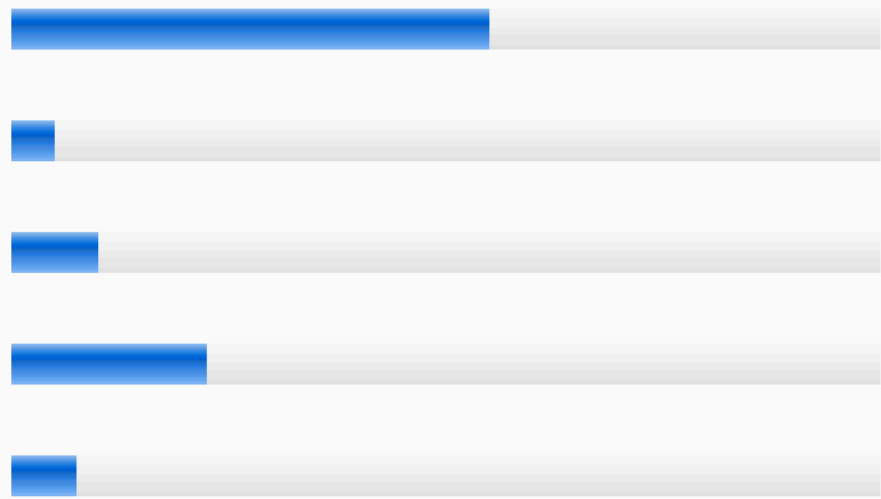


4.27%		21	(87.5 %)
7.26%		2	(8.33 %)
3.78%		0	(0 %)
5.14%		1	(4.17 %)
18.7%		0	(0 %)

Average Grade:	0.88 / 1 (87.5 %)
Standard Deviation	n/a
Point Biserial	n/a
Discrimination Index	n/a

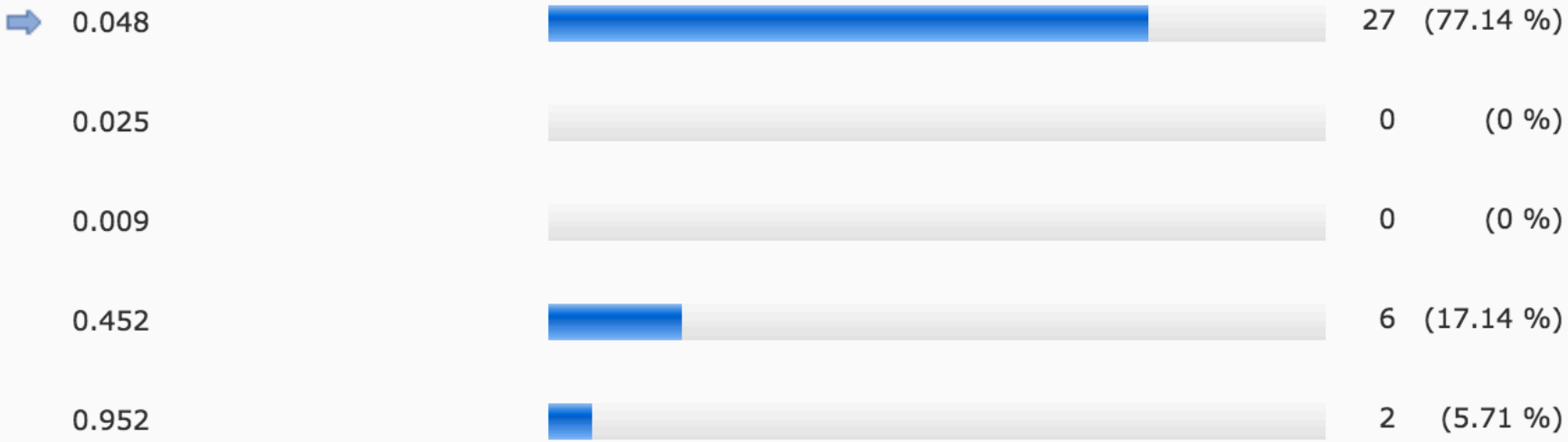
If a True or False test has 50 questions, and Bart Simpson guesses every answer, the normal approximation to the binomial suggests that the probability, to four decimal points (eg, a number like 0.8392) of Bart answering 30 questions or higher correctly is:

➡	0.1015	22	(55 %)
	0.5562	2	(5 %)
	0.8985	4	(10 %)
	0.3985	9	(22.5 %)
	0.0354	3	(7.5 %)



Average Grade:	0.55 / 1 (55 %)
Standard Deviation	n/a
Point Biserial	n/a
Discrimination Index	n/a

