|  |
| --- |
| THE PENNSYLVANIA STATE UNIVERSITY |
| STAT 500 Project Report |
| Project Report |

|  |
| --- |
| 8/8/2015 |

Team 6: Daljeet Maken  
Devin Smock  
Ruthann Herstek

Table of Contents

[Section 1 Quantitative Data Cleaning 1](#_Toc342482827)

[Section 2 Qualitative Data Cleaning 4](#_Toc342482830)

[Section 3 Test Of Two Means Of Independent Samples 5](#_Toc342482833)

[Section 4 Test Of Two Means Of Paired Samples 7](#_Toc342482836)

[Section 5 Sample Size Determination For Testing A Population Mean 9](#_Toc342482839)

[Section 6 Chi-Square Test Of Independence 10](#_Toc342482842)

Section 7 List of Figures (data output)…………………………………………………………...12  
Section 8 List of Tables (graphs)…………………………………………………………………19  
Section 9 Trailing Appendix……………………………………………………………………...22

1. You will use descriptive methods such as graphical methods to locate entry mistakes for quantitative variables that you may use for the inference part. Note that data may be outliers but not errors. A data entry mistake may be a head circumference of 20 cm since it is not a realistic measurement for an adult head circumference. You only need to provide one such quantitative variable for this question and detail how you resolve the problem. However, when you proceed to work on the other problems, you should use the data that had been cleaned by you. Note that if you determine to remove a data value that is a clear error, you should replace it by \* in Minitab, denoting missing data. The reason is if you eliminate the whole row when only one variable has a mistake, you would also be removing useful information for other variables.

**Data Entry Mistakes**

The box plots for the various columns are:









The data indicates the following mistakes:

* The height column has two entries with values 5.3 each. We have marked these entries with a \*.
* The ideal-height column also has issues in the same rows as the height. The values are: 5.8 and 5.5. We have marked these entries with a \*.
  + There is another outlier row of ideal height= 80. Kept this data since it is plausible
* There is a GPA row with 56.8. We have marked these entries with a \*.
* There are three rows with head 1 size between 22 and 24. Also two rows have head 2 as 22. We have marked these entries with a \*.
* One entry for column Hours TV is 24 (assuming the unit is hrs, this is an error). Not marked this entry with \* as it is plausible. (not replaced these entries with \*)

**Resolutions of Data Entry Mistakes**

We have marked the entries highlighted above with a \*.

1. Use a graphical method to locate entry errors for qualitative variables you may use in the following problems. You only need to give one example of a qualitative variable with entry error and you should detail how to fix the problem.

Mistakes:



As shown in the bar graph for Hair Color, we have entries for color brown and Brown (difference in case). We will change all the entries with brown to be Brown. So we will get the bar chart as:

Corrected:



1. Test of two means, independent samples: Consider a scenario where a two-sample t-test is the appropriate tool of analysis. Set up a hypothesis for a one-tailed test. Show your work on checking the assumptions in order to use the two-sample t-test. Also, give some justifications for using the separate variances test or the pooled variances test that you use. What is the p-value of the test? Draw a conclusion for α=0.01. Give an interpretation of the interval. (Note: two-sample t-test is for independent samples, two measurements from the same subject usually form paired data, not independent samples)

Intro: Hypothesis: In general male students tend to be taller than female students i.e. μ(F) - μ(M) < 0

Set up the hypotheses:

**Null Hypothesis**: The average height of male students tends to be equal or smaller than the average height of female students i.e. μ(F) - μ(M) ≥ 0 or μ(F) ≥ μ(M)

**Alternate Hypothesis**: The average height of male students tend to be taller than the average height of female students i.e. μ(F) - μ(M) < 0 or μ(F) < μ(M)

Whether we can use 2-sample t-test:

Assumption 1: Are these independent samples? Yes, since the samples are from male and female students so they are not related.

Assumption 2: Are these large samples or a normal population?

**Descriptive Statistics: Height**

Variable Gender N N\* Mean SE Mean StDev Minimum Q1 Median

Height

F 23 2 64.674 0.646 3.099 60.000 62.500 64.000

M 18 0 69.958 0.696 2.952 66.000 68.000 69.000

Variable Gender Q3 Maximum

Height

F 67.000 73.000

M 71.250 77.000

We have number of males = 18 < 30, number of females = 23 < 30. We do not have large enough samples and thus we need to check the normality assumption from both populations.



