LAB ASSIGNMENT 4

INFERENCES FOR NUMERICAL DATA

In this lab assignment, you will analyze data from a study to compare weight and length at birth of infants born to smoking or non-smoking mothers. First, you will examine the study design and the variables that might affect the growth characteristics. Then you will apply graphical, numerical, and inferential tools in StatCrunch to compare the distributions of weight and length for each group. The assumptions required to make those inferences valid will be discussed as well. Before you start working on the assignment, you should review the course material about designing statistical studies, analysis of variance, and comparing two population means.

Smoking during Pregnancy

Medical research shows that adverse environmental conditions can affect fetal development. In particular, cigarette smoke contains thousands of different compounds that may be harmful to a developing fetus. In order to examine the relationship between maternal smoking and such growth parameters like weight and length of a newborn, three groups of mothers and their infants were examined. The mothers were 94 healthy volunteers of similar age, body mass index and education level registered in a large clinic. The first group consisted of mothers who were non-smokers (NS), the second group were moderate smokers (MS) during pregnancy (5-10 cigarettes per day), and the third group were heavy smokers (HS) (at least 15 cigarettes per day). The body length and weight measurements for full-term infants born to the three groups of mothers were determined at birth. The weight of the infants was also determined six days later.

The data from the study are available in the StatCrunch file *lab4.txt* posted on *STAT 151* Laboratories web site at http://www.stat.ualberta.ca/statslabs/stat151/index.htm (click *Stat 151* link, and *Data* for *Lab 4*). The data are not to be printed in your submission. The following is a description of the variables in the data file:

Column	Name of Variable	Description of Variable
1	LENGTH,	Length at birth (in cm),
2	WEIGHT	Weight at birth (in grams),
3	SMOKING	NS=non-smoker, MS= moderate smoker, HS= heavy smoker;
4	WEIGHT6	Weight of newborns six days later (in grams),

1. Is this an observational study or a randomized experiment? Can the data be generalized to a broader population? If infants born to smoking mothers in the study turned out to have lower birth weight than those born to non-smoking mothers, could this be used as proof that smoking lowers birth weight? What factors may be associated with smoking during pregnancy?

   There are no treatments imposed on the subjects in the study. The status of each infant (whether or not, he or she was born to a smoking or non-smoking mother) is established beyond the control of the investigator. Therefore, the study is an example of an observational study.

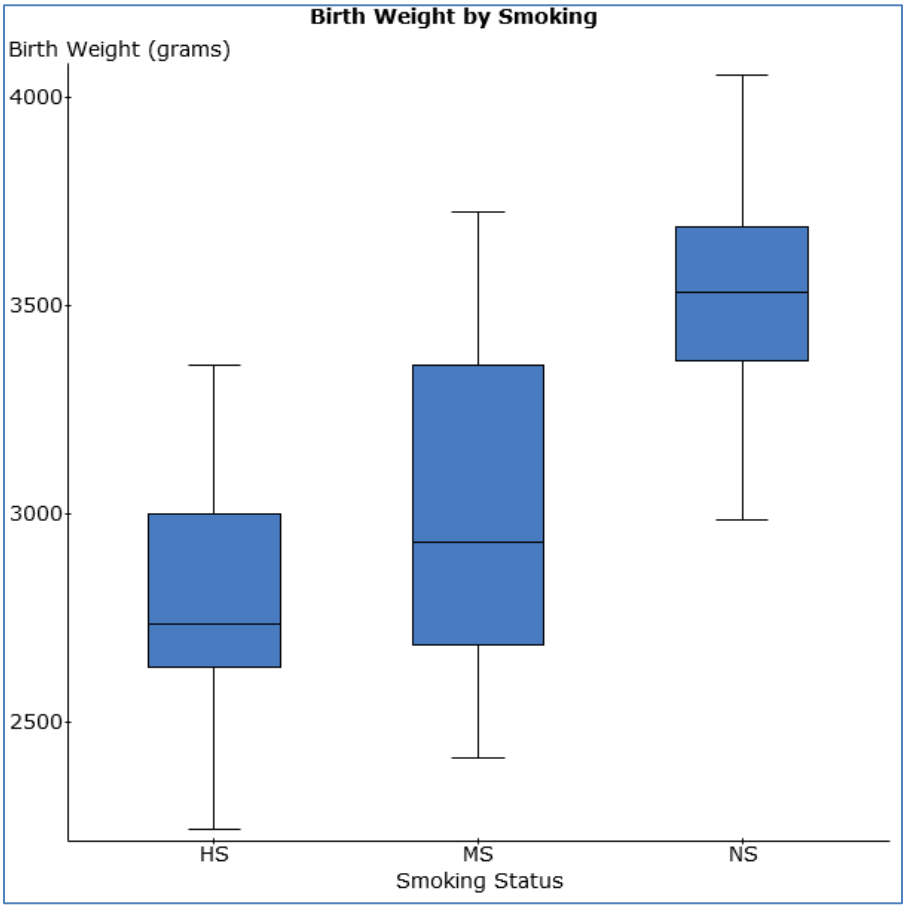
   Since the 94 infants were not randomly sampled from any well-defined population, inference to a broader population is not justified. Moreover, as the study is observational, no causal inferences about

