Alex Smith
February 7, 2014
W203: Exploring & Analyzing Data

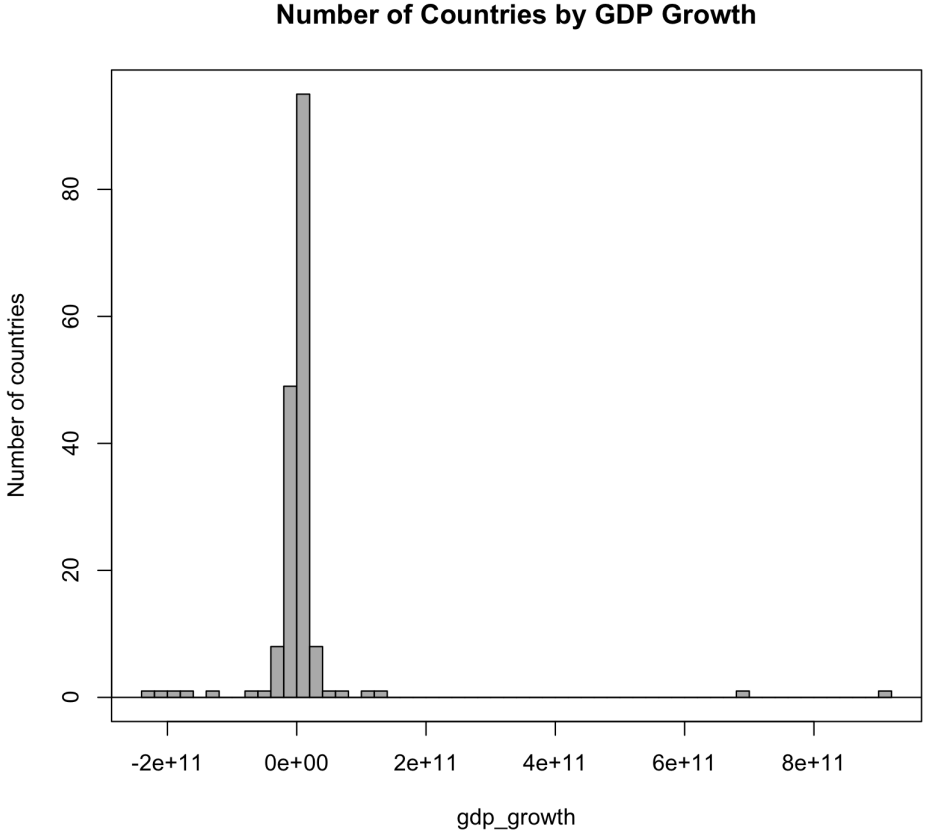
LAB 1

PART I.

1. e) ordinal
2. a) measuring a potentially interval or ordinal variable as a binary variable
3. b) The chance that a single draw from a population falls within one standard deviation of the mean is always the same for any population.
4. b) stratified random sampling
5. d) It depends on the standard deviation of the population.
6. b) For large samples, it suggests that the normal distribution is a good model for the distribution of the mean and other statistics.
7. c) A smaller variance of the sampling distribution of the mean.
8. d) The distribution of your age variable is platykurtic.
9. b) $.01 \leq P(HIA1) < .02$

PART IIA.

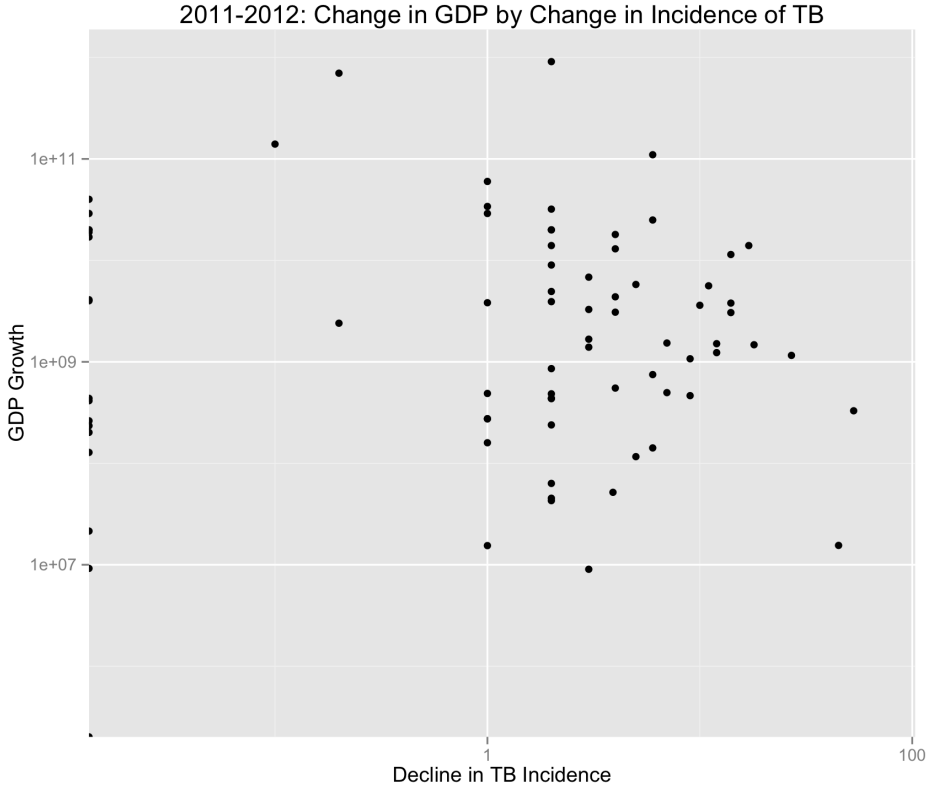
10. Mean of gdp_growth: 7172376796
11. The distribution of gdp_growth is **unimodal** with positive kurtosis (**leptokurtic**) and a **positive skew**.



