summut— title: “test” author: “Nicholas Nagle” date: “January 26, 2015” output: pdf_document — # Question 1 This is a question about how the axes affects how we interpret plots.

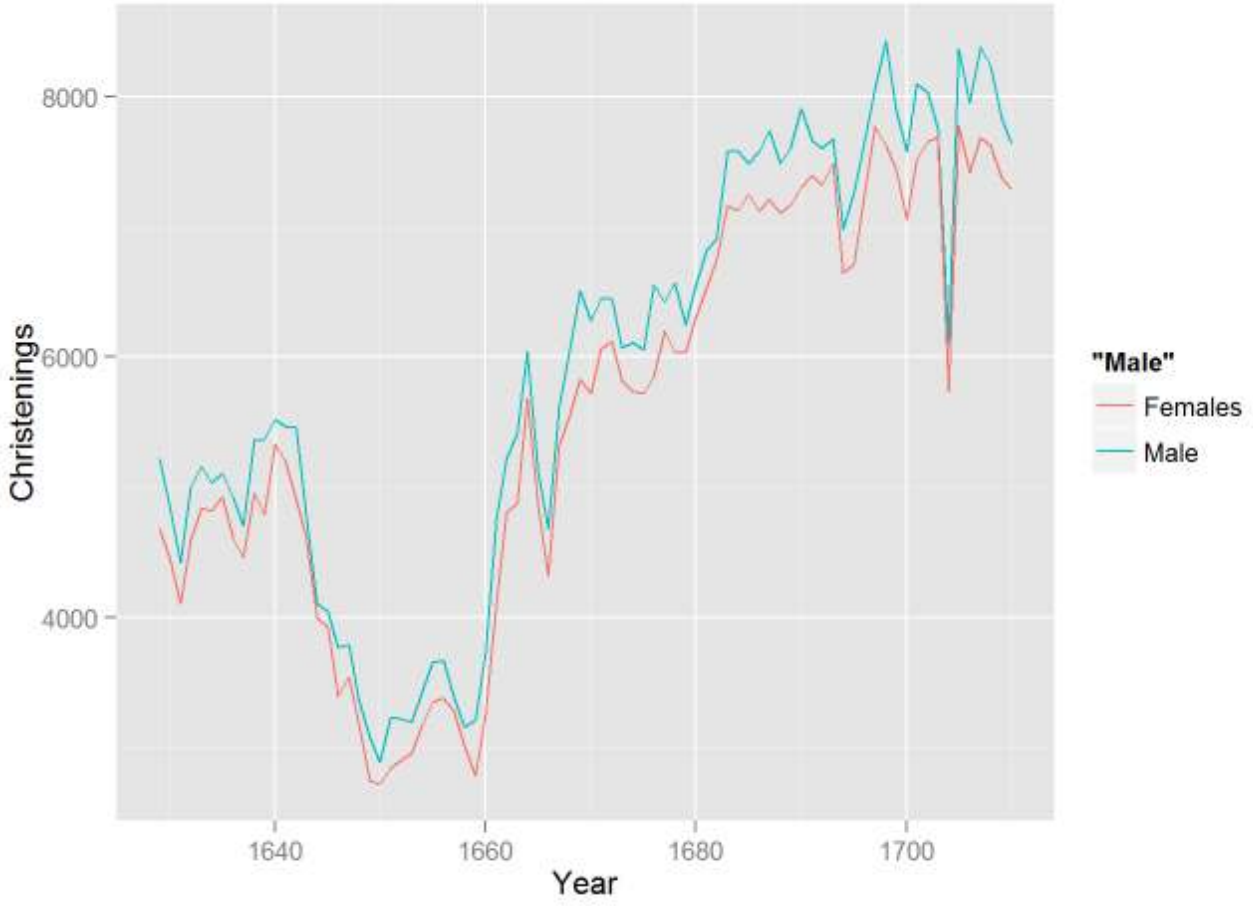
Create two plots of Male-Female Christenings, one in which the y-axis scale is set by default, and one in which the y-axis extends all the way to zero. Yes, I know that was in the tutorial. I want to see it here.