the effect of smoking on the growth parameters can be made based on the data. In other words, based on the data alone we cannot conclude that mother's smoking during pregnancy caused their offspring to be smaller. Medical research can be used to establish the causal relationship between the growth characteristics and mothers' smoking during pregnancy.

Studies found that age, education, ethnicity, marital status, alcohol consumption, work status, and the mother's reproductive history are associated with smoking during pregnancy.

2. In this part you will display the distributions of birth weight for the three groups.

   (a) Obtain and paste the side-by-side boxplots of birth weight for the three groups. Use fences to identify outliers. Paste the plots into your report.



   (b) Comment on the center and spread of the three distributions. Comment about the most likely shape (symmetric, skewed or neither) of each distribution. Are there any outliers?

   The side-by-side boxplots show that the median birth weight of infants born to moderate smokers and that of infants born to heavy smokers are considerably lower than the median birth weight of infants born to non-smokers.

The median birth weight of infants born to heavy smokers is slightly lower than the median birth weight of infants born to moderate smokers.

The spreads of the distributions measured by the length of their corresponding boxes (IQR) are similar among the three groups. Birth weight for moderate smokers exhibits slightly larger spread than birth weight for each of the other groups.

The distributions for heavy smokers and moderate smokers seem to be slightly skewed to right. The boxplot of birth weight for infants born to non-smokers is consistent with the assumption of symmetry for the distribution (the median is halfway between the first and third quartiles and whiskers are of approximately the same length).

There are no outliers in any of the three distributions.

3. Now you will obtain numerical summaries of birth weight for the three groups.

   (a) Use the (Stat→Summary Statistics→Column) feature to compute the mean, median, standard deviation and the interquartile range for the birth weight for each of the three groups. Paste the output into your report. Comment briefly.

Summary statistics for WEIGHT:
Group by: SMOKING

SMOKING	n	Mean	Std. dev.	Median	IQR
HS	36	2764.8889	274.6476	2737	367.5
MS	19	3013.3158	396.29759	2934	671
NS	39	3534.6154	248.59644	3531	320

The mean birth weight for HS group of 2764.9 grams is considerably lower than the median for MS group of 3013.3 grams which is lower than the mean for NS group of 3534.6 grams. The ordering of the three medians exhibits similar pattern: medians for HS, MS and NS groups are respectively, 2737, 2934 and 3531 grams.

The spread measured by standard deviation and interquartile range is similar for HS and NS groups but much higher for MS group.

   (b) For each of the three groups, what birth weight corresponds to the 80^{th} percentile? Paste the appropriate output into your report. Comment briefly.

In order to obtain the weight exceeded by 20% of infants born to mothers in the three groups, we need to obtain 80^{th} percentile for each group.

Summary statistics for WEIGHT:
Group by: SMOKING

SMOKING	n	80th Per.
HS	36	3036
MS	19	3378
NS	39	3753

The 80^{th} percentile for HS group of 3036 grams is considerably lower than the 80^{th} percentile for MS group of 3378 grams which is lower than the 80^{th} percentile for NS group of 3753 grams.

4. In this part you will compare confidence intervals for birth weight for the three groups.

   (a) Use one sample tools in StatCrunch to obtain 95% confidence intervals for the mean birth weight for each group. Paste the output into your report. Which interval is the widest? Explain why it is the widest. Do the confidence intervals overlap?