From the normality plots, we conclude that both populations may come from normal distributions.

Assumption 3: Do the populations have equal variance? We will perform the Test of Equal Variances in Minitab:

**Test for Equal Variances: Height versus Gender**

Method

Null hypothesis All variances are equal

Alternative hypothesis At least one variance is different

Significance level α = 0.05

F method is used. This method is accurate for normal data only.

95% Bonferroni Confidence Intervals for Standard Deviations

Gender N StDev CI

\* 1 \* ( \*, \*)

F 23 3.09915 (2.31413, 4.62847)

M 18 2.95248 (2.13063, 4.71579)

Individual confidence level = 97.5%

Tests

Test

Method Statistic P-Value

F 1.10 0.850

The p-value = 0.850 > α = 0.01, so we fail to reject the null hypothesis and assume equal variance.

Next we perform the two sample t test with pooled variance in Minitab:

**Two-Sample T-Test and CI: Height, Gender**

Two-sample T for Height

Gender N Mean StDev SE Mean

F 23 64.67 3.10 0.65

M 18 69.96 2.95 0.70

Difference = μ (F) - μ (M)

Estimate for difference: -5.284

95% upper bound for difference: -3.675

T-Test of difference = 0 (vs <): T-Value = -5.53 P-Value = 0.000 DF = 39

Both use Pooled StDev = 3.0361  
  
Since the p-value = 0.000 < α = 0.01, we reject the null hypothesis. At 1% level of significance, the data provides sufficient evidence that the mean height of male students is greater than the mean height of the female students.

1. Test of two means, paired samples: Consider a scenario where a paired t-test is the appropriate tool of analysis. Set up a hypothesis for a two tailed test. Show your work on checking the assumptions in order to use the paired t-test. What is the p-value of the test? Draw a conclusion for α=0.05.

Intro: Hypothesis: The average graduate student wants to weigh different than their current weight i.e. µ(Ideal Weight) ≠ µ(Weight) or µ(Ideal Weight) - µ(Weight) ≠ 0

Set up the hypotheses:

**Null Hypothesis**: The average graduate student doesn’t want to weigh different than their current weight i.e. µ(Ideal Weight) = µ(Weight) or µ(Ideal Weight) - µ(Weight) = 0

**Alternate Hypothesis**: The average graduate student wants to weigh different than their current weight i.e. µ(Ideal Weight) ≠ µ(Weight) or µ(Ideal Weight) - µ(Weight) ≠ 0

We can assume normal distribution since the sample size > 30 and the probability plot of the difference is as follows:



The plot suggests that the differences may come from a normal distribution. We can proceed with a paired t-test.

Set up the hypotheses:

Ho : µ(Ideal Weight) - µ(Weight) = 0 versus Ha : µ(Ideal Weight) - µ(Weight) ≠0

**Paired T-Test and CI: Ideal-weight, Weight**

Paired T for Ideal-weight - Weight

N Mean StDev SE Mean

Ideal-weight 43 136.35 27.98 4.27

Weight 43 144.28 29.05 4.43

Difference 43 -7.93 12.82 1.96

95% CI for mean difference: (-11.88, -3.98)

T-Test of mean difference = 0 (vs ≠ 0): T-Value = -4.06 P-Value = 0.000  
  
Based on the Minitab output above, we have computed t value = - 4.06 with the p-value < 0.001 for the one-tailed test. Since p-value < 0.001 < 0.05, we can reject the null hypothesis and conclude the alternative. We conclude that the ideal-weight of the student is on average different than their actual weight.

1. Suppose you want to test whether on an average graduate students want to weigh less than what their weights actually are. How many students would you sample so that you may get a power of 0.80 if the minimum detectable difference is 1 lb and α=0.05?

*(Pay attention to estimation of the sample standard deviation, whether the data is paired or independent. Is this a one-tailed or a two tailed test?)*

Intro: The hypothesis to test here is µ(Ideal-Weight) < µ(Weight) or equivalently µ(Ideal-Weight) - µ(Weight) < 0. This is one-tailed test.

