BIOSTATISTICS 600

Global Activity Two

Male Circumcision and HIV Infection in African populations: Revisited

ANSWER KEY

INTRODUCTION

Many studies have investigated the association between *Male Circumcision* and *HIV Prevalence*. Global Activity One explored this relationship in 37 African countries using a continuous measure for *Male Circumcision* and for *HIV Prevalence*.

A more recent article (Drain 2006) explores the relationship between *Male Circumcision* and *HIV Prevalence* in a somewhat different way. *HIV Prevalence* was estimated from the Joint United Nations Programme on HIV/AIDS. *Male circumcision* rates were categorized as either "Low" (<20%), "Intermediate" (20-80%), or "High" (>80%). Investigators found a significantly higher prevalence of HIV in African countries with "Low" rate of male circumcision compared to African countries with a "High" rate of male circumcision from a two- sample t-test. Investigators also explored the relationship between the two factors using univariate linear regression and correlation. Using the data provided (primarily updated values from the Joint United Nations: Report on the Global HIV/AIDS epidemic 2008), students will investigate the relationship between *Male Circumcision* and *HIV Prevalence* in Africa using methods based on the (Drain 2006) article. Students will reproduce some results from the original article as well as develop some new results.

SOURCES

Drain PK, Halperin DT, Hughes JP, Klausner JD, Bailey RC. 2006. Male circumcision, religion and infectious diseases: an ecologic analysis of 118 developing countries. BMC Infectious Diseases 6:172.
Open Access: www.biomedcentral.com/1471-2334/6/172

Joint United Nations Programme on HIV/AIDS: Report on the Global HIV/AIDS Epidemic 2004 Geneva: UNAIDS; 2004.

Joint United Nations Programme on HIV/AIDS: Report of the Global HIV/AIDS Epidemic 2008 Mexico City: UNAIDS; 2008

Note: Data for questions based on the (Drain 2006) article were obtained from http://www.unaids.org/en/KnowledgeCentre/HIVData/GlobalReport/Archive.asp (accessed June 2, 2009)

QUESTIONS

For the following problems, refer to the data set "GA_Two_Drain.xls". Questions are based on the methods and findings from the article Drain (2006). Students' answers may be different in some cases from the original article due to updated data sources and other factors. Forty-four African nations were grouped into three categories based on the percentage of males circumcised: *Low (<20%), Intermediate (20-80%)* or *High (>80%).* Unless otherwise stated, use the column labeled *2007 HIV prevalence percent Adult (15–49)* as the outcome variable.