One sample T confidence interval:
Group by: SMOKING
μ: Mean of WEIGHT

95% confidence interval results:

SMOKING	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
HS	2764.8889	45.774599	35	2671.9615	2857.8163
MS	3013.3158	90.916904	18	2822.3065	3204.3251
NS	3534.6154	39.807289	38	3454.0297	3615.201

The 95% confidence interval for birth weight of infants born to moderate smokers is the widest. The interval is wider than each of the two other intervals due to much larger standard error of the sample mean for the group (which is a consequence of the larger standard deviation and smaller sample size).

In general, if there was a substantial overlap for the three groups, we could claim that there is no sufficient evidence of any differences among the three groups. However, in this particular case a very small overlap exists between the confidence intervals for the moderate and heavy smokers groups relative to the widths of the two confidence intervals. The 95% confidence interval for NS groups is clearly separated from the other two confidence intervals with the lower bound higher than the right bounds of MS and HS groups. This will most likely indicate a significant difference between the mean for NS group and the means for the other two groups.

Note that the above statements about comparing the three groups cannot be made with the same level of confidence as those for individual intervals.

(b) Based on your intervals in part (a), does it appear that birth weight of infants born to smoking mothers was lower than the birth weight of infants born to non-smoking mothers?

The output in part (a) indicates that birth weight of infants born to smoking mothers (heavy or moderate) is lower than the birth weight of infants born to non-smoking mothers.

(c) A normal birth weight is considered to be 3000 g or higher. Do the data provide any evidence that the mean weight at birth of infants born to smoking mothers is less than 3000 g? Carry out the appropriate test with StatCrunch and paste the output into your report. State the null and alternative hypotheses, specify the distribution of the test statistic under the null hypothesis, report the value of the test statistic, and the p-value of the test. Answer the above question.

One sample T hypothesis test:
Where: SMOKING<>NS
μ : Mean of variable
$H_0 : \mu = 3000$
$H_A : \mu < 3000$

Hypothesis test results:

Variable	Sample Mean	Std. Err.	DF	T-Stat	P-value
WEIGHT	2850.7091	45.816753	54	-3.2584349	0.001

Define the null and alternative hypotheses as follows:

$$H_0 : \mu_{HS\ OR\ MS} = 3000,$$

$$H_A : \mu_{HS\ OR\ MS} < 3000.$$

According to the above output, the value of the test statistic, to three decimal places, is -3.258 while the output shows that the *P*-value of 0.001. The null distribution of the test statistic is a *t*-distribution with 54 degrees of freedom. With a *P*-value that small, there is convincing evidence against the null hypothesis. Thus, we reject H_0 and there is convincing evidence that birth weight of infants born to smoking mothers is less than 3000 grams, on average.

5. Now you will use two sample tools in StatCrunch to compare the average birth weight of infants born to smoking mothers and non-smoking mothers.

(a) Is there significant evidence that the average birth weight of infants born to smokers is less than the average birth weight of infants born to non-smokers? Answer the question by carrying out an appropriate test in StatCruch. Paste the output into your report. Justify your choice of the test. State the null and alternative hypotheses, specify the distribution of the test statistic under the null hypothesis, report the value of the test statistic, and the p-value of the test. State your conclusion.

Define the null and alternative hypotheses as follows:

$$H_0 : \mu_{NS} - \mu_{HS\ OR\ MS} = 0 \quad vs. \quad H_0 : \mu_{NS} - \mu_{HS\ OR\ MS} > 0$$

Two sample T hypothesis test:
μ_1 : Mean of WEIGHT where SMOKING=NS
μ_2 : Mean of WEIGHT where SMOKING<>NS
μ_1 - μ_2 : Difference between two means
H_0 : μ_1 - μ_2 = 0
H_A : μ_1 - μ_2 > 0
(without pooled variances)

Hypothesis test results:

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
μ_1 - μ_2	683.90629	60.694276	91.88896	11.268053	<0.0001

According to the above output, the value of the test statistic, to three decimal places, is 11.268 while the output shows that the *P*-value is less than 0.0001. The null distribution of the test statistic is a *t*-distribution with 91.89 degrees of freedom. With a *P*-value close to zero, there is convincing evidence against the null hypothesis. Thus, we reject H_0 and there is sufficient evidence that the average birth weight of infants born to smokers is less than the average birth weight of infants born to non-smokers

ALTERNATIVE SOLUTION:

As the ratio of the larger to the smaller standard deviation is 339.786/248.596 <2, the test with the assumption of equal standard deviations can also be used. Students *must* show that this is the case; otherwise 2 points for this part will be deducted.