Setup the hypothesis:

**Null Hypothesis**: H0 : µ(Ideal-Weight) >= µ(Weight) or µ(Ideal-Weight) - µ(Weight) >= 0

**Alternative Hypothesis:** Ha : µ(Ideal-Weight) < µ(Weight) or µ(Ideal-Weight) - µ(Weight) < 0

Here we are working with a single sample of differences rather than a single sample of y-values. Therefore we will need n samples with each providing their weight and ideal-weight. For a One-Sided test we can calculate the number of samples by using:



Δ = -1 (since alternate is <), β = 1 – power = 0.2, α=0.05

Since we are talking about paired data, the sample standard deviation to be used here will be for the difference = µ(Ideal-Weight) - µ(Weight)

**Descriptive Statistics: IdealWeight-Weight**

Variable N N\* Mean SE Mean StDev Minimum Q1 Median

IdealWeight-Weight 43 1 -7.93 1.96 12.82 -45.00 -12.00 -9.00

Variable Q3 Maximum

IdealWeight-Weight 0.00 20.00

From the descriptive statistics we get sample standard deviation = 12.82

Using Minitab > Power and Sample Size > Paired t (with alternate <)

**Power and Sample Size**

Paired t Test

Testing mean paired difference = 0 (versus < 0)

Calculating power for mean paired difference = difference

α = 0.05 Assumed standard deviation of paired differences = 12.82

Sample Target

Difference Size Power Actual Power

-1 1018 0.8 0.800181

So we get the sample size = 1018. We will have to get the weight and ideal-weight of at least 1018 students to get the desired result.

1. Pick two qualitative variables that you are interested in and give a scenario for using the chi-square test of independence. Perform the hypothesis test and draw a conclusion for α=0.05. Check conditions to see whether you can use the chi-square test.

The qualitative variables we have chosen are Gender and Hair Dyed?. Since it is customary to define the rows using the explanatory variable and the columns using the response variable, we will make the following selections:

**Explanatory variable: Gender**

**Response variable: Hair Dyed?**

Intro: Are the two variables, Gender and Hair Dyed? independent or dependent?

Setup the hypothesis:

**Null Hypothesis**: The two categorical variables i.e. Gender and ‘Hair Dyed?’are independent.

**Alternative Hypothesis**: The two categorical variables i.e. Gender and ‘Hair Dyed?’are dependent.

The cross-tabulation from Minitab is as follows:

**Tabulated Statistics: Gender, Hair dyed?**

Rows: Gender Columns: Hair dyed?

no yes All

\* 1 0 1

F 13 12 25

M 15 3 18

All 29 15 44

Cell Contents: Count

We will exclude the row with \* as that represents missing data.

Before we proceed to the chi-square test of independence, we evaluate the condition for using the Chi-square Test. The condition used by some statisticians is to not use the chi-square test if more than 20% of the cells have expected frequencies below five. Let’s check our table:

**Tabulated Statistics: Gender\_Excl Missing, Hair dyed?\_Excl Missing**

Rows: Gender\_Excl Missing Columns: Hair dyed?\_Excl Missing

no yes All

F 13 12 25

16.279 8.721

M 15 3 18

11.721 6.279

All 28 15 43

Cell Contents: Count

Expected count

So we can see that all of our expected counts are > 5

Further, when performing chi-square test of association on rows excluding the missing data, we get:

**Tabulated Statistics: Gender\_Excl Missing, Hair dyed?\_Excl Missing**

Rows: Gender\_Excl Missing Columns: Hair dyed?\_Excl Missing

no yes All

F 13 12 25

16.279 8.721

0.6605 1.2329

M 15 3 18

11.721 6.279

0.9174 1.7124

All 28 15 43

Cell Contents: Count

Expected count

Contribution to Chi-square

Pearson Chi-Square = 4.523, DF = 1, P-Value = 0.033

Likelihood Ratio Chi-Square = 4.781, DF = 1, P-Value = 0.029  
  
We can see that Chi-Sq = 0.6605 + 1.2329 + 0.9174 + 1.7124 = 4.5232

DF = 1, P-Value = 0.033

Since the P-Value = 0.033 < α = 0.05, we reject the null hypothesis and conclude that Gender and ‘Hair Dyed?’ are dependent.

**List of Figures (Optional)**

**Our Data Outputs:**

**————— 7/22/2015 10:49:25 AM ————————————————————**

Welcome to Minitab, press F1 for help.

Retrieving project from file: ‘C:\study\psu\git\500\Projects\ourdata.MPJ’

**Chart of Hair color**

**Chart of Race**

**Chart of College**

**Chart of College**

**————— 7/23/2015 12:51:29 PM ————————————————————**

Welcome to Minitab, press F1 for help.