There are tens of thousands of cabs in Toronto. The odometre reading in each cab is not normally distributed, but is known to have a mean of 68,000 km and a standard deviation of 18,000 km. If you took a sample of 900 cabs, the central limit theorem suggests that the probability, to three decimal points (eg a number like 0.839), that your sample will have a mean less than 67,000 is about:



➡	0.048	27	(77.14 %)
	0.025	0	(0 %)
	0.009	0	(0 %)
	0.452	6	(17.14 %)
	0.952	2	(5.71 %)

Average Grade: 0.77 / 1 (77.14 %)

Standard Deviation n/a

Point Biserial n/a

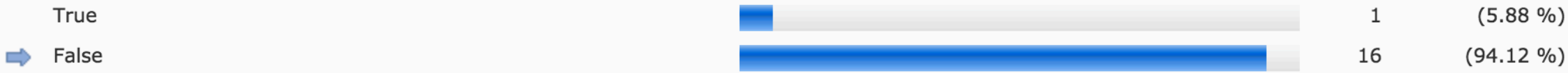
Discrimination Index n/a

Important Topics

- Binomial Distribution, Poisson Distribution,
- Uniform distribution, normal (and standard normal) distribution
- Normal Approximation to Binomial distribution
- Central limit theorem

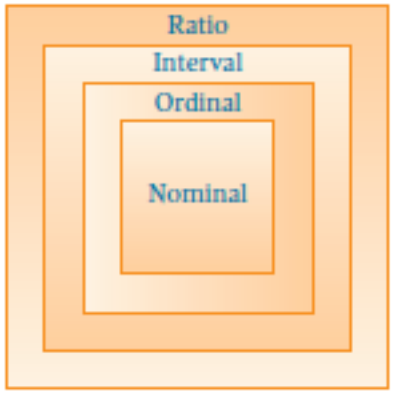
Quiz 1

Interval level data is the highest level of data measurement.



Types of Data

There are four types of data arranged in a hierarchy by information content: nominal, ordinal, interval, and ratio:



Nominal:

- Categories
- Numeric values assigned to categories have no meaning
- *Example:* single / married

Ordinal:

- Rank is meaningful
- Distance between values is not equal
- *Example:* low, medium, high

Interval:

- Distance between values has meaning
- There is ***no*** **absolute zero**
- *Example:* temperature in Celsius

Ratio:

- Distance between values has meaning
- There is an **absolute zero**
- *Example:* Starting salary at graduation

Apparel.xlsx

In the data file, there are customer satisfaction scores for several firms. The scores for Hanes are more narrowly distributed around their mean than for any other firm.

Descriptive Statistics with Excel

- **'Data, Data Analysis'**
  choose 'Descriptive Statistics, Summary Statistics'

Distance Travelled KM	
Mean	8.820755
Standard Error	0.500278
Median	8
Mode	8
Standard Deviation	5.150677
Sample Variance	26.52947
Kurtosis	-0.67942
Skewness	0.742313
Range	17
Minimum	3
Maximum	20
Sum	935
Count	106

Descriptive Statistics with Excel

Instead of using Data Analysis, you can also use Excel's functions for some descriptive statistics, such as:

=AVERAGE()

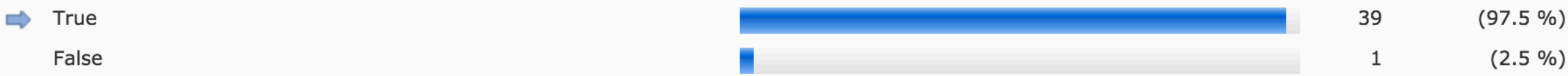
=MEDIAN()

=MODE()

=STDEV()