- Describe the visual appearance of the two plots. Do they “feel” like they describe the same data?
- Describe how you might be able to mislead readers by changing the scaling on graphs.
- Which plot seems more appropriate here? Why?

Answer:

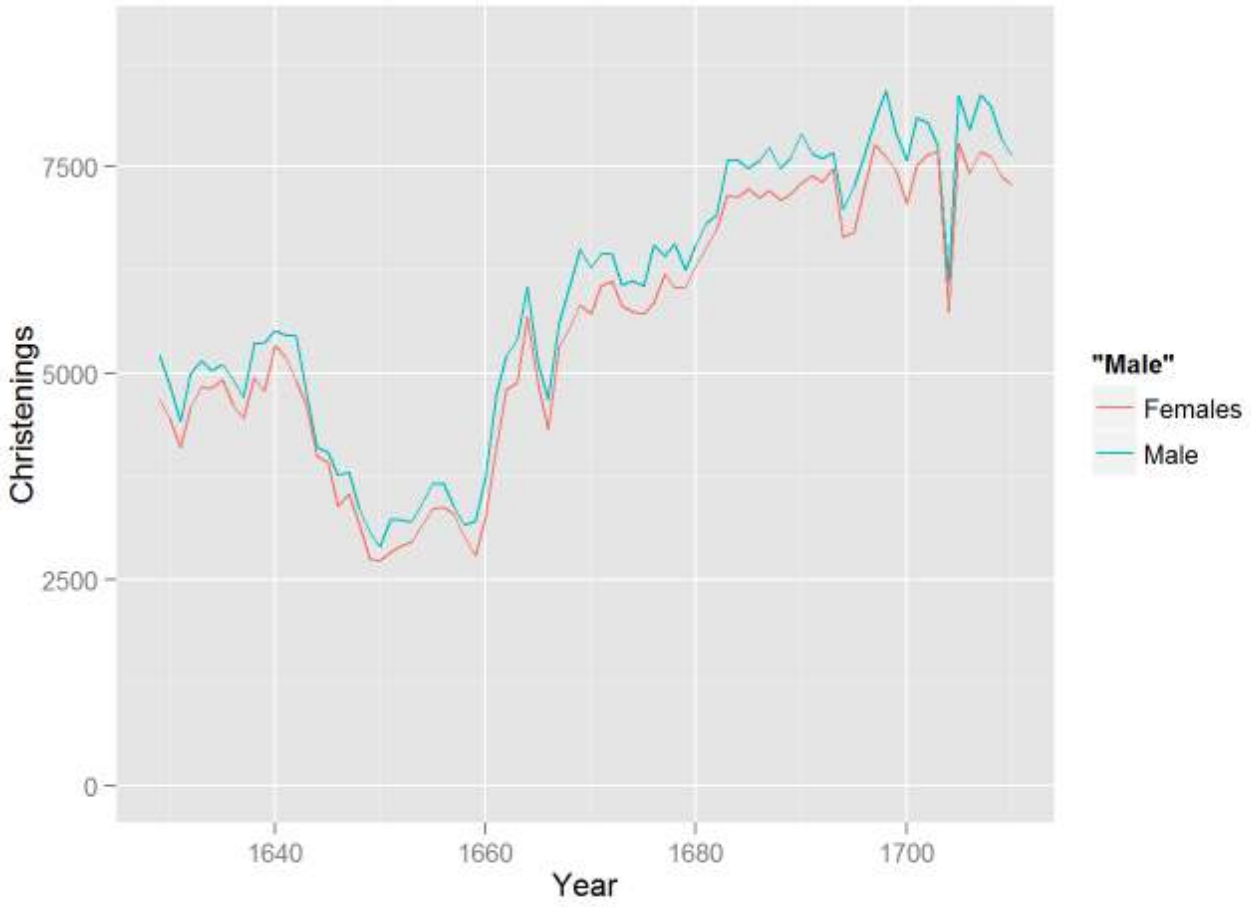
The plot of Male-Female Christenings in which the Y-axis is set by default is given below:

```
library(ggplot2)
ggplot(data=arbuthnot) +
    geom_line(aes(x=Year, y=Males, color='Male')) +
    geom_line(aes(x=Year, y=Females, color='Females')) + ylab('Christenings')
```



and the one in which the Y-axis extends all the way to zero is:

```
ggplot(data=arbuthnot)+
    geom_line(aes(x=Year, y=Males, color='Male'))+
    geom_line(aes(x=Year, y=Females, color='Females'))+
    scale_y_continuous(limits=c(0,9000))+
    ylab('Christenings')
```