12. 39 countries have above average growth and 40 countries have below average growth. The mean splits the peak of the histogram above. About half of the countries are to the right of the mean and about half are to the left. There are a few countries that showed very large growth that are pulling the mean up. There are multiple countries that showed modest declines in growth. These countries center the mean in the distribution of country growths.

PART IIB.

13.



We first plotted GDP growth (from 2011 to 2012) against the decline in tuberculosis incidence (from 2011 to 2012). We chose to plot these variables against each other because there has been considerable research on how public health impacts economies [1]. We could not learn much from the graph because the data points of both variables was clustered around 0. We took the log of both variables (as demonstrated in the Async material, 5.5). We understand that by taking the log, we are losing potentially valuable negative data points. However, taking the log of both variables gives us a different picture of the data. We can see that, to the extent that there is a trend, those countries with lower declines in TB incidence saw greater economic growth. One possible explanation is that countries with stronger economies have lower levels of TB incidence. These countries may not have lowered their already low incidence of TB but may have continued to grow fastest. The data definitely sparks interest in further research. For further research, we may also want to consider using the proportional growth of GDP and decline of TB rather than the nominal change.

[1] Ashraf, Quamrul, Ashley Lester, and David Weil. “When Does Improving Health Raise GDP?” (University of Chicago Press). *NBER Macroeconomics Annual 2008*, Volume 23. http://www.nber.org/chapters/c7278.pdf

Alex Smith
February 7, 2014
W203: Exploring & Analyzing Data

APPENDIX.

```
# import the data and set it to a data frame
WB_gdp <- read.csv("GDP_World_Bank.csv",header = TRUE)

# add a column for gdp growth between 2012 and 2011
WB_gdp$gdp_growth <- WB_gdp$gdp2012 - WB_gdp$gdp2011

# find the mean growth, while avoiding "NA" values
mean_growth <- mean(WB_gdp$gdp_growth, na.rm = TRUE)
mean_growth

# plot a histogram of the growths
library("Rcmdr")
with(WB_gdp, Hist(gdp_growth, scale="frequency", breaks=50, col="darkgray", xlab="gdp_growth", ylab="Number of
  countries", main="Number of Countries by GDP Growth"))

# create a high_growth column that yields true for countries with above average growth and false
# for countries with below average growth
WB_gdp$high_growth = WB_gdp$gdp_growth > mean_growth

# find the number of countries with above and below average growths
num_true <- length(WB_gdp$gdp_growth[WB_gdp$gdp_growth==TRUE])
num_true
num_false <- length(WB_gdp$gdp_growth[WB_gdp$gdp_growth==FALSE])
num_false

# load TB incidence rates, source: World Bank (http://data.worldbank.org/indicator/SH.TBS.INCD)
# IMPORTANT: must delete frist two rows of World Bank data so that R can properly read in headers,
# change header for "Country Name" to "Country" to merge with existing data frame
WB_TB <- read.csv("TB_data.csv", header = TRUE)

# create new data frame for both GDP and TB data
WB_gdp_tb = merge(WB_gdp,WB_TB, by="Country", all=TRUE)

# create new variable, TB_decline, that measures how TB incidence dropped from 2011 to 2012
WB_gdp_tb$TB_decline <- WB_gdp_tb$X2011 - WB_gdp_tb$X2012

# plot GDP growth against TB incidence, make sure you have GGPlot package installed
library(ggplot2)
graph_gdp_tb = ggplot(WB_gdp_tb, aes(TB_decline,gdp_growth)) + ggtitle("2011-2012: Change in GDP by Change
  in Incidence of TB")
graph_gdp_tb + geom_point()

# because we find that points are largely clustered around the zero values on the x- and y-axises, we choose to scale
# both variables by log10, recognizing that we're losing those values that are negative
# (source: Async lecture 5.5)
graph_gdp_tb + geom_point() + scale_y_log10() + scale_x_log10() + labs(x="Decline in TB Incidence", y="GDP
  Growth")
```