**ANOVA**

**Problem Set**

The goal of this activity is to use ANOVA to determine if there is a difference of means and use multiple comparisons to determine where the significant differences are.

In this activity, you will carry out hypothesis testing for comparing three or more means.

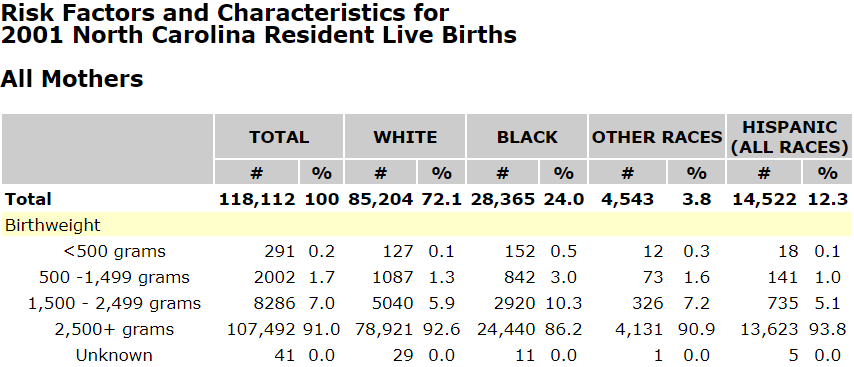
**REVIEW DATA REPORT**

North Carolina keeps extensive records on all births that take place in the state. The summaries of this data in 2001 are available here:

https://schs.dph.ncdhhs.gov/schs/births/matched/2001/all.html

The first part of the table on the website explores birthweights by race/ethnicity (see below). Some key points:

* All people who gave birth were women.
* The birthweight categories are medically relevant: a baby who is less than 2500 grams at birth (5.5 pounds) is considered low birthweight, and under 1500 grams (3.3 pounds) is very low birthweight.
* Unlike in class where we had 5 race/ethnicity categories: American Indian, Asian/Pacific Islander, Black, Hispanic, and White, the state of North Carolina chooses to group American Indian mothers and mothers of Asian and Pacific Islander heritage into one group “Other Races” and analyzes Hispanic women separately.



1. What trends do you see in the table from North Carolina?
2. Based on your answer to the previous question, do you think there are differences in birthweight by race/ethnicity?
3. What are the advantages of summarizing birthweight using categories (i.e., < 500 grams, 500-1,499 grams, etc) versus using the natural numerical measurement (i.e., 486 grams, 1,289 grams, etc.)?
4. What are the disadvantages of summarizing birthweight using categories (i.e., < 500 grams, 500-1,499 grams, etc) versus using the natural numerical measurement (i.e., 486 grams, 1,289 grams, etc.)?

**INFERENCE FOR COMPARING THREE OR MORE MEANS**

Let’s test to see if the average birthweight differs by race/ethnicity. A random sample of 1409 births that took place in North Carolina in 2001 is available on our Canvas site under the name NCbirths.csv. The data dictionary (with a description of the research) is also available under the name NCbirths-doc.docx.

1. What is the population of interest for this dataset? Do you think this sample is representative of that population?
2. Looking at the description of the research, what study design did they use? Was it a clinical trial, a cross-sectional study, a cohort study, or a case control study? Explain your reasoning.
3. What sampling method was used for this data? Explain your reasoning.

This data and the data documentation of the 1409 randomly selected births are available on Canvas, titled NCbirths.csvand NCbirths-doc.docx, respectively.

* Load the data into R or SAS.

You will be investigating two research questions in this section, comparing maternal and baby characteristics between the races/ethnicities.