Retrieving project from file: ‘C:\study\psu\git\500\Projects\ourdata.MPJ’

**————— 7/27/2015 1:20:56 PM ————————————————————**

Welcome to Minitab, press F1 for help.

Retrieving project from file: ‘C:\study\psu\git\500\Projects\ourdata.MPJ’

\* NOTE \* Distribution could not be fit. The number of distinct rows of data

\* must be greater than or equal to the number of estimated

\* distribution parameters.

**Probability Plot of Height**

**Test for Equal Variances: Height versus Gender**

Method

Null hypothesis All variances are equal

Alternative hypothesis At least one variance is different

Significance level α = 0.05

F method is used. This method is accurate for normal data only.

95% Bonferroni Confidence Intervals for Standard Deviations

Gender N StDev CI

\* 1 \* ( \*, \*)

F 23 3.09915 (2.31413, 4.62847)

M 18 2.95248 (2.13063, 4.71579)

Individual confidence level = 97.5%

Tests

Test

Method Statistic P-Value

F 1.10 0.850

**Two-Sample T-Test and CI: Height, Gender**

\* ERROR \* There must be exactly two distinct subscripts.

**Two-Sample T-Test and CI: Height, Gender**

Two-sample T for Height

Gender N Mean StDev SE Mean

F 23 64.67 3.10 0.65

M 18 69.96 2.95 0.70

Difference = μ (F) - μ (M)

Estimate for difference: -5.284

95% upper bound for difference: -3.675

T-Test of difference = 0 (vs <): T-Value = -5.53 P-Value = 0.000 DF = 39

Both use Pooled StDev = 3.0361

**Two-Sample T-Test and CI: Height, Gender**

Two-sample T for Height

Gender N Mean StDev SE Mean

F 23 64.67 3.10 0.65

M 18 69.96 2.95 0.70

Difference = μ (F) - μ (M)

Estimate for difference: -5.284

95% lower bound for difference: -6.894

T-Test of difference = 0 (vs >): T-Value = -5.53 P-Value = 1.000 DF = 39

Both use Pooled StDev = 3.0361

**Two-Sample T-Test and CI: Height, Gender**

Two-sample T for Height

Gender N Mean StDev SE Mean

F 23 64.67 3.10 0.65

M 18 69.96 2.95 0.70

Difference = μ (F) - μ (M)

Estimate for difference: -5.284

95% upper bound for difference: -3.675

T-Test of difference = 0 (vs <): T-Value = -5.53 P-Value = 0.000 DF = 39

Both use Pooled StDev = 3.0361

**————— 7/27/2015 2:15:00 PM ————————————————————**

Welcome to Minitab, press F1 for help.

Retrieving project from file: ‘C:\study\psu\git\500\Projects\ourdata.MPJ’

**Results for: Worksheet 2**

**Descriptive Statistics: Q3-Height**

Variable Q3-Gender N N\* Mean SE Mean StDev Minimum Q1 Median

Q3-Height F 23 0 64.674 0.646 3.099 60.000 62.500 64.000

M 18 0 69.958 0.696 2.952 66.000 68.000 69.000

Variable Q3-Gender Q3 Maximum

Q3-Height F 67.000 73.000

M 71.250 77.000

**Probability Plot of Q3-Height**

**Test for Equal Variances: Q3-Height versus Q3-Gender**

Method

Null hypothesis All variances are equal

Alternative hypothesis At least one variance is different

Significance level α = 0.05

F method is used. This method is accurate for normal data only.

95% Bonferroni Confidence Intervals for Standard Deviations

Q3-Gender N StDev CI

F 23 3.09915 (2.31413, 4.62847)

M 18 2.95248 (2.13063, 4.71579)

Individual confidence level = 97.5%

Tests

Test

Method Statistic P-Value

F 1.10 0.850

**Two-Sample T-Test and CI: Q3-Height, Q3-Gender**

Two-sample T for Q3-Height

Q3-Gender N Mean StDev SE Mean

F 23 64.67 3.10 0.65

M 18 69.96 2.95 0.70

Difference = μ (F) - μ (M)

Estimate for difference: -5.284

95% upper bound for difference: -3.675

T-Test of difference = 0 (vs <): T-Value = -5.53 P-Value = 0.000 DF = 39

Both use Pooled StDev = 3.0361

**Results for: Worksheet 1**

**Test for Equal Variances: Height versus Gender**

Method

Null hypothesis All variances are equal

Alternative hypothesis At least one variance is different

Significance level α = 0.05

F method is used. This method is accurate for normal data only.