Summary statistics for WEIGHT:

Group by: SMOKING MS OR HS

SMOKING MS OR HS	n	Mean	Std. dev.
false	39	3534.6154	248.59644
true	55	2850.7091	339.78613

Two sample T hypothesis test:

μ_1 : Mean of WEIGHT where SMOKING=NS
μ_2 : Mean of WEIGHT where SMOKING=MS OR SMOKING=HS
μ_1 - μ_2 : Difference between two means
H_0 : μ_1 - μ_2 = 0
H_A : μ_1 - μ_2 > 0
(with pooled variances)

Hypothesis test results:

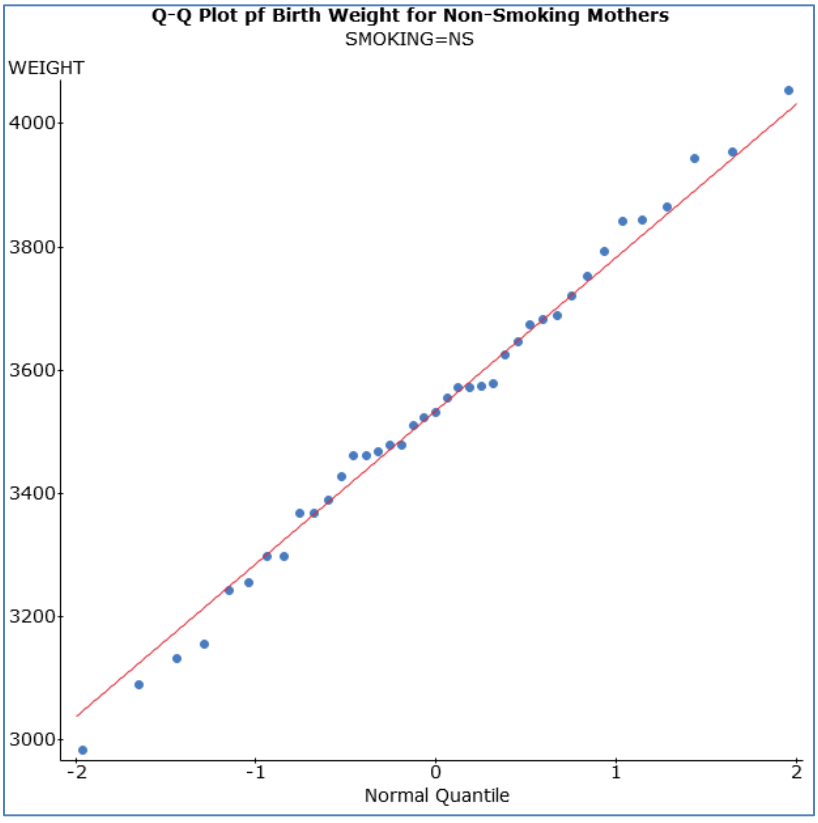
Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	683.90629	63.940323	92	10.696009	<0.0001