**Inference for Comparing Average Birthweight by Race/Ethnicity**

Let us investigate the following research question in this part:

**Research Question: Is there a difference in average birthweight between the mother’s race/ethnicity categories?**

1. Using your software of choice, create a graphical display for the BirthWeightGm comparing the four race/ethnicity categories used by North Carolina (NCRace) (Note: **Do not** use the variables titled RaceMom or MomRace). Share your code and output.
2. Using your software of choice, calculate summary statistics for the BirthWeightGmvariable by the race/ethnicity variable used in North Carolina, NCRace. Fill in the values in the table below. Share your code and output.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Race/Ethnicity | n |  | Mean | Median | Q1 | Q3 | IQR | Range | Min | Max | SD |
| Black | 324 |  | 3132.41 | 3231.9 | 2835 | 3515.4 | 680.4 | 4280.85 | 340.2 | 4621.05 | 660.82 |
| Hispanic | 153 |  | 3360.68 | 3345.3 | 3005.1 | 3685.5 | 680.4 | 2664.9 | 2012.85 | 4677.75 | 514.48 |
| Other | 47 |  | 3327.81 | 3345.3 | 3033.45 | 3784.73 | 751.28 | 2097.9 | 1984.5 | 4082.4 | 502.03 |
| White | 885 |  | 3350.78 | 3402 | 3061.8 | 3770.55 | 708.75 | 4677.75 | 453.6 | 5131.35 | 627.84 |

1. Using the summary statistics, the graph, and the sample sizes, are the conditions met for the Central Limit Theorem for this variable? Explain.
2. Do you think we can assume equal variance? Explain why or why not.
3. What would the hypotheses be for the test? Write the hypothesis in proper notation and explain what each part of the notation means.
4. Using your software of choice, find the *F*-test statistic and the *p-*value for the test. Share your code and output.
5. Using α=0.05, what decision and conclusion would you make about the null hypothesis? Make sure to state your conclusion in the context of the problem.
6. If you rejected the null hypothesis, what would the next step in your analysis be? Based on what you suggest, about what issues would you have to be careful?

**Post-Hoc Analysis**

1. Using your software of choice*,* determine which pairwise comparisons are significantly different using the Bonferroni correction. Share your code and output.
2. Which races/ethnicities are significantly different at α = 0.05 using the Bonferroni correction? Explain your reasoning.
3. Using your software of choice, determine which pairwise comparisons are significantly different using Tukey’s HSD intervals. Share your code and output.
4. Which races/ethnicities are significantly different at α = 0.05 using the Tukey’s HSD intervals? Explain your reasoning.
5. Based on the Tukey’s HSD intervals, which race/ethnicity(s) has a significantly higher average birth weight than the others?
6. For those Tukey’s HSD intervals that were significantly different, do they appear to be practically significant? In other words, does the average difference seem notable or meaningful? Explain your reasoning.
7. Refer back to the first table from the North Carolina State Center for Health statistics (on page 1 of this activity). Do your results support or contradict the information in the table?
8. Use the information from questions 9, 13, and 20 to complete the following table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Infant birthweight (gm)  Mean (SD) | *p*-value | Post-hoc Resultsa |
| Race/Ethnicity |  | 9.52e-07 |  |
| Black |  |  | Hispanic-Black: Hispanic  White-Black: Black |
| Hispanic |  |  |
| Other |  |  |
| White |  |  |

a In the post-hoc analysis column, list estimates of the size of the difference for those groups (if any) that are significantly different and indicate which groups those estimates are for.

**Inference for Comparing Birthweight by Race/Ethnicity for the 5 Race/Ethnicity categories used in class**

Let’s compare the results using the 4 race/ethnicity categories listed on North Carolina’s website to an analysis using the 5 race/ethnicity categories used in class.

1. Using your software of choice, create a graphical display for the BirthWeightGm comparing the five race/ethnicity categories used in class (MomRace). Share your code and output.
2. Using your software of choice, calculate summary statistics for the BirthWeightGmvariable by the race/ethnicity variable for the parent giving birth, MomRace. Fill in the values in the table below. Share your code and output.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Race/Ethnicity | n | Mean | Median | Q1 | Q3 | IQR | Range | Min | Max | SD |
| American Indian |  |  |  |  |  |  |  |  |  |  |
| Asian / Pacific Is |  |  |  |  |  |  |  |  |  |  |
| Black |  |  |  |  |  |  |  |  |  |  |
| Hispanic |  |  |  |  |  |  |  |  |  |  |
| White |  |  |  |  |  |  |  |  |  |  |

