Unit 10 – Two Sample Inference
Homework
Solutions

1. The National Health and Nutrition Examination Survey of 1975-1980 give the following data on serum cholesterol levels in US males.

Group	Age, years	Population Mean, μ	Population Standard Deviation, σ
1	20-24	180	43
2	25-34	199	49

Suppose the distribution of serum cholesterol is normal in each age group. If you draw simple random samples of size 50 from each of the two groups, what is the probability that the difference between the two sample means (Group 2 mean – Group 1 mean) $= \mu_2 - \mu_1$ will be more than 25?

Source: National Center or Health Statistics, R. Fulwood, W. Kalsbeek, B. Rifkind, et al. "Total serum cholesterol levels of adults 20-74 years of age: United States, 1976-1980". Vital and Health Statistics Series 11, No. 236. DHHS Pub. No (PHD) 86-1686, Public Health Service, Washington, DC, US Government Printing Office, May 1986. Cited in Daniel (p 140, 5.4.1 Copyright 1999 by John Wiley & Sons, Inc. By permission of John Wiley.

Answer: .257

Solution:

On pp 21-22 of the supplemental notes for unit 8, we learn that

$$(\bar{X}_2 - \bar{X}_1) \text{ is distributed Normal}[\ (\mu_2 - \mu_1), \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)\ ]$$

Let $\bar{X}_1$ = Average among age 20-24. It is distributed Normal(μ_1=180, $\sigma^2_{\bar{X}_1} = \frac{\sigma_1^2}{n_1} = \frac{43^2}{50}$)

$\bar{X}_2$ = Average among age 25-34. It is distributed Normal(μ_2=199, $\sigma^2_{\bar{X}_2} = \frac{\sigma_2^2}{n_2} = \frac{49^2}{50}$)

Thus, $Y=(\bar{X}_2 - \bar{X}_1)$. is distributed Normal with

$$\mu_Y = (\mu_{\bar{X}_2} - \mu_{\bar{X}_2}) = 19$$

$$\sigma_Y^2 = \left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right] = \left[\frac{43^2}{50} + \frac{49^2}{50}\right] = 85$$

Now we use the z-score method that we learned in Unit 7, Normal Distribution, and in particular the z-score standardization that is found on page 22 under "(1)", we have that

Probability group 2 mean – group 1 mean will be more than 25

$= \Pr[Y > 25]$

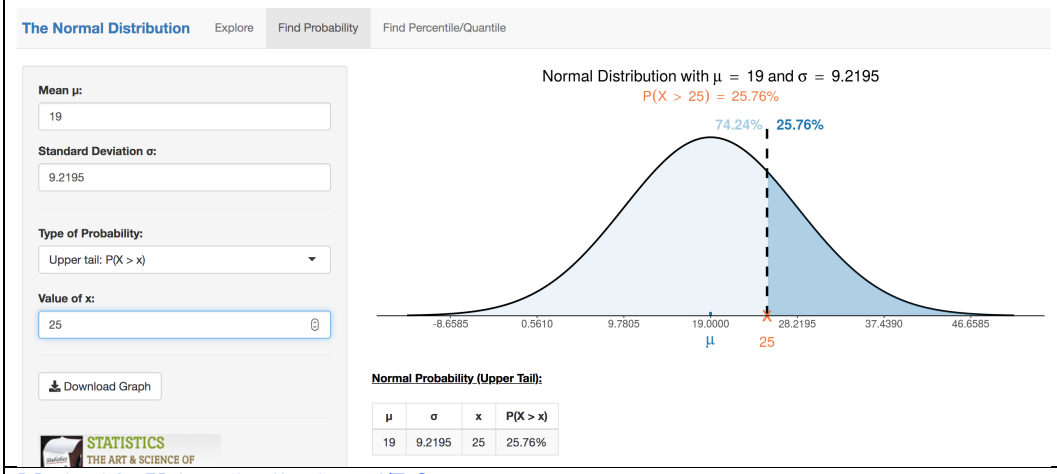
$$= \Pr\left[\frac{Y-\mu_Y}{\sigma_Y} > \frac{25-19}{9.2195}\right]$$