95% Bonferroni Confidence Intervals for Standard Deviations

Gender N StDev CI

\* 1 \* ( \*, \*)

F 23 3.09915 (2.31413, 4.62847)

M 18 2.95248 (2.13063, 4.71579)

Individual confidence level = 97.5%

Tests

Test

Method Statistic P-Value

F 1.10 0.850

**Descriptive Statistics: Height**

Variable Gender N N\* Mean SE Mean StDev Minimum Q1 Median

Height \* 1 0 68.000 \* \* 68.000 \* 68.000

F 23 2 64.674 0.646 3.099 60.000 62.500 64.000

M 18 0 69.958 0.696 2.952 66.000 68.000 69.000

Variable Gender Q3 Maximum

Height \* \* 68.000

F 67.000 73.000

M 71.250 77.000

\* NOTE \* Distribution could not be fit. The number of distinct rows of data

\* in Height (for Gender = \*) must be greater than or equal to the

\* number of estimated distribution parameters.

**Probability Plot of Height**

**Probability Plot of IdealW-W**

**Probability Plot of IdealWeight-Weight**

**Paired T-Test and CI: Ideal-weight, Weight**

Paired T for Ideal-weight - Weight

N Mean StDev SE Mean

Ideal-weight 43 136.35 27.98 4.27

Weight 43 144.28 29.05 4.43

Difference 43 -7.93 12.82 1.96

95% upper bound for mean difference: -4.64

T-Test of mean difference = 0 (vs < 0): T-Value = -4.06 P-Value = 0.000

**Test for Equal Variances: Height versus Gender**

Method

Null hypothesis All variances are equal

Alternative hypothesis At least one variance is different

Significance level α = 0.01

F method is used. This method is accurate for normal data only.

99% Bonferroni Confidence Intervals for Standard Deviations

Gender N StDev CI

\* 1 \* ( \*, \*)

F 23 3.09915 (2.16205, 5.18331)

M 18 2.95248 (1.97618, 5.39487)

Individual confidence level = 99.5%

Tests

Test

Method Statistic P-Value

F 1.10 0.850

**————— 7/29/2015 11:34:37 AM ————————————————————**

Welcome to Minitab, press F1 for help.

Retrieving project from file: ‘C:\study\psu\git\500\Projects\ourdata.MPJ’

**Descriptive Statistics: IdealWeight-Weight**

Variable N N\* Mean SE Mean StDev Minimum Q1 Median

IdealWeight-Weight 43 1 -7.93 1.96 12.82 -45.00 -12.00 -9.00

Variable Q3 Maximum

IdealWeight-Weight 0.00 20.00

**Power and Sample Size**

Paired t Test

Testing mean paired difference = 0 (versus < 0)

Calculating power for mean paired difference = difference

α = 0.05 Assumed standard deviation of paired differences = 12.82

Sample Target

Difference Size Power Actual Power

-1 1018 0.8 0.800181

**Power Curve for Paired t Test**

**Power and Sample Size**

2-Sample t Test

Testing mean 1 = mean 2 (versus <)

Calculating power for mean 1 = mean 2 + difference

α = 0.05 Assumed standard deviation = 12.82

Sample Target

Difference Size Power Actual Power

-1 2033 0.8 0.800015

The sample size is for each group.

**Power Curve for 2-Sample t Test**

**Power and Sample Size**

Paired t Test

Testing mean paired difference = 0 (versus < 0)

Calculating power for mean paired difference = difference

α = 0.05 Assumed standard deviation of paired differences = 12.82

Sample Target

Difference Size Power Actual Power

-1 1018 0.8 0.800181

**Power Curve for Paired t Test**

**Paired T-Test and CI: Ideal-weight, Weight**

Paired T for Ideal-weight - Weight

N Mean StDev SE Mean

Ideal-weight 43 136.35 27.98 4.27

Weight 43 144.28 29.05 4.43

Difference 43 -7.93 12.82 1.96

95% upper bound for mean difference: -4.64

T-Test of mean difference = 0 (vs < 0): T-Value = -4.06 P-Value = 0.000

**Paired T-Test and CI: Ideal-weight, Weight**

Paired T for Ideal-weight - Weight

N Mean StDev SE Mean

Ideal-weight 43 136.35 27.98 4.27

Weight 43 144.28 29.05 4.43

Difference 43 -7.93 12.82 1.96

95% CI for mean difference: (-11.88, -3.98)

T-Test of mean difference = 0 (vs ≠ 0): T-Value = -4.06 P-Value = 0.000

**List of Tables (Optional)**

**————— 8/6/2015 3:19:52 PM ————————————————————**

Welcome to Minitab, press F1 for help.