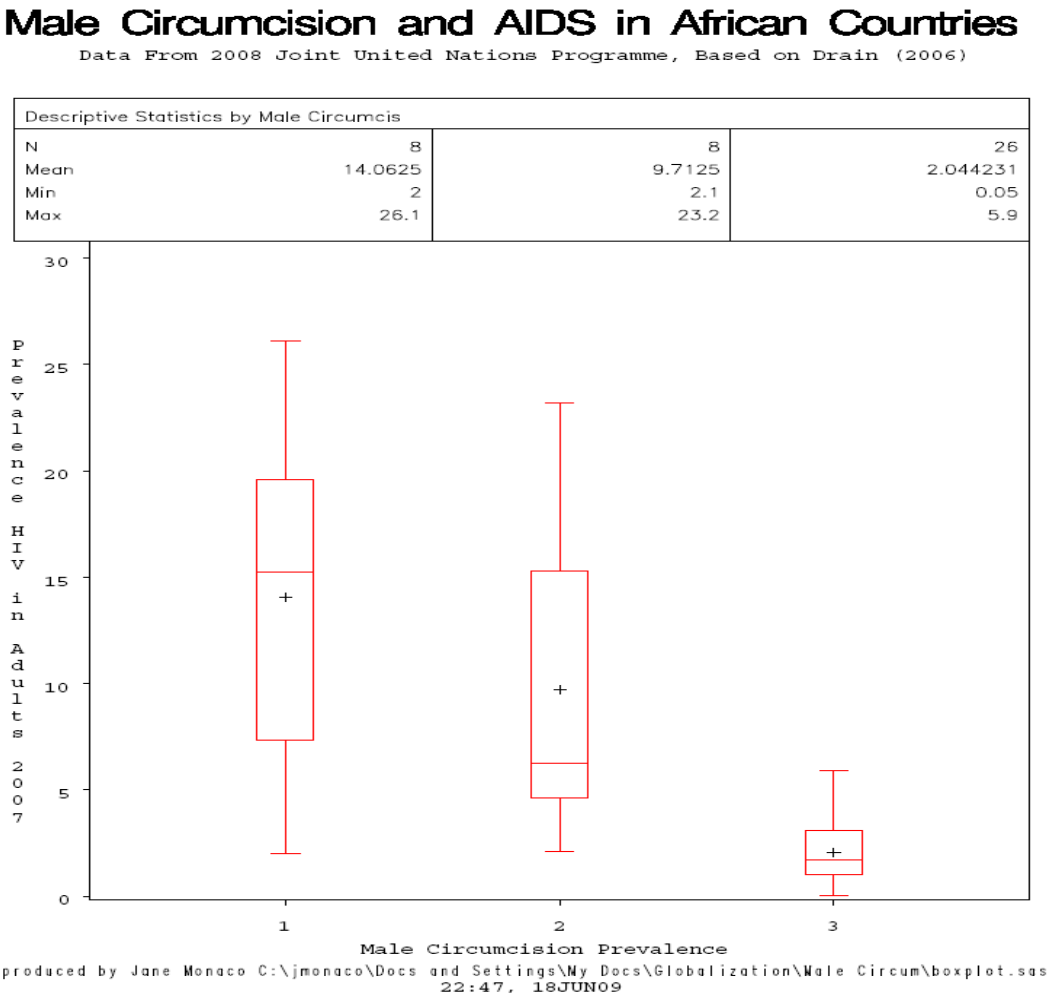
1. Calculate the mean, standard deviation, median, and IQR for the *2007 HIV Prevalence Percent* for each *Male Circumcision Category*. Report your results in a well-organized table.

Table 1: Descriptive Statistics for HIV Prevalence Percent in 2007 by Male Circumcision Category for 42 African Countries

Male Circumcision Category	Number of Countries	Mean (S.D.)	Median (IQR)
Low(<20%)	8	14.1 (8.6)	15.3 (9.6,17.5)
Medium(20-80%)	8	9.7 (7.5)	6.3 (5.0,13.9)
High(>80%)	26	2.0 (1.4)	1.7 (1.1,3.1)

2. Construct side-by-side boxplots for *2007 HIV Prevalence Percent* for the three *Male Circumcision Categories* (by hand or software). Describe the relationship.

2.



2 (cont). In SAS:

```
proc boxplot data=ga_two;          * Use Global Activity Two data;
      * HIV Prevalence 2007 (continuous) and Male Circumcision (3 categories);
      plot hivpr_07*mc_num /boxstyle=schematic;
      insetgroup n mean min max /header='Descriptive Stats by Male Circ
            Category';  * display descriptive statistics;
      run;
```

Or you can calculate the 5 number summary in Excel and draw the boxplots by hand.

Countries with the lowest male circumcision rates in general had higher percentage of adults with HIV (median =15.3%) and their HIV prevalence varied widely (IQR = (9.6-17.5%)). Countries with the highest male circumcision rates (>80%) had a lower percentage of HIV prevalence (median = 1.7%) and these rates were less variable (IQR = (1.1 -3.1%)).

3. Conduct a statistical test to compare the average *2007 HIV Prevalence* in the two *Male Circumcision Categories: Low*. Vs. *High*. Include these steps: State the null and alternative hypothesis. Check the assumptions. Calculate an appropriate test statistic and corresponding *p*-value. Interpret the *p*-value for a nonstatistician.