$= \Pr[\text{Normal}(0,1) > 0.65] = .257$

Dear Class - using www.artofstat.com, I can get the required probability in either of 2 ways: (top panel) for the distribution of Y directly, OR (bottom panel) or using its standardization to a Z-score.

www.artofstat.com > Online Web Apps > Normal Distribution > tab: Find Probability

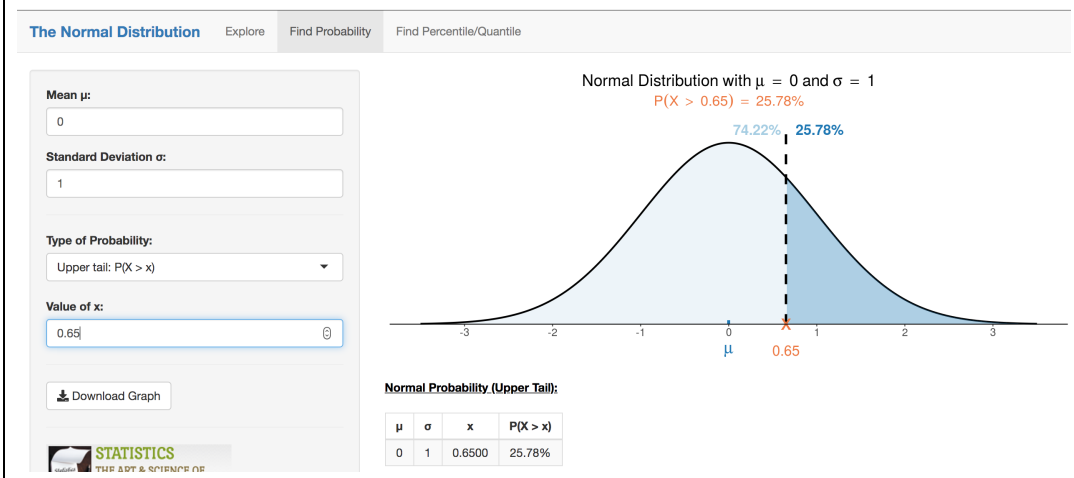
Method 1: Using distribution of $Y = \left(\bar{X}_2 - \bar{X}_1\right)$



Normal Probability (Upper Tail):

μ	σ	x	P(X > x)
19	9.2195	25	25.76%

Method 2: Using distribution of Z-Score



Normal Probability (Upper Tail):

μ	σ	x	P(X > x)
0	1	0.6500	25.78%

R

```
# Method 1:  Distribution of Y:  Calculate Pr[Normal(19, 9.2195^2) > 25]
> pnorm(25,mean=19,sd=9.2195,lower.tail=FALSE)
[1] 0.2575896

# Method 2:  Z-Score Method:  Calculate Pr[Normal(0,1) > 0.65]
> pnorm(0.65,lower.tail=FALSE)
[1] 0.2578461
```

2. A possible environmental determinant of lung function in children is the amount of cigarette smoking in the home. To study this question, two groups of children were studied. Group 1 consisted of 23 nonsmoking children aged 5-9 both of whose parents smoke in the home. Group 2 consisted of 20 nonsmoking children aged 5-9 neither of whose parents smoke. The sample mean (sample SD) of FEB1 for group 1 is 2.1 L (0.7) and for the Group 2 children, the sample mean (sample SD) of FEV1 is 2.3 L (0.4). Under the assumption of normality, compute a 95% confidence interval for the true mean difference in FEV1 between 5-9 year old children whose parents smoke and comparable children whose parents do not smoke. In developing your answer assume that the variances are NOT equal.

Answer: (-0.55, +0.15)
Solution:

The correct standard error formula to use is "Scenario #3" on page 21 of the unit 10 notes.

$$\bar{X}_1 - \bar{X}_2 = 2.1 - 2.3 = -0.2$$

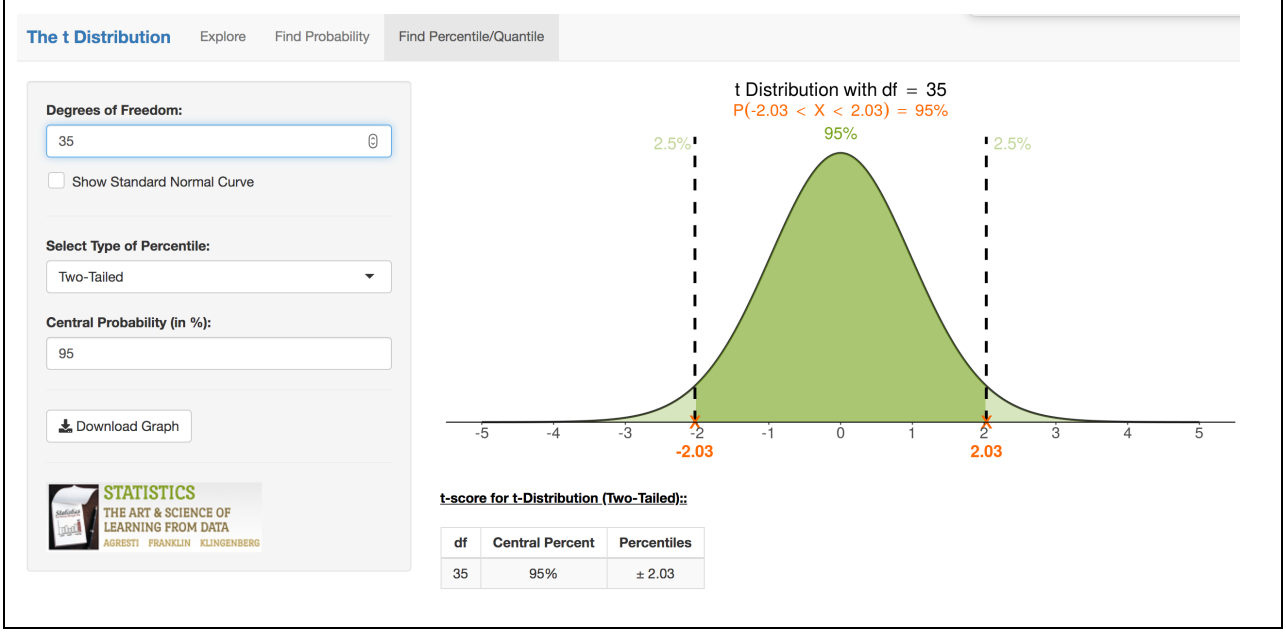
$$\hat{SE}\left[\bar{X}_1 - \bar{X}_2\right] = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = \sqrt{\frac{0.7^2}{23} + \frac{0.4^2}{20}} = 0.1712$$

$$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{\left[\frac{S_1^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{S_2^2}{n_2}\right]^2}{n_2 - 1}\right)} = \frac{\left(\frac{0.7^2}{23} + \frac{0.4^2}{20}\right)^2}{\left(\frac{\left[\frac{0.7^2}{23}\right]^2}{22} + \frac{\left[\frac{0.4^2}{20}\right]^2}{19}\right)} = \frac{0.0008587}{0.000024} = 35.78 \approx 35 \text{ by rounding DOWN}$$

$$t_{1-\alpha/2;f} = t_{.975;DF=35} = 2.03$$

$$95\%CI = \left(\bar{X}_1 - \bar{X}_2\right) \pm \left(t_{.975;DF=35}\right) \hat{SE}\left[\left(\bar{X}_1 - \bar{X}_2\right)\right] = (-0.2) \pm (2.03)(0.1712) = (-0.5475, +0.1475)$$

www.artofstat.com > Online Web Apps > t Distribution > tab: Find Percentile/Quantile



R

```
# Obtain 97.5th percentile of Student-t with DF=35
> qt(.975,df=35)
[1] 2.030108
```

3. The setting is the same as question #2. A possible environmental determinant of lung function in children is the amount of cigarette smoking in the home. To study this question, two groups of children were studied. Group 1 consisted of 23 nonsmoking children aged 5-9 both of whose parents smoke in the home. Group 2 consisted of 20 nonsmoking children aged 5-9 neither of whose parents smoke. The sample mean (sample SD) of FEB1 for group 1 is 2.1 L (0.7) and for the Group 2 children, the sample mean (sample SD) of FEV1 is 2.3 L (0.4). Under the assumption of normality, construct a 95% confidence interval for the ratio of the variances of the two groups. What is your conclusion regarding the reasonableness of the assumption of equality of variances?

Answer: (1.24, 7.37)

Solution:

Point Estimates

$S_1^2 = 0.7^2$
$S_2^2 = 0.4^2$

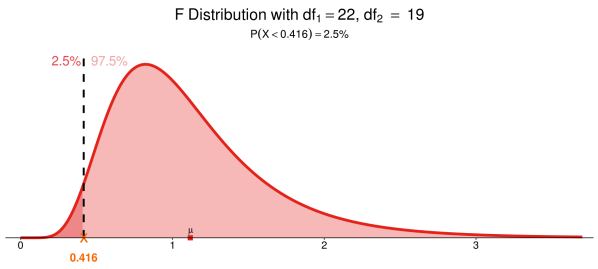
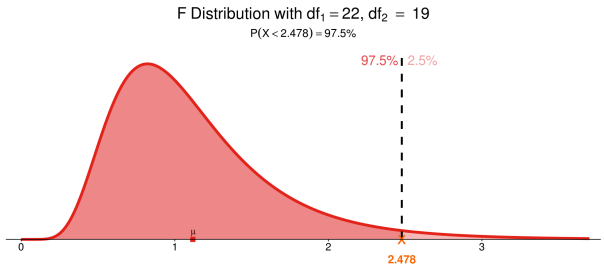
Degrees of Freedom

Numerator degrees of freedom $df_1 = (n_1 - 1) = (23 - 1) = 22$

Denominator degrees of freedom $df_2 = (n_2 - 1) = (20 - 1) = 19$

Confidence Coefficient Multipliers From the F Distribution:
For desired confidence = .95, we want the 2.5th and 97.5th percentiles

www.artofstat.com > Web Apps > t Distribution > tab: Find Percentile

Solution for 2.5th percentile of F-distribution Numerator df=22 and denominator df=19	**Solution for 97.5th percentile of F-distn Numerator df=22 and denominator df=19**
	
$F_{n_1-1,n_2-1;\alpha/2} = F_{22,19;.025} = 0.416$	$F_{n_1-1,n_2-1;1-\alpha/2} = F_{22,19;.975} = 2.478$

R

```
# Obtain 2.5th percentile of F with df1=22 and df2=19
> qf(.025,df1=22,df2=19)
[1] 0.4155022

# Obtain 97.5th percentile of F with df1=22 and df2=19
> qf(.975,df1=22,df2=19)
[1] 2.4782875
```

Putting it all together: Solution for Confidence Interval Limit Values:

$$\text{Lower limit} = \left(\frac{1}{F_{n_1-1;n_2-1;(1-\alpha/2)}}\right)\left[\frac{S_1^2}{S_2^2}\right] = \left(\frac{1}{2.478}\right)\left[\frac{0.7^2}{0.4^2}\right] = 1.2359$$

$$\text{Upper limit} = \left(\frac{1}{F_{n_1-1;n_2-1;\alpha/2}}\right)\left[\frac{S_1^2}{S_2^2}\right] = \left(\frac{1}{0.416}\right)\left[\frac{0.7^2}{0.4^2}\right] = 7.3618$$

Since the confidence interval has lower limit equal to 1.2350, a number that is above 1, these data are not consistent with the assumption of equal variances. (Logic – if the variances are equal, then their ratio is equal to 1. It then follows that, if the confidence interval for the ratio does not include 1, then the data are not consistent with the assumption of equal variances).