BIOSTATISTICS 600

Global Activity One

Male Circumcision and HIV Infection in African populations

ANSWER KEY

INTRODUCTION

Many studies have investigated the association between male circumcision and HIV prevalence. One of the first studies (Bongaarts 1989) examined the relationship between percentage of males who were circumcised in 37 African countries and the HIV seroprevalence in those countries based on estimates from the capital city. The authors report the two factors were strongly correlated ($p<0.001$). In the following exercises, students will explore the relationship between *Percent of Males Circumcised* and *HIV Seroprevalence* by reproducing many of the calculations in the original journal article and expanding on those findings.

SOURCE

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

QUESTIONS

Use the data provided (published in Bongaart (1989) and given in GA_One_Bongaarts.xls) to complete the following questions:

1. Provide descriptive statistics for *x= Percent Males Circumcised* and for *y= Percent HIV Seroprevalence* in one neat, well-labeled table. Explain your reasoning for your choice of descriptive statistics.

1.

Table 1: Descriptive Statistics for 37 African Countries

	Median (IQR)
Male Circumcision Rate	95% (65 – 100%)
HIV Seroprevalence	1.6% (0.0 - 3.9%)

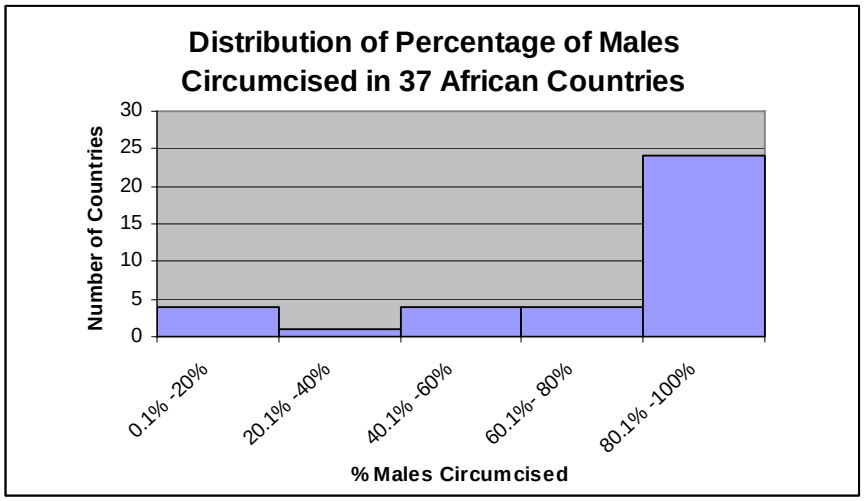
Using the median and the IQR are preferred (rather than the mean and SD) because the data are skewed.

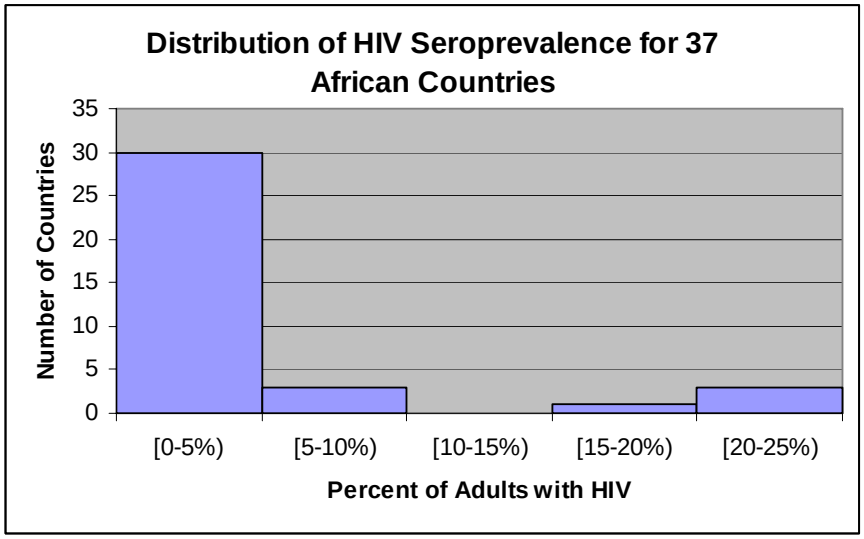
2. a. Produce a histogram for *x=Percent Males Circumcised*. Do the data appear to be normally distributed?

b. Produce a histogram for *y=Percent HIV Seroprevalence*. Do the data appear to be normally distributed?

c. Which do you prefer to describe the factors, the descriptive statistics (1.) or the histograms (2a.,b.)? Why?

2.





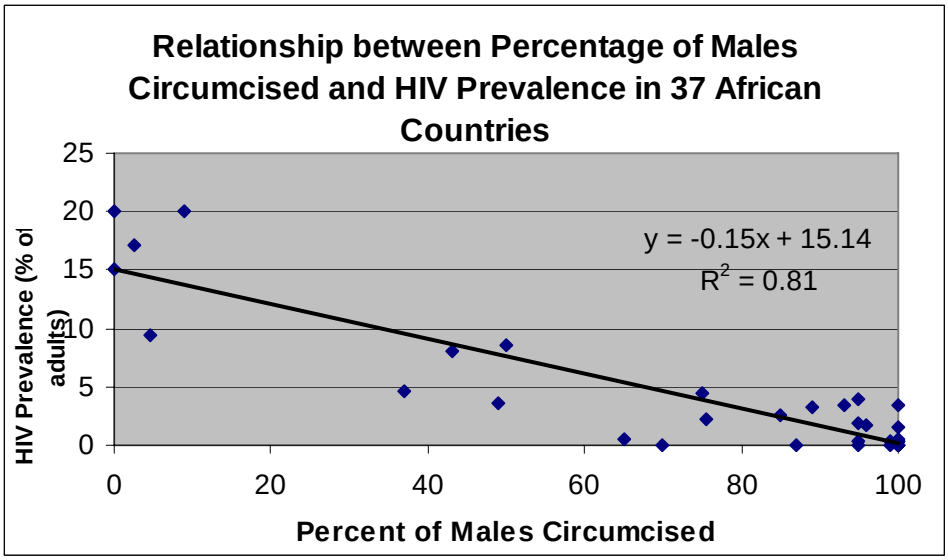
Although the histograms take more room to display and (marginally) more time to produce, they provide more detailed information about the distribution of the factors.

Neither factor is approximately normally distributed. The *Percent Circumcised* values are very left skewed, while the *HIV Seroprevalence* values are very right skewed.

The endpoints of the intervals are a bit tricky on these graphs. Recall for a histogram it is important that the interval widths are equal. It would be helpful to have more decimals in the data presented from the article. (For example, I have made the assumption that the value "0" for Rwanda is actually greater than 0.1. This would need to be confirmed.)

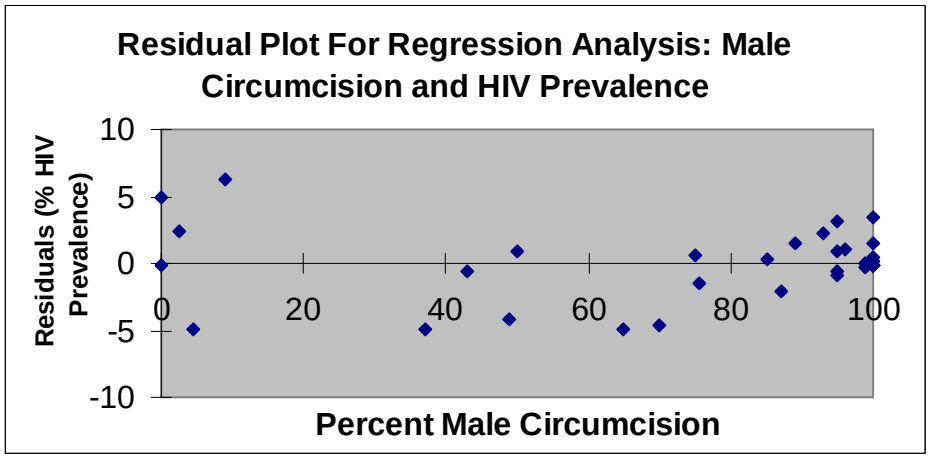
3. Produce a neat scatter plot for the given data with *x=Percent Males Circumcised* and *y = Percent HIV Seroprevalence*. Include the equation of the least squares regression line.

3.



4. a. Check the assumptions for the linear regression analysis. Produce a residual plot and comment on the normality assumption.

b. Do *Percent Males Circumcised* and *Percent HIV Seroprevalence* need to be normally distributed? Explain.



4. a.
We need to check that:
1. The relationship is approximately linear?
Based on the original plot, the data do appear to be approximately linear. Also from the residual plot, we see no "pattern" to the residuals,

such as "U" or "n" shaped.
2. The errors are approximately normally distributed with constant variance?
Based on the residual plot, the errors appear approximately normally distributed with (arguably) constant variance. The variance of the errors appears slightly bigger for *x* between 0 and 10, but not alarmingly so.
3. The sample is a SRS?
The data are not from a SRS. The population in this problem is likely "all African countries". This sample is "countries from which the data are available" – so it is unclear how we can make inference about the entire population or whether the sample is representative. (More discussion about this in question #8.)

4.b.
The *x* and *y* values are clearly NOT normally distributed. *x* is strongly left skewed and *y* is right skewed. This is not problematic for the regression analysis because the assumption of normality is for the errors, not the original data.

5. Are there any countries which, in your opinion, are outliers? [a) with respect to x, b) with respect to y, c) with respect to the overall data distribution or d) an influential point.

5.
There are no outliers in this data set with respect to *x*, *y*, the overall data distribution or influential points.

6. Find the predicted *Percent HIV Seroprevalence* when the *Percent of Circumcised Males* is 75%.

6.
The regression equation is $y = 15.15 - 0.15x$.
When $x = 75$:

$$y = 15.15 - 0.15(75)$$
$$= 3.85$$

When the *Percent of Circumcised Males* is 75% the *Predicted HIV Seroprevalence* is 3.85%.
[Use all decimals from the original regression analysis in the calculations, but results are rounded for presentation.]

7. Find the *Percent of Circumcised Males* predicted by the regression line when the *Percent HIV* is 10%.

7. The regression equation is $y = 15.15 - 0.15x$.
Substitute $y=10$ and solve for x.

$$10 = 15.15 - 0.15x$$
$$\frac{10 - 15.15}{-0.15} = x$$
$$34.2 = x$$

Approximately 34.2% of Males would be Circumcised when the *HIV Prevalence* is 10% for a country on the regression line.

8. Calculate r and r^2. Explain the interpretation of r^2. Conduct a statistical test for $\rho = 0$ and interpret the p-value.

8. r=-0.9 and r^2= 0.81.
From the coefficient of variation, r^2, we know that *Percentage of Males Circumcised* explains approximately 81% of the variation in *HIV Prevalence* for these countries.

H_0: $\rho = 0$ H_a: $\rho \neq 0$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.9\sqrt{37-2}}{\sqrt{1-0.81}} = -12.23$$

When t=-12.23 with degrees of freedom = n-2=35, the two-sided p-value is < 0.001.
If the true population correlation is zero, then the probability of obtaining a sample statistic as extreme as r=-0.9 is practically zero.
We have strong evidence against the population correlation coefficient being zero.

NOTE 1: Our sample is not a Simple Random Sample and may not be representative. The countries not included in the sample may be very different from the countries that were included.
NOTE 2: Is the p-value even needed or useful for this scenario? The population is the set of all African countries. Our data set consists of 37 out of 52 African countries.
Since this is such a large sample, relative to the population, we should question the usefulness of this significance test (even if we had a SRS of countries). For example, if we had data for all countries in Africa then there would be no reason to conduct a hypothesis test – we could calculate ρ, that actual correlation for the population (of countries), rather than an estimate, r.
What is more interesting than the significance test in this problem (how likely is the result due to chance under the null hypothesis?) is the size of the estimate (how strongly are the two factors correlated? r= -0.9… the factors are **very** strongly correlated).
Unfortunately, many authors and readers expect to see a p-value, regardless of the utility of the value.

9. Interpret the estimate for the slope. Conduct a hypothesis test for population slope equal to zero and interpret the *p*-value carefully.

Regression Statistics	
Multiple R	0.900173
R Square	0.810311
Adjusted R Square	0.804891
Standard Error	2.514634
Observations	37

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	945.4243	945.4243	149.5125	3.43E-14
Residual	35	221.3184	6.323382		
Total	36	1166.743			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	15.14251	1.021428	14.82484	1.18E-16	13.0689	17.21612
X Variable 1	-0.15046	0.012305	-12.2275	3.43E-14	-0.17544	-0.12548

9. The slope, b=-0.15, indicates a ten percent increase in *male circumcision* in a country is associated with a 1.5% decrease in *HIV prevalence*.

From the regression analysis, we can test:
H_0: $\beta = 0$ H_a: $\beta \neq 0$

We find t=-12.23 with df = 35 and p<0.001.
Assuming the slope of the population regression line is equal to zero, the probability we'd observe a sample slope as extreme as b=-0.15 is practically zero.

Again, we note that the sample is not a SRS from the population. Both the factors are measured with considerable error. And the test of significance is likely not even needed because the sample is so large relative to the population (See Note 2 in Problem 8). Of more interest is the size of the slope, rather than the probability that it is due to chance. Recall that this p-value doesn't address "is this result important?" – it addresses "if the population slope is zero, is our sample slope likely this extreme due to chance, as a result of sampling?"

10. Note that *x= Percent Males Circumcised* was estimated for each country while *y=Percent HIV Prevalence* was estimated based on the data from each capital city. The paper states, "the resulting errors produce an underestimate of the true level of correlation between the proportion of males who are circumcised and the level of HIV infection, because deviations from the regression line in a number of countries are due to discrepancies between urban and national circumcision practices." Explain.

10. If we were able to obtain HIV prevalence based on the entire country we'd expect a stronger correlation (further from 0). Our results are "biased towards the null".

11. The design of this study is an ecologic study with the unit of observation, "country". Discuss some limitations of an ecologic study design. For more information about ecologic studies see http://www.bmj.com/epidem/epid.6.html.
[Later studies have addressed this association using stronger study designs. We'll explore more about the topic in Global Activity Two.]

11.
In general, ecologic studies are considered considerably weaker than many other study designs. Often, ecologic studies are the first studies to be conducted when an association is first explored because they are relatively cheaper and easier. Some issues with the study include:
a) Data are at the 'country level,' so we can't extrapolate results to an individual level. Read more about 'ecologic fallacy' in most epidemiology textbooks.
b) There is likely considerable mismeasurement of the variables (which would bias towards the null).
c) Timing of circumcision and sexual activity is unable to determine.
d) This study does not adjust for religion, which is strongly associated with male circumcision in African countries.

BIOSTATISTICS TOPICS:
SIMPLE LINEAR REGRESSION, CORRELATION, DESCRIPTIVE STATISTICS.