3. We will test the population average *HIV Prevalences* equal in the two groups using a two-sample *t*-test. Note the unit of observation is "country" not individual patients.

State Null and alternative hypotheses:

H_0: The average *HIV Prevalence* in all African countries with *Low* male circumcision rate is equal to the average *HIV Prevalence* in all African countries with *High* male circumcision rate.

H_a: The averages are different in these two groups of countries.

$H_0: \mu_{Low}=\mu_{High}$, $H_a: \mu_{Low}\neq\mu_{High}$

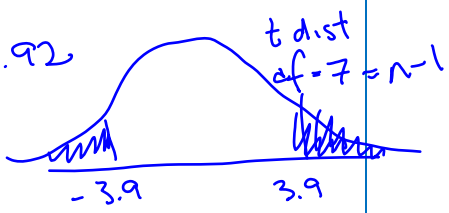
Check Assumptions:

- The two groups of countries are distinct groups (no overlap)
- The responses in each country are independent
- The sample of countries should be a simple random sample from all possible countries in the population (of countries in Africa). This assumption is not met. More on this later.
- HIV Prevalence should be normally distributed in the two groups. This assumption is also not met based on the descriptive statistics and boxplots above. The data are moderately skewed. However we know that the two-sample t-test is robust – meaning that with sufficient sample size the assumption can be relaxed. One common "rule of thumb", is that if $n_1 + n_2 > 39$, we can proceed with the two-sample t-test, even if the population values are not normally distributed. (See your textbook for complete description.)

Calculate test statistic

By hand

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{14.06 - 2.04 - 0}{\sqrt{\frac{9.64^2}{8} + \frac{1.44^2}{26}}} = 3.92$$



p-value: *2*Pr(t>3.92)= 0.006*

Or with software

t-Test: Two-Sample Assuming Unequal Variances

Testing Equality of Average *HIV Prevalence*

	Low Male Circumcision	*High Male Circumcision*
Mean	14.0625	2.044231
Variance	74.72268	2.084865
Observations	8	26
df	7	
t Stat	3.915655	
P(T<=t) two-tail	0.005781	

(Check that the values for mean and variance are the same as we calculated in the boxplot analysis. When calculating by hand, carry ALL decimals but round your final answers for presentation.)

Interpret Results:

We find *p*= 0.006. If the average *HIV prevalence* in all African countries with *low male circumcision* rates were equal to the average *HIV prevalence* in all African countries with *high circumcision rates*, the probability that we'd observe two sample average prevalences as extreme as ours (14% vs. 2%) is extremely small. We'd expect this to happen by chance only 0.6% of the time.

However, recall our sample of countries is not a Simple Random Sample. It is a sample of most African countries – we need to investigate which countries were excluded and why. Specifically we would like to know how countries left out of the sample are different from the countries in the sample with respect to the two factors of interest.

Note that this study would ideally like to include ALL African countries (rather than a sample), and it comes very close! If estimates were available in all African countries, then could compute the ACTUAL population means, μ_{Low} *and* μ_{High}, rather than just estimates (xbar values). In this case, it would not be necessary to compute a statistical test to compare averages in the two groups….we would have the actual population means. One could argue in this example that we have such a large sample of countries relative to all the countries, that computing a statistical test to compare the means is not informative. However, most readers expect to see this *p*-value, regardless of this point. The original article did compute such a *p*-value.

1

4. Calculate a 95% Confidence interval for the difference in average 2007 *HIV prevalence* in the two groups, *Low* vs. *High Male Circumcision Categories*. Carefully interpret the confidence interval for someone who has no statistical training.

4.

CI: $\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 12.01 \pm 2.365\ (3.069) = (4.8\%, 19.3\%)$

The *t*-value is based on *degrees of freedom = min(n_1-1,n_2-1) = 7.* Approximately 95% of confidence intervals computed in this way will contain the true difference in means in the two groups. So if we repeatedly took samples from the population of African countries and computed the difference in mean *Prevalence of HIV* for each sample and the associated CI, about 95% of CI will contain the true difference. So we can be reasonably confident that the true difference in means is between 4.8% and 19.3%. Note that 0% is not in the confidence interval.