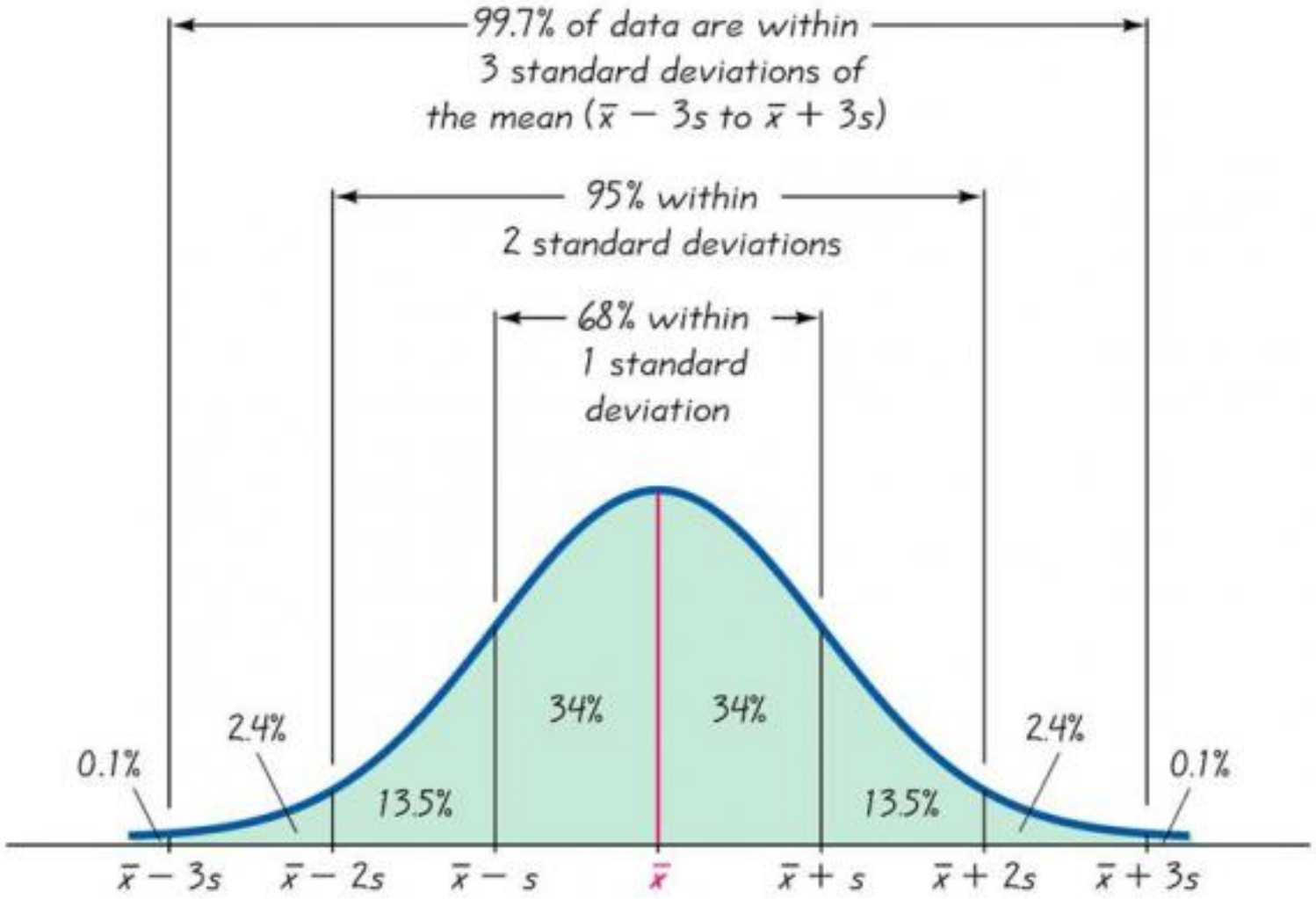
=VAR()

The empirical rule says that if the data are normally distributed, about 68% of the values will lie within plus or minus 1 standard deviations of the mean.



The **empirical rule** says that if a distribution is bell-shaped (i.e. approximately ***Normal***):

- 68% of all observat
- 95% within +/- **2σ**
- 99.7% with +/- **3σ**



A researcher's aim is to ensure that the frame used for a sample is as different from the target population as possible.

	True	0	(0 %)
➡	False	40	(100 %)

Percentiles

- The 1st to 99th **percentiles** provide information on the distribution of data that goes beyond where it is centered.

- Percentiles are defined as the value such that **a certain percentage of the observations are lower**:
  - Example: The 32nd percentile is a value such that at least 32 percent of the data lies below that value, and no more than 68% of the data lies above that value..

Important Topics

- Data type, level of measurement.
- Descriptive statistics
- Percentile
- Empirical rule

Outline of the Course

We cover descriptive statistics, distributions, estimation, hypothesis testing, and regression:

Descriptive Statistics:

- Provide insightful summaries
- Develop graphs and charts to tell managerially relevant stories

Distributions:

- Understand commonly occurring discrete and continuous distributions
- Utilize distributions to determine the probability of various events for both statistical and managerial decisions

Estimation:

- Make inferences about populations from samples
- Establish confidence intervals around estimates

Hypothesis Testing:

- Use statistics to make managerial decisions with a measure of confidence
- Understand the types of errors one can make and how to manage them

Regression:

- Integrate all the previous material
- Develop, assess, test and use regression models to gain insights into complex processes in a variety of fields