The two graphs one of which has a system default Y-axis and the other one having user defines Y-axis represent the Male-Female Christenings over the time period as given in the input data. At a first glance the two plots gives the idea that they represent the same data. A closer look reveals the smoothness of the second plot. In both these plots, the X-axis scaling remains the same, thus the change of plot dimension will affect interpretation of these two graphs. In the first graph, th esudden drop around year 1650-1660 and 1705 looks more severe compared to the second graph. Still, the second graph still has visual similarity to the first one, as the Y-axis scaling has not been changed drastically.

In his book “How to Lie With Statistics” (1954), author Darrell Huff pointed out how misrepresentation of statistical data can lead to false interpretation. Line graphs such as the ones in question here may be used to mislead readers if the dimension is changed drastically. For example, the plots convey that there is a gradual increase in chirstenings over the time period inspite of some undulations. If the Y-axis range is increased to say 0-20,000 range, we will only gat a feel that there has been a gradual increase but not the sudden drops. Or if the X-axis range changed to 1200-2010, then the gradual increase will suddenly seem like a rapid increase.

To me, the second plot looks more appropriate as it still retains the important information as the first one while doing away with the exaggerated undulations of the first plot.

Question 2

This question is designed to give you some practice with ggplot as well as describing plots in words.

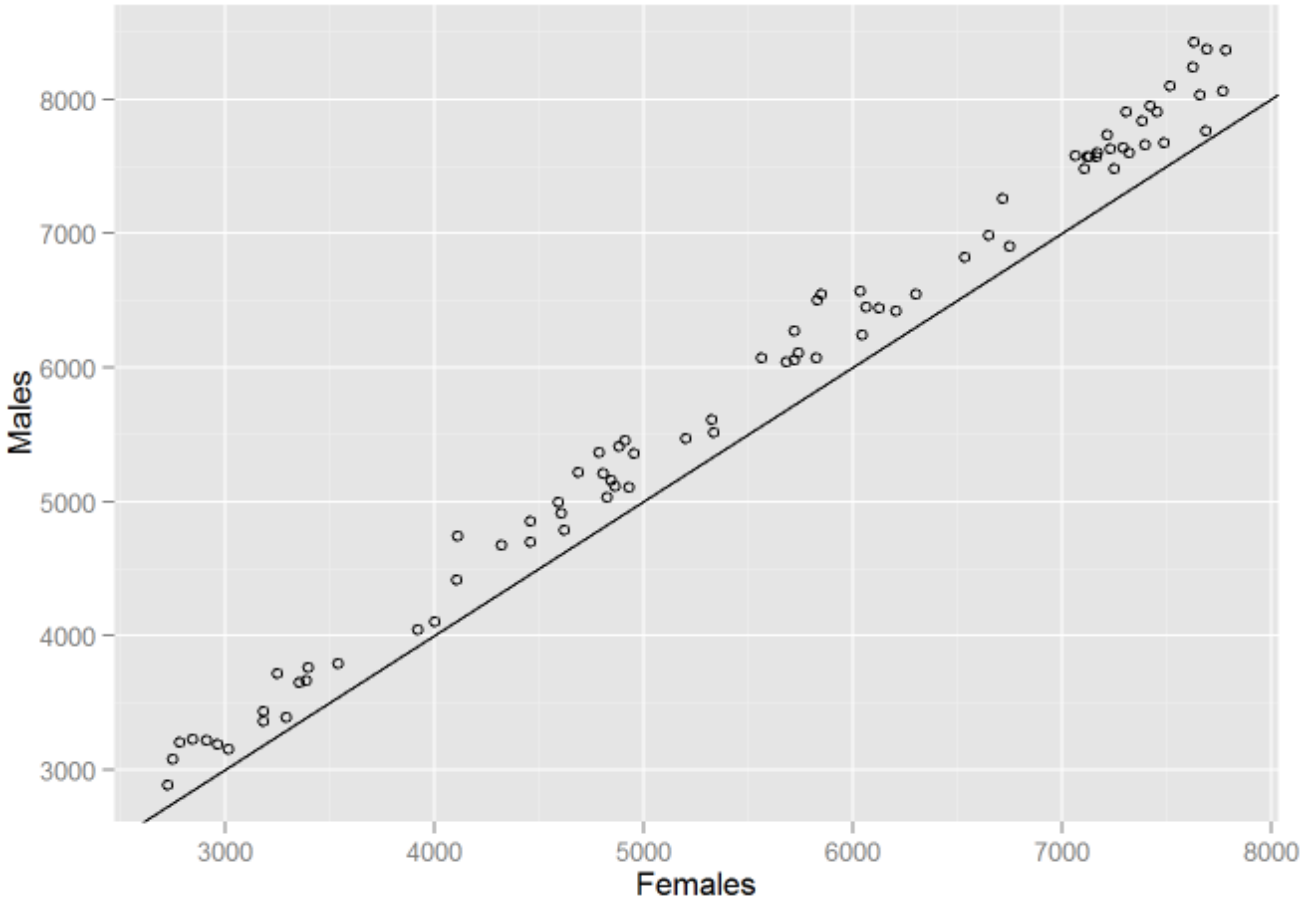
Use ggplot to create a scatterplot that has Female christenings on the x-axis and Male christenings on the y-axis. Draw a 45 degree line (using geom_abline) to show the line where Male and Female christenings are equal. Use this figure to describe the relationship and distribution of Male and Female christenings.

Answer:

The scatterplot showint Female Christenings on the X-axis and the Male Christenings on the Y-axis along with a 45

degree line showing where Male and Female christenings are equal is given below:

```
ggplot(data=arbuthnot, aes(x=Females, y=Males))+
    geom_point(shape=1)+
    geom_abline(intercept=1, slope=1)
```



From the scatterplot, we can see that the distribution is skewed towards the Males. This could be interpreted from the fact the data points are on the top of the 1:1 line (45 degree line), the zone which is skewed towards the male numbers. Similarly, if the points were below th eline, then th edistribution would have been skewed towards Females and if they were on the line, then the distribution would have been equal.

Question 3

This is a question about population, sample, representativeness and generalizability.

How do you think the christenings-based sample would compared to a births-based sample? Similar? Different? Why? Arbuthnot's data probably included most every christening in London during this period; they probably aren't any unreported christenings. Is this fact important? Why or why not? Would a christenings based sample be appropriate now, in the 21st century? (Hint, this last question is trickier than it might seem. Think about what causes Male/Female Births, what causes people to christen their children, and any relations or not between these)

Answer

Given the fact the Arbuthnot conducted his study in 17th Century London, his sampling could have represented a rational representation of the newborns. Having said that, there could have been people who did not follow the christening ceremonies hence their babies would not have been covered by this study. Thus, a birth based sampling would have been better if one was to capture information about all th enewborn across different sections of the society.

The fact that there has been no unreported christenings during the period is important if the study was meant to cover only christened newborns, but as the study ws supposed to cover all newborns this fact does not hold much value.

In the 21st Century, the birth based sampling would be more appropriate. As the population becoming more heterogenious in terms of beliefs and culture, christening based sampling will capture information of a subset of total population and the results from such studies may not reflect the ground realities.

Question 4

What does “sex ratio at birth” mean? How does it depend on biology, cuture and technology. Consider the three cases of 1) late 17th century England, 2) 21st century US, and China under the one-child policy (http://en.wikipedia.org /wiki/One-child_policy).

Answer:

Sex ratio at Birth may be defined as the number of boys born alive per 100 girls born alive (http://stats.oecd.org /glossary/detail.asp?ID=2447 (http://stats.oecd.org/glossary/detail.asp?ID=2447))

Factors affecting human sex ratio at birth have been an actively pursuaded area by researchers over time. extensive studies suggest that sex ration may vary widely depending on varios social, economic factors such as paternal age, maternal age, birth order, parents health history and psychological stress.

Recent studies have found this ratio to generally be 1.02 to 1.08, meaning an excess of male babies compared to female babies being born alive, (James, 2008). It should also be npted that there has been observable variations from this range due to various reasons.

Biological factors such as stress during gestation informs of malnutrition in mother may increase fetal deaths especially of males, thereby lowring the ratio. Whereas Hepatitis B infections have been believed to increase the male to female ratio.

Social factors such as gender selective abotions and infanticide result in skewed male to female retio especially in China and India, where the ration is higher than the mean ration recorded in United States.Often religious and cultural beliefs affect number of children per couple, since birth order plays a role in sex ratio these cultural factors also play a role in sex ratio determination. Change in demographics may also lead to variation in sex ratio as seen by some researchers in California. Culture also determines the general age of marriage. Though researchers have not found any relation between maternal age and sex ratio but have noted that parental age affetcs the same significantly. Thus early marriage in some cultures may have played a role in sex ratio.(en.wikipedia.org/wiki/Human_sex_ratio)

Technological factors such as prenatal sex determination combined with a general preference for boys have significantly changed sex ratio in some countries such as India as highlighted by Madan and Breuning, 2013.(http://www.nature.com/gim/journal/v16/n6/full/gim2013172a.html (http://www.nature.com/gim/journal/v16/n6 /full/gim2013172a.html))

Late 17th Century London sex ratio shows an excess of male babies compared to females (shown in the scatterplot), which may have been resulted from birth orders and cultural influences. A study by Mathews et al, 2005 showed significant transitions in sex ratio from 1940-2002. The affecting factors were identified as maternal age, birth order and race (http://www.cdc.gov/nchs/data/nvsr/nvsr53/nvsr53_20.pdf (http://www.cdc.gov/nchs/data/nvsr/nvsr53 /nvsr53_20.pdf)).

The one chile policy adopted by China was adopted to control the ever increasing population and has been found to be effective in achieving that goal. But have also shown its effect in a skewed sex ratio. Sex ratio reached significantly higher levels and some studies predict a dempgraphy with marked excess of men which may lead to social problems in future. In case th efirst child is a girl and the couple get to conceive a second baby, they have been seen to be taking measures to concieve a boy. Similarly couple with two or more boys have been shown to have a tendancy to conceive girls.

Question 5

The purpose of this quetion is to give you a little practice using standard deviation.

Answer:

The average height of young women ages 18-24 is 64.5 in. The distribution of heights is approximately normal (Gaussian) with a standardard deviation of $\sigma = 2.5$. Complete this sentence: Approximately 95% of women have a height greater than 59.5 in and less than 69.5 in.

Question 6

The purpose of this question is to help you understand the variance.

The formula for sample variance σ^2 of a dataset is:

$$\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

Describe in words what each part of this equation is, and using the concept of “distance” describe what the sample variance measures.

Answer:

$$x_i$$

: This describes individual observations in the sample datasets while i could take any value from 1 to N where n is the number of individual observations.

$$\bar{x}$$

: Represents the mean of all the individual observations.

$$\sum_{i=1}^{N}(x_i - \bar{x})^2$$

: Represents the differences between each observation and mean squared and then summed for all the observations. This is the numerator in the equation. N represents

$$N-1$$

: N is number of individual samples. The value (N-1) is used as a denominator in the equation

The variance signifies how the individual observations are distributed around their mean. If the variance is high then the samples are spread far apart from the mean and from each other while low value signifies a close knit distribution. Thus, the distance from the mean line is higher is case of a high variance while low variance means the distance from the mean line is also low. In case of a variance 0 means all the observations are identical.