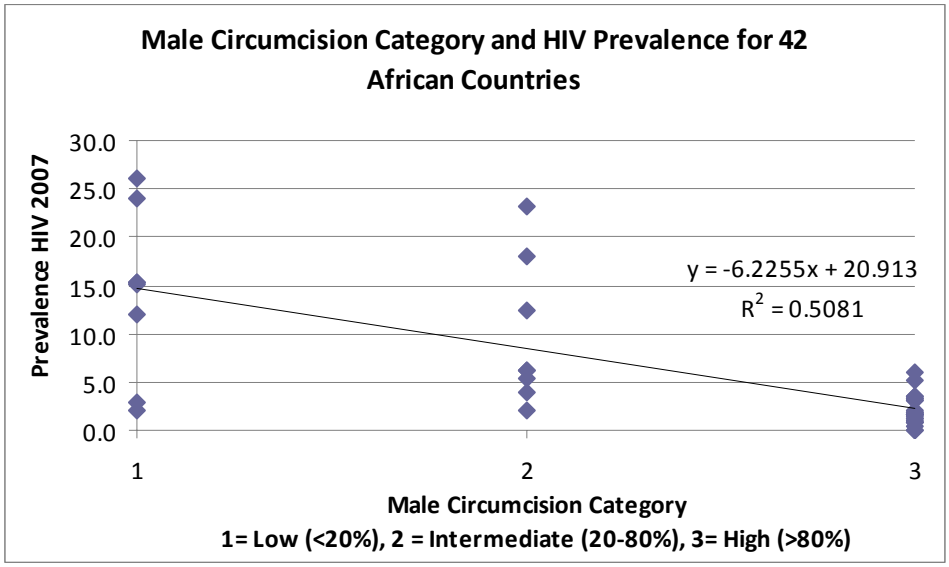
5. i) The three categories for *Male Circumcision* were coded as *1=Low, 2=Intermediate, 3= High* in the original journal article (Drain 2006). Those values are provided in the given dataset. Produce a scatterplot with *x= Numerical Category of Male Circumcision* and *y= 2007 Prevalence of HIV in Adults.* Include the linear regression equation fit to the data.

ii) Conduct a linear regression analysis for *x=Numerical Category of Male Circumcision* and *y= 2007 Prevalence of HIV in Adults*. Include a plot of the residuals. Check the assumptions. What assumption(s) are violated?

iii) There are more countries with *Numerical Category of Male Circumcision =3* than the other categories. Is this problematic?

iv) Consider the coding of the categories as *1=Low, 2=Intermediate, 3= High*. Do you think it is valid to assume equal spacing for these three categories?

5. i)



5. ii)

Regression Statistics	
Multiple R	0.71284
R Square	0.50814
Observations	42

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1018.766	1018.766	41.32399	1.18E-07
Residual	40	986.1253	24.65313		
Total	41	2004.891			

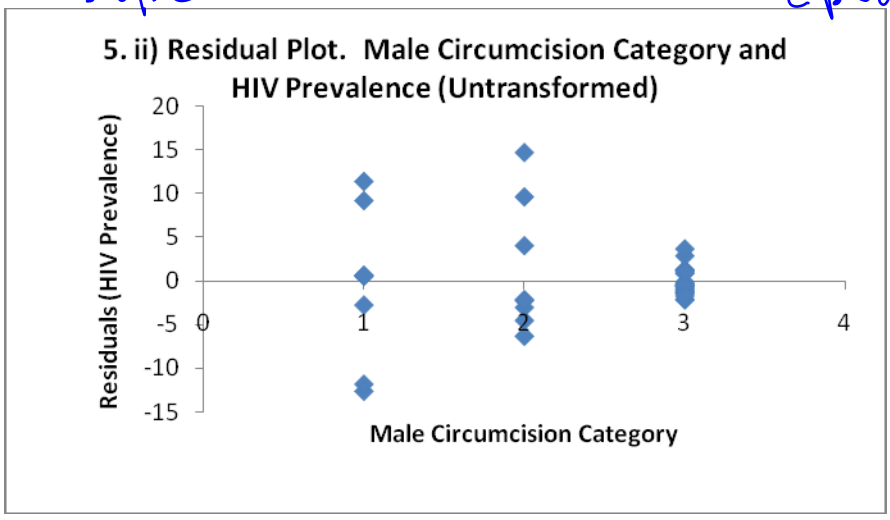
df

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	20.91322	2.473585	8.454622	1.93E-10	15.91392	25.91253
X Variable 1	-6.22554	0.968448	-6.42837	1.18E-07	-8.18285	-4.26824

y intercept

slope

p value



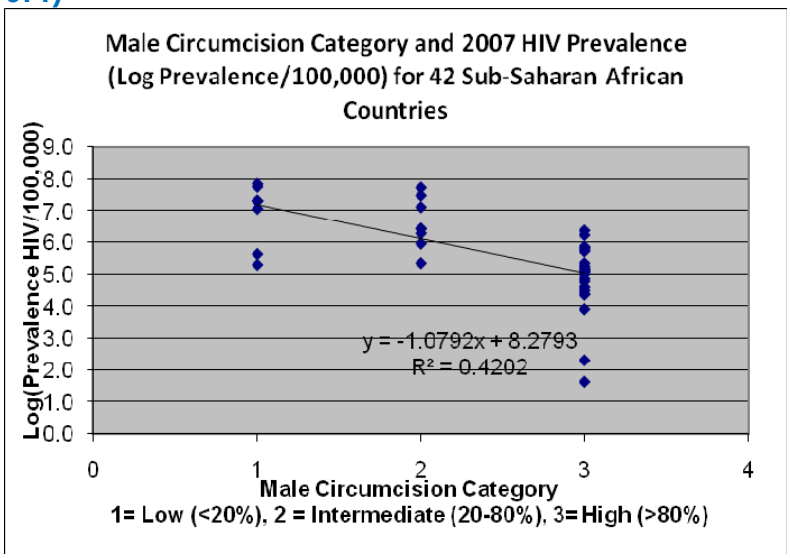
The equation of the regression line is $y = 20.9 - 6.2x$.
Checking assumptions: the relationship is approximately linear. The errors are approximately normally distributed. However, the errors do not have constant variance. The variance is larger for $x = 1$ and 2 than when $x = 3$. In order to address this problem, we will transform the y variable in Question 6 by log transforming the *HIV Prevalence* values.
Also, as noted previously the countries are not a SRS from all African countries.

5. iii) Having more data points when $x=3$ (*Male Circumcision=High*) is not problem. No assumptions are made about the data being evenly distributed across the x values.

5. iv) Opinions may differ here. The authors of the this paper (Drain 2006) take a different way of viewing the male circumcision categorically rather than continuously (Bongaarts 1989). One advantage to this method is that errors in measurement of the x variable have less impact when viewed categorically. The assumption that the three categories are equally spaced may not be valid. We could investigate the impact of the assumption by using different spacing for the three categories and seeing the difference in the results.

6. i) Produce a scatterplot with *x= Numerical Category of Male Circumcision* and *y= ln(2007 Prevalence of HIV per 100,000 Adults).* These *y* values are calculated for you in the given dataset. Include the linear regression equation fit to the data.
ii) Conduct a linear regression analysis for *x=Numerical Category of Male Circumcision* and *y= ln(2007 Prevalence of HIV per 100,000 Adults).* Include a plot of the residuals. Check the assumptions. Do the log transformed data in this problem meet the linear regression assumptions better than the original data above?

6. i)



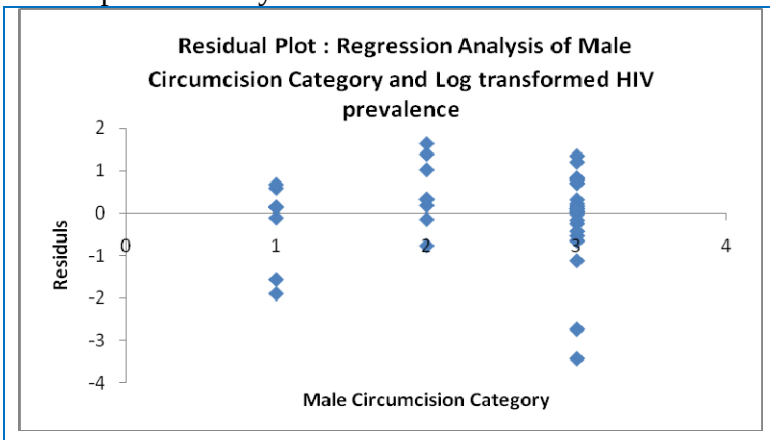
6. ii)

Regression Statistics	
Multiple R	0.64823
R Square	0.420203
Observations	42

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	30.61613	30.61613	28.98962	3.46E-06
Residual	40	42.24426	1.056107		
Total	41	72.8604			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	8.279333	0.51197	16.17152	3.82E-19	7.244603	9.314063
X Variable 1	-1.07923	0.200444	-5.3842	3.46E-06	-1.48435	-0.67412

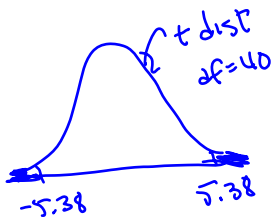


In checking the assumptions, the data appear approximately linear. The errors are approximately normally distributed, with the variance somewhat larger when $x=3$. The transformed data do appear to meet the linear regression assumptions better than the untransformed data.
The equation for the regression line is $y=8.28-1.08x$, with $p<0.001$ for testing the slope equal to zero.
Our results using the 2007 data ($n=42$, $beta = -1.08$, $r^2=0.42$) are similar to the original findings by Drain 2006 ($n=38$, $beta=-0.90$, $r^2=0.51$).

7. Calculate the correlation coefficient and the coefficient of variation between *Numeric Category of Male Circumcision* and *log(2007 HIV Prevalence Percent in Adults)*. (Use the transformed data, as in Question 6). Conduct a statistical test for the correlation coefficient equal to zero and carefully interpret the *p*-value.

7.
From the regression analysis above, $r=-0.65$, $r^2=0.42$.
Significance test:
H_0: $\rho = 0$, H_a: $\rho \neq 0$,

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.65\sqrt{40}}{\sqrt{1-0.42}} = -5.38$$



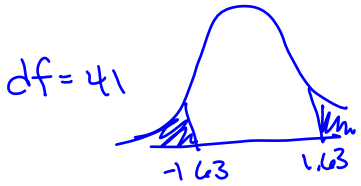
$t=-5.38$, $df=40$, $p<0.001$. Notice the t statistic value is the same as in #6.
If ρ, the correlation coefficient between *Male Circumcision Category* and *Log(HIV Prevalence)* were equal to zero in the population of all African countries (n=approximately 56 countries) then the probability that we'd find a sample correlation as extreme as $r=-0.65$ is very small. This is strong evidence the true correlation is likely not equal to zero.
[See previous problems for discussion about the assumptions.]

8. Investigators are interested in the change in average *HIV Prevalence* between the years 2001 and 2007. Conduct a statistical test to address this question.

8.
We will conduct a matched pairs t-test. We are given, essentially, "before and after" measures on the same units of observations (countries). Therefore, we can calculate the *2007 HIV prevalence* minus the *2001 HIV Prevalence* for each country. Then conduct a statistical test for the average of those differences equal to zero by computing the average difference (*dbar*) and its standard deviation (s_{dbar}) and corresponding test statistic.

$H_0: \mu_{2001}=\mu_{2007}$, $H_a: \mu_{2001}\neq\mu_{2007}$

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{-0.469 - 0}{1.8646/\sqrt{42}} = 1.63$$



p=2 Pr(t< -1.63) = 0.11
If there were no difference in the average *HIV Prevalence* in 2001 and 2007, we'd expect a sample average difference as extreme as -0.47 about 11% of the time. Our results are not unusual if there is no difference in the two time points.

9. Suppose investigators wished to compare the three mean *HIV Prevalences* for all the *Male Circumcision Categories* (rather than just *Low* vs. *High* in previous problem and original article). Conduct a statistical test to compare the three averages (*Low*, *Intermediate* and *High*).

9. We will compare the three average *HIV Prevalences* using One-Way ANOVA.

$H_0: \mu_{Low}= \mu_{Intermediate}= \mu_{High}$,
H_a: Not all the means are the same in the three groups of countries.

We first check the assumptions for the One-Way ANOVA.

In SAS:

```
proc glm data = ga_two;
        class mc_num;                *Male Circumcision is 3 level categorical variable;
        model hivpr_07= mc_num ;     *HIV Prevalence 2007 is continuous variable;
        means mc_num;                *Displays the averages and var for each MC level;
        run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1035.361877	517.680939	20.82	<.0001
Error	39	969.529135	24.859721		
Corrected Total	41	2004.891012			

R-Square	Coeff Var	Root MSE	hivpr_07 Mean
0.516418	86.05301	4.985952	5.794048

Source	DF	Type I SS	Mean Square	F Value	Pr > F
mc_num	2	1035.361877	517.680939	20.82	<.0001

Level of mc_num	N	hivpr_07	
		Mean	Std Dev
1	8	14.0625000	8.64422805
2	8	9.7125000	7.50570021
3	26	2.0442308	1.44390629

Or in Excel: Cut and paste the *HIV Prevalence Values* into three columns with one column for each *Male Circumcision Level*. Then select: >Tools > Data Analysis > ANOVA: Single Factor. Highlight the values in the three columns with the *HIV Prevalence Values*. Click <OK>.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	8	112.5	14.0625	74.72268
Column 2	8	77.7	9.7125	56.33554
Column 3	26	53.15	2.044231	2.084865

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1035.362	2	517.6809	20.82408	7.03E-07	3.238096
Within Groups	969.5291	39	24.85972			
Total	2004.891	41				

We find strong evidence against that the assumption that all three population mean *HIV Prevalences* are the same. (Not a surprising result, since we previously found a significant difference between the *Low* and *High* groups.) If the mean *HIV Prevalence* were the same in the three populations, the probability we have sample averages as extreme as in this case (14.1%, 9.7%, and 2.0%) is very small ($p<0.0001$).

We could continue to explore this relationship by testing which pairs of means are different.

Additional Sources regarding the relationship between *HIV* and *Male Circumcision* in Africa:

Auvert B, Taljaard P, Lagarde E, Sobngwi-Tarbekou J, Sitta R, Puren A. 2005. Randomized, Controlled Intervention Trial of Male Circumcision for Reduction of HIV Infection Risk: The ANRS 1265 Trial. PLoS Med 2(11):e298, 1112-22. www.plosmedicine.org

Bailey RC, Moses S, Parker CB, Agot K, Maclean I, Krieger JN, Williams CFM, Campbell RT, Ndinya-Achola JO. 2007. Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomized, controlled trial. The Lancet 369: 643-56.

Bailey RC, Plummer FA, Moses S. 2001. Male circumcision and HIV prevention: current knowledge and future research directions. The Lancet Infectious Disease 1:223-31.

Bongaarts J, Reining P, Way P, Conant F. 1989. The relationship between male circumcision and HIV infection in African populations. AIDS 3:373-7.

BIOSTATISTICS TOPICS:

DESCRIPTIVE STATISTICS, TWO-SAMPLE T-TEST, CONFIDENCE INTERVAL FOR DIFFERENCE IN MEANS, LINEAR REGRESSION, DATA TRANSFORMATION, CORRELATION, MATCHED PAIRS T-TEST, ANOVA.