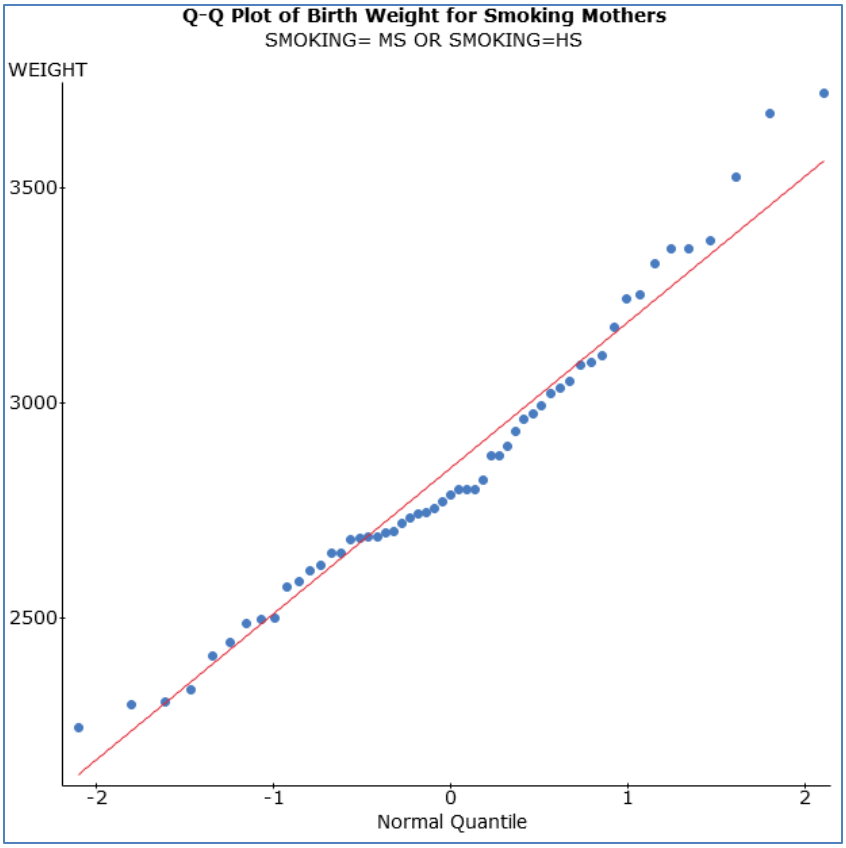
According to the above output, the value of the test statistic, to three decimal places, is 10.696 while the output shows that the *P*-value is less than 0.0001. The null distribution of the test statistic is a *t*-distribution with 92 degrees of freedom. With a *P*-value close to zero, there is convincing evidence against the null hypothesis. Thus, we reject H_0 and there is convincing evidence that the average birth weight of infants born to smokers is less than the average birth weight of infants born to non-smokers

(b) What assumptions must be satisfied to justify the procedure in part (a)? Are the assumptions satisfied in this case? Obtain a Q-Q plot for non-smoking mothers and then another Q-Q plot for smoking mothers. Paste the plots into your report and comment.

The procedure is valid as long as the subjects are randomly selected from the related population. In particular, the observations must be independent (this assumption is satisfied as long as only single births were included and mothers were not related) and no outliers are present in each group. Note that the assumption of normality is not required as the sample sizes are large (39 for NS group and 55 for the combined HS and MS groups) and the Central Limit Theorem can be applied in this case.

The Q-Q plots obtained below confirm the assumptions of normality for the two groups.





As almost all points in each plot lie close to a straight line, the data are approximately normal. The assumption of normality is feasible for the data.

(c) Report the 95% confidence interval for the difference in average birth weight between the two groups. Is your interval consistent with the outcome of the test in part (a)?

Two sample T confidence interval:
μ_1 : Mean of WEIGHT where SMOKING=NS
μ_2 : Mean of WEIGHT where SMOKING<>NS
$\mu_1 - \mu_2$: Difference between two means
(without pooled variances)

95% confidence interval results:

Difference	Sample Diff.	Std. Err.	DF	L. Limit	U. Limit
$\mu_1 - \mu_2$	683.90629	60.694276	91.88896	563.36028	804.45231

With 95% confidence, the difference in average birth weights of infants born to non-smoking and smoking mothers is between 563.360 and 804.452, to three decimal places. Since both endpoints of the interval are positive, the interval is consistent with the outcome of the test in part (a).

ALTERNATIVE SOLUTION:

A 95% confidence interval for the difference in the two means when the assumption of equal standard deviations is used is shown below:

Two sample T confidence interval:

μ_1 : Mean of WEIGHT where SMOKING=NS
μ_2 : Mean of WEIGHT where SMOKING=MS OR SMOKING=HS
$\mu_1 - \mu_2$: Difference between two means
(with pooled variances)

95% confidence interval results:

Difference	Sample Diff.	Std. Err.	DF	L. Limit	U. Limit
$\mu_1 - \mu_2$	683.90629	63.940323	92	556.91529	810.89729

With 95% confidence, the difference in average birth weights of infants born to non-smoking and smoking mothers is between 556.915 and 810.897, to three decimal places. Since both endpoints of the interval are positive, the interval is consistent with the outcome of the test in part (a).

6. In this question you will compare the birth weights of infants born to heavy and moderate smokers. In particular:

(a) Is there significant evidence that the average birth weight of infants born to heavy smokers is smaller than the average birth weight of infants born to moderate smokers? Answer the question by carrying out an appropriate test in StatCruch with the level of significance α=0.05. Justify your choice of the test. Paste the output into your report. State the null and alternative hypotheses, specify the distribution of the test statistic under the null hypothesis, report the value of the test statistic, and the p-value of the test. State your conclusion.

Define the null and alternative hypotheses as follows:

$$H_0 : \mu_{MS} - \mu_{HS} = 0 \quad \text{vs.} \quad H_A : \mu_{MS} - \mu_{HS} > 0 .$$

Two sample T hypothesis test:
μ_1 : Mean of WEIGHT where SMOKING=MS
μ_2 : Mean of WEIGHT where SMOKING=HS
$\mu_1 - \mu_2$: Difference between two means
$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 > 0$
(without pooled variances)

Hypothesis test results:

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	248.4269	101.78997	27.377515	2.4405834	0.0107

According to the above output, the value of the test statistic, to three decimal places, is 2.441 while the output shows that the *P*-value of 0.0107. The null distribution of the test statistic is a *t*-distribution with 27.38 degrees of freedom. With a *P*-value less than α=0.05, there is strong evidence against the null hypothesis. Thus, we reject H_0 and there is strong evidence that the average birth weight of infants born to heavy smokers is smaller than the average birth weight of infants born to moderate smokers.

ALTERNATIVE SOLUTION:

Based on the analysis in Question 3, the ratio of the standard deviations for the two groups is 396.298/276.648 <2, the test with the assumption of equal standard deviations is justified. Students *must* show that this is the case; otherwise 2 points for this part will be deducted.

Two sample T hypothesis test:

μ_1 : Mean of WEIGHT where SMOKING=MS
μ_2 : Mean of WEIGHT where SMOKING=HS
$\mu_1 - \mu_2$: Difference between two means
$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 > 0$
(with pooled variances)

Hypothesis test results:

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	248.4269	91.073284	53	2.7277692	0.0043

In this case, the value of the test statistic, to three decimal places, is 2.728 while the output shows that the *P*-value of 0.0043. The null distribution of the test statistic is a *t*-distribution with 53 degrees of freedom. With a *P*-value less than α=0.05, there is strong evidence against the null hypothesis. Thus, we reject H_0 and there is convincing evidence that the average birth weight of infants born to heavy smokers is smaller than the average birth weight of infants born to moderate smokers.

(b) Report the 95% confidence interval for the difference in average birth weight between the two groups. Is your interval consistent with the outcome of the test in part (a)?

Two sample T confidence interval:
μ_1 : Mean of WEIGHT where SMOKING=MS
μ_2 : Mean of WEIGHT where SMOKING=HS
$\mu_1 - \mu_2$: Difference between two means
(without pooled variances)

95% confidence interval results:

Difference	Sample Diff.	Std. Err.	DF	L. Limit	U. Limit
$\mu_1 - \mu_2$	248.4269	101.78997	27.377515	39.705815	457.14799

With 95% confidence, the difference in average birth weights for moderate and heavy smokers is between 39.706 and 457.148, to three decimal places. Since both interval endpoints are positive, the interval is consistent with the outcome of the test in part (a).

ALTERNATIVE SOLUTION:

Two sample T confidence interval:

μ_1 : Mean of WEIGHT where SMOKING=MS
μ_2 : Mean of WEIGHT where SMOKING=HS
$\mu_1 - \mu_2$: Difference between two means
(with pooled variances)

95% confidence interval results:

Difference	Sample Diff.	Std. Err.	DF	L. Limit	U. Limit
$\mu_1 - \mu_2$	248.4269	91.073284	53	65.757027	431.09677

With 95% confidence, the difference in average birth weights for moderate and heavy smokers is between 65.757 and 431.097, to three decimal places. Since both interval endpoints are positive, the interval is consistent with the outcome of the test in part (a). Note that the interval is much narrower than the confidence interval without the assumption of equal variances.

7. Do the data provide evidence that the weight of infants after six days born to non-smokers increased more than the weight of infants born to smokers, on average? Obtain a new variable DIFF6 which is the weight gain after six days and carry out the appropriate test to answer the question. Justify your choice of the test. Paste the output into your report. State the null and alternative hypotheses, specify the distribution of the test statistic under the null hypothesis, report the value of the test statistic, and the p-value of the test. State your conclusion.

Define the null and alternative hypotheses as follows:

$$H_0 : \mu_{NS} - \mu_{HS\ OR\ MS} = 0 \quad vs. \quad H_0 : \mu_{NS} - \mu_{HS\ OR\ MS} > 0,$$

where μ is the mean change in birth weight after 6 days.

As the ratio of the larger to the smaller standard deviation, 40.207/14.947>2 (see the output below), the test without the assumption of equal variances can only be used in this case.

Summary statistics for DIFF6:

Group by: SMOKING MS OR HS

SMOKING MS OR HS	n	Mean	Std. dev.
false	39	212.10256	14.946915
true	55	139	40.207333

Two sample T hypothesis test:
μ_1 : Mean of DIFF6 where SMOKING=NS
μ_2 : Mean of DIFF6 where SMOKING=HS OR SMOKING=MS
$\mu_1 - \mu_2$: Difference between two means
$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 > 0$
(without pooled variances)

Hypothesis test results:

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	73.102564	5.9263593	73.150908	12.335156	<0.0001

According to the above output, the value of the test statistic, to three decimal places, is 12.335 while the *P*-value is less than 0.0001. The null distribution of the test statistic is a *t*-distribution with 73.15 degrees of freedom. As the *P*-value is less than 0.0001, we reject H_0 and there is convincing evidence that infants born to non-smoking mothers gain more weight after six days than those born to smoking mothers, on average.

8. Now you will compare the mean length of infants born to the three groups of mothers (make sure that you use the variable LENGTH not WEIGHT).
   (a) Do the data provide evidence that not all means are equal? Answer the question by carrying out an appropriate test in StatCrunch. Paste the output into your report. State the null and alternative hypotheses, specify the distribution of the test statistic under the null hypothesis, report the value of the test statistic, and the p-value of the test. State your conclusion. What is the pooled estimate of the variance?

Analysis of Variance results:
Responses: LENGTH
Factors: SMOKING

Response statistics by factor

SMOKING	n	Mean	Std. Dev.	Std. Error
HS	36	48	3.6009523	0.60015871
MS	19	49.3	2.5024433	0.57409985
NS	39	51.1	2.8974581	0.46396461

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
SMOKING	2	181.37489	90.687447	9.3188167	0.0002
Error	91	885.58	9.7316484		
Total	93	1066.9549			

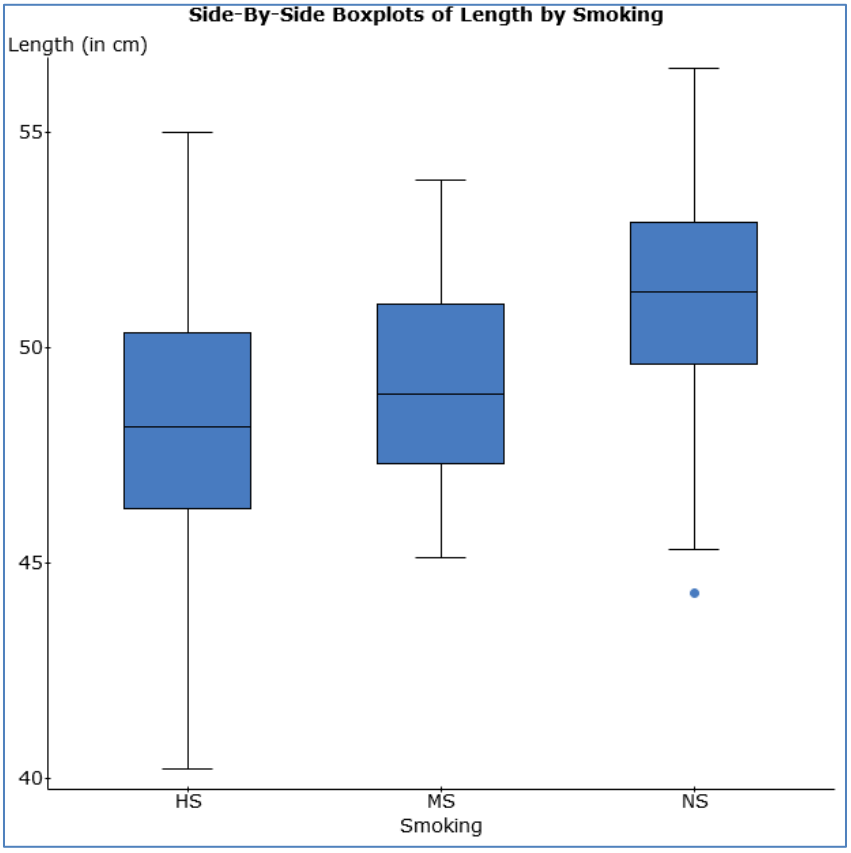
Let the means of the three goups be denoted as μ_{NS}, μ_{HS} and μ_{MS}. The null and alternative hypothesis for the One-Way ANOVA are defined as follows:

$H_0 : \mu_{NS} = \mu_{HS} = \mu_{MS}$ (all three means are equal),
$H_A :$ at least one μ_j is different (there is some difference between the means).

The value of the test statistic is 9.319 and the test statistic follows an *F* distribution with 2 and 91 degrees of freedom. The *P*-value of the test is reported as 0.0002.

There is convincing evidence that there are differences among the three group means. In other words, there are differences in birth length times among the three groups. The pooled estimate of the variance is 9.7316.

(b) Obtain side-by-side boxplots of LENGTH for the three groups. Paste the output into your report. What assumptions about the three distributions are required to make the inferences in part (a) valid for the data? Explain briefly.



The ANOVA test in part (a) is valid as long as the three groups are independent random samples, with roughly the same spread and each follows approximately a normal distribution. The side-by-side boxplots show groups with approximately the same spread (as measured by IQR) and approximately symmetric which is consistent with the assumption of normality for the data.

LAB 4 ASSIGNMENT: MARKING SCHEMA

Question 1 (8)

Observational study or a randomized experiment: 2 points
Population inferences: 2 points
Causal inferences: 2 points
Factors associated with smoking during pregnancy: 2 points

Question 2 (11)

(a) Side-by-side boxplots: 4 points
(b) Center, spread, and shape of each distribution: 6 points (2 points each)
Outliers: 1 point

Question 3 (8)

(a) Summary statistics output: 2 points
Comment: 2 points
(b) Birth exceeded by 20% of infants output: 3 points
Comment: 1 point

Question 4 (20)

(a) 95% confidence interval for the birth weight output: 3 points
Which interval is the widest and why it is widest: 2 points
Overlap and its meaning: 2 points
(b) Comparison of confidence intervals: 3 points
(c) One-sample *t*-test output: 3 points
Null and alternative hypotheses: 2 points
Value of the test statistic: 1 point
Null distribution: 1 point
P-value: 1 point
Conclusion in plain language: 2 points

Question 5 (24)

(a) Two-sample *t*-test output: 3 points
Justification of the test applied: 2 points
Null and alternative hypotheses: 2 points
Value of the test statistic: 1 point
Null distribution: 1 point
P-value: 1 point
Conclusion in plain language: 2 points
(b) Assumptions for the data: 2 points
Q-Q plots: 4 points (2 points each)
Comments: 2 points
(c) 95% confidence interval for the difference: 2 points
Consistency with the outcome of the test in part (a): 2 points

Question 6 (16)

(a) Justification of the test applied: 2 points
Two-sample *t*-test output: 3 points
Null and alternative hypotheses: 2 points
Value of the test statistic: 1 point
Null distribution: 1 point
P-value: 1 point
Conclusion in plain language: 2 points

(b) 95% confidence interval for the difference: 2 points
Consistency with the outcome of the test in part (a): 2 points

Question 7 (12)

Justification of the test applied: 2 points
Output for the test: 3 points
Null and alternative hypotheses: 2 points
Value of the test statistic: 1 point
Null distribution: 1 point
P-value: 1 point
Conclusion in plain language: 2 points

Question 8 (18)

(a) ANOVA output: 3 points
Hypotheses: 2 points
The value of the test statistic: 1 point
The distribution of the test statistic: 2 points
P-value: 1 point
ANOVA conclusion: 2 points
Pooled estimate of the variance: 1 point

(d) Assumptions for the data: 2 points
Boxplots: 2 points
Conclusions: 2 points

TOTAL = 117

Henryk Kolacz
University of Alberta
March 2018