1. Based on the graphs, summary statistics, and sample size, are all of the conditions met for carrying out an Analysis of Variance (ANOVA) hypothesis test? Explain your reasoning.
2. What would the hypotheses be for the test? Write the hypothesis in proper notation and explain what each part of the notation means.
3. Using your software of choice, find the *p-*value for the test. Share your code and output.
4. Using α=0.05, what decision and conclusion would you make about the null hypothesis? Make sure to state your conclusion in the context of the problem.
5. Based on your answer to question 29, explain whether a post-hoc analysis should or could be carried out. If you think post-hoc analysis should be carried out, and your test allows you to do so, then using your software of choice, perform that post-hoc analysis using either Tukey’s HSD or Bonferroni (present only one here per person) and determine which groups are statistically different and how. Share your code and output (if applicable).
6. Use the information from questions 25, 28, and 30 to complete the following table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Birthweight (gm)  Mean (SD) | *p*-value | Post-hoc Resultsa |
| Race/Ethnicity |  |  |  |
| American Indian |  |  |  |
| Asian or Pacific Is. |  |  |
| Black |  |  |
| Hispanic |  |  |
| White |  |  |

a In the post-hoc analysis column, list estimates of the size of the difference for those groups (if any) that are significantly different and indicate which groups those estimates are for, if you are able to do so given your test.

1. Examine the results using North Carolina’s four race/ethnicity categories (that grouped two different populations, American Indians and people of Asian or Pacific Islander heritage, into one “Other” category) to the five race/ethnicity category analysis done in questions 24 – 31. What results are similar? What results differ?

**FURTHER EXPLORATION**

Recall that, in class, we tested to see if there was a difference in average age for the parent giving birth by five races/ethnicities. Let us now examine the birthing parent’s ages for the four race/ethnicities used by North Carolina.

1. Using your software of choice, create a graphical display for the MomAge comparing the four race/ethnicity categories used by North Carolina (NCRace). Share your code and output.
2. Using your software of choice, calculate summary statistics for the MomAgevariable by the race/ethnicity variable, NCRace. Fill in the values in the table below. Share your code and output.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Race/ethnicity | n | Mean | Median | Q1 | Q3 | IQR | Range | Min | Max | SD |
| Black |  |  |  |  |  |  |  |  |  |  |
| Hispanic |  |  |  |  |  |  |  |  |  |  |
| Other |  |  |  |  |  |  |  |  |  |  |
| White |  |  |  |  |  |  |  |  |  |  |

1. Based on the graphs, summary statistics, and sample size, are all of the conditions met for carrying out an Analysis of Variance (ANOVA) hypothesis test? Explain your reasoning.
2. What would the hypotheses be for the test? Write the hypothesis in proper notation and explain what each part of the notation means.
3. Using your software of choice, find the *p-*value for the test. Share your code and output.
4. Using α=0.05, what decision and conclusion would you make about the null hypothesis? Make sure to state your conclusion in the context of the problem.
5. Based on your answer to question 38, explain whether or not a post-hoc analysis should or could be carried out. If you think post-hoc analysis should be carried out, and your test allows you to do so, then using your software of choice, perform that post-hoc analysis using either Tukey’s HSD or Bonferroni (present only one here per person) and determine which groups are statistically different and how. Share your code and output (if applicable).
6. Use the information from questions 34, 37, and 39 to complete the following table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mother’s age (years)  Mean (SD) | *p*-value | Post-hoc Resultsa |
| Race/ethnicity |  |  |  |
| Black |  |  |  |
| Hispanic |  |  |
| Other |  |  |
| White |  |  |

a In the post-hoc analysis column, list estimates of the size of the difference for those groups (if any) that are significantly different and indicate which groups those estimates are for, if you are able to do so given your test.

1. Compare these results to those found in lecture. What are the dangers of combining race/ethnicity categories into one “Other” category?