Retrieving project from file:

‘C:\Users\Owner\AppData\Local\Microsoft\Windows\Temporary Internet

Files\Content.IE5\582AVSTP\ourdata (1).MPJ’

**Our Cleaned Data File**

**Gender Height Ideal-Hight Weight Ideal-weight IdealWeight-Weight Hair color Hair dyed? Pairs of jeans Hub food Race Left wrist Right wrist Left arm Right arm Head 1 Head 2 GPA Hours st>> Hours TV**

**\* 68.00 68 180 136 -44 Brown no 0 \* Caucasian 17.0 16.5 70.0 71.0 59.5 59.0 3.75 9.0 9.0**

**F 63.00 66 108 115 7 Brown yes 14 good Caucasian 14.5 15.0 69.0 71.0 55.0 54.7 3.45 5.0 3.0**

**F 63.00 65 114 110 -4 Black no 4 good Asian 14.4 15.0 63.5 63.0 55.0 56.2 3.10 8.0 10.0**

**M 72.00 72 180 170 -10 Black no 4 \* \* 17.5 17.5 74.0 73.5 59.0 58.5 3.45 15.0 7.0**

**F \* \* 140 119 -21 Brown yes 8 good Other 6.0 6.0 24.0 24.0 \* \* 3.90 5.0 12.0**

**F 70.00 71 155 145 -10 Brown no 4 good Caucasian 15.5 15.7 74.5 70.5 56.2 58.5 \* 6.0 9.0**

**M 68.00 74 145 135 -10 Brown no 3 good Caucasian 16.5 16.7 74.0 73.0 57.0 58.0 3.89 5.0 10.0**

**F 62.50 \* 104 100 -4 Black no 4 good Asian 12.7 13.1 63.0 66.3 55.5 55.5 \* 0.5 0.0**

**F 62.00 70 108 99 -9 Black yes 7 good Asian 13.9 14.1 61.0 59.5 54.5 54.5 \* 1.0 0.0**

**F 65.00 65 120 120 0 Black no 2 good Asian 15.0 16.0 67.0 67.0 55.0 55.0 3.70 5.0 5.0**

**M 67.00 72 120 140 20 Black no 1 good Asian 15.5 15.7 70.5 70.5 58.3 58.4 \* 5.0 0.0**

**F 64.00 72 110 110 0 Black yes 5 poor Asian 15.0 16.1 67.0 67.5 53.7 57.8 \* 7.0 7.0**

**F 63.00 63 115 100 -15 Black yes 4 good Asian 15.0 14.5 62.0 63.0 57.5 58.0 \* \* \***

**F 62.00 65 118 110 -8 Blonde no 10 good Caucasian 15.0 14.5 61.0 64.0 51.0 52.0 3.78 4.0 10.0**

**F 73.00 73 155 155 0 Brown no 5 good Caucasian 16.0 16.0 79.0 79.0 58.0 58.0 \* 3.0 2.0**

**M 68.90 80 165 160 -5 Black no 2 good Asian 17.0 17.5 68.0 69.0 60.0 60.0 3.65 5.0 7.0**

**F 60.00 70 132 120 -12 Black no 4 good Asian 14.0 14.0 67.5 67.5 54.5 56.0 3.33 10.0 7.0**

**M 76.00 76 200 185 -15 Brown no 4 poor Caucasian 17.0 17.0 74.0 24.0 \* \* \* 0.0 \***

**M 71.00 72 150 150 0 Black no 6 good Asian 17.0 17.0 73.0 76.0 57.0 58.0 3.52 3.0 10.0**

**F 63.50 65 125 120 -5 Black no 6 good Asian 15.0 15.5 66.0 67.0 57.0 57.5 \* 6.0 5.0**

**F 65.00 68 114 105 -9 Blonde yes 40 good Caucasian 15.5 15.5 68.5 68.5 54.5 56.0 3.62 9.0 4.0**

**M 69.00 72 165 147 -18 Black no 6 good Asian 16.0 17.0 78.0 77.5 59.5 59.5 \* 16.0 6.0**

**M 77.00 74 218 216 -2 Blonde yes 4 poor Other 18.0 18.0 83.0 83.0 57.0 58.0 3.61 1.0 0.0**

**M 68.00 72 200 180 -20 Black no 8 good African American/Black 17.0 17.0 74.0 75.0 62.0 62.5 3.60 7.0 7.0**

**F 67.00 72 153 145 -8 Brown yes 12 excellent Other 17.0 18.0 64.0 69.0 57.0 56.5 3.00 3.0 5.0**

**F 67.00 70 170 125 -45 Blonde no 7 good Caucasian 16.1 16.0 73.5 71.0 56.8 54.8 3.84 10.0 12.0**

**F 62.00 \* 126 110 -16 Black yes 5 good \* 14.4 14.5 67.3 67.3 57.0 56.5 \* 3.0 5.0**

**M 72.00 72 160 171 11 Black no 2 good Asian 16.0 16.0 76.0 75.5 57.0 56.5 3.30 5.0 5.0**

**F \* \* 101 106 5 Black yes 6 good Asian 14.0 14.5 62.0 62.0 55.5 55.0 3.80 3.0 0.0**

**M 68.00 72 140 150 10 Black no 6 good Asian 15.5 15.5 66.5 68.5 58.0 58.0 3.78 2.0 10.0**

**M 69.00 72 178 170 -8 Black no 10 good Asian 16.5 16.5 72.0 73.0 58.1 58.0 3.50 3.0 10.0**

**M 70.00 72 160 180 20 Brown yes 3 poor Caucasian 15.5 16.0 68.0 67.0 57.5 57.5 3.73 2.0 6.0**

**F 64.00 70 113 105 -8 Black no 8 good Asian 5.5 5.5 26.5 26.5 \* 22.0 3.70 2.0 5.0**

**F 63.00 72 130 120 -10 Black no 8 poor Asian 14.5 15.5 64.0 64.0 55.0 55.5 3.76 15.0 12.0**

**M 71.00 72 175 175 0 Brown yes 6 good Caucasian 16.0 16.0 72.5 72.5 57.0 56.5 3.40 6.0 10.0**

**F 68.00 68 155 140 -15 Brown yes 11 good Caucasian 15.5 15.5 67.5 67.5 56.0 56.0 3.62 5.0 24.0**

**F 65.00 67 145 135 -10 Blonde yes 5 good Caucasian 15.5 15.5 64.0 63.5 58.0 57.5 3.50 14.0 14.0**

**M 68.45 70 170 140 -30 Black no 0 good Asian 18.0 18.0 67.5 67.0 59.5 60.0 \* 6.0 0.0**

**M 66.90 66 135 132 -3 Black no 6 poor Asian 18.0 18.0 68.0 70.0 58.0 59.0 3.95 6.0 10.0**

**F 61.00 66 110 100 -10 Brown yes 4 poor Asian 14.5 14.5 64.5 66.5 56.0 57.0 3.85 5.0 10.0**

**F 66.00 69 130 \* \* Blonde no 3 good Caucasian 14.0 15.0 64.5 66.5 55.5 55.5 3.87 4.0 5.0**

**M 71.00 72 170 160 -10 Black no 5 good Asian 19.0 18.5 68.0 69.0 60.0 60.3 3.70 6.0 5.0**

**M 66.00 67 140 130 -10 Blonde no 4 good Caucasian 16.0 16.0 66.0 66.0 56.0 56.0 3.70 4.0 4.0**

**F 68.50 67 132 122 -10 Black no 5 good Asian 15.0 15.0 68.0 68.0 57.0 57.5 \* 8.0 1.5**

**Trailing Appendix**

Additional discoveries for Question Two:

We also have issues with column: Hub Food



The data entry form has no entry for unknown, but we have occurrences for this value. We will mark these as \*.

We again have issues with having same value in different cases within the race column:



We will change all “other” to “Other”

There are many discrepancies in the column “Chart of College”. This type of column is best handles by providing options.



There are numerous entries that refer to same entity but are stated differently:

* Agricultural Sciences, College of Agricultural Sciences, College of Agriculture. Merged as College of Agricultural Sciences
* Education and College of Education. Merged as College of Education
* Engineering and College of Engineering. Merged as College of Engineering
* College of Health and Human Development vs College of Human and Health Development. There is also an entry marked “Health”. Merged all as College of Human and Health Development
* College of Science vs College of Science, IGPG. Merged as College of Science

The bar chart after making the above corrections is:

