Problem set 1

Erik Parker

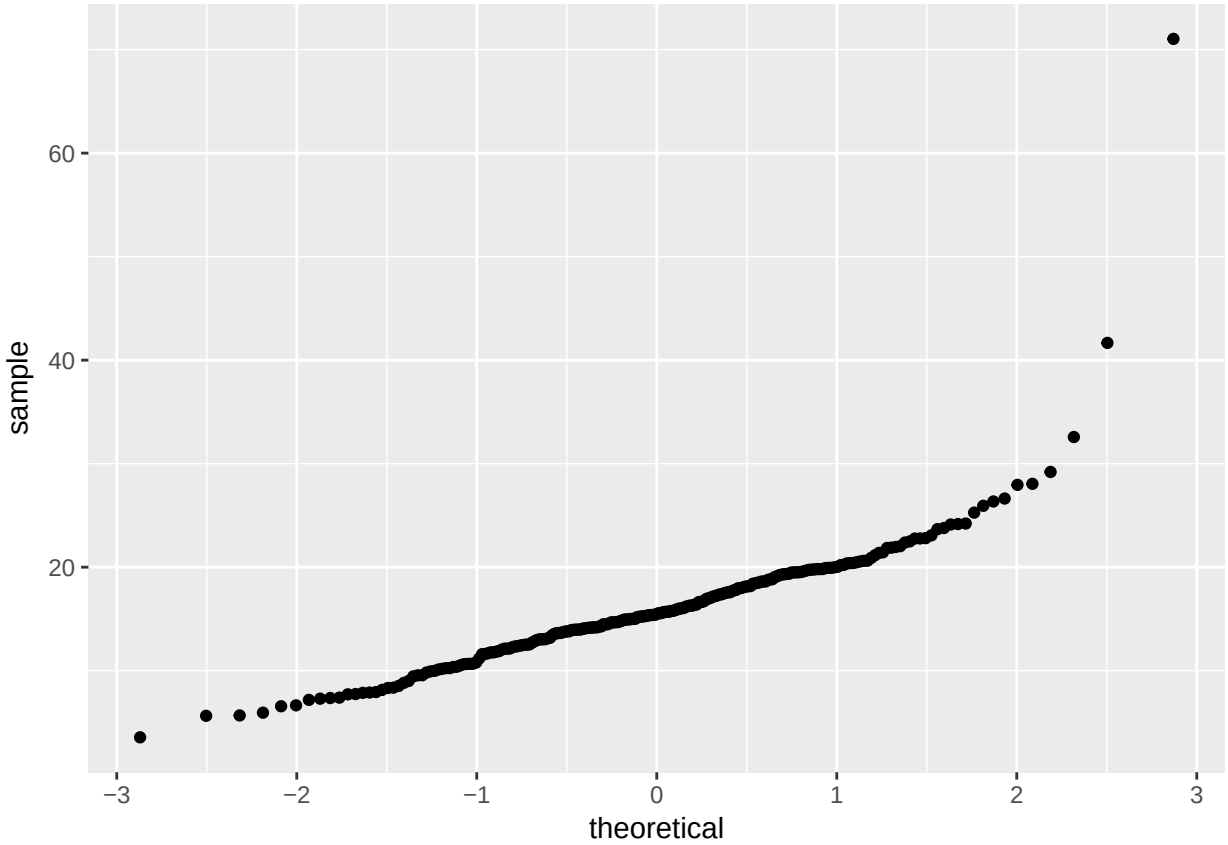
January 17, 2017

```
tips <- read.table("/media/removable/USB Drive/S670/tips.txt", header = TRUE)
tip.percent <- (tips$tip / tips$total_bill)*100
require(ggplot2)
```

```
## Loading required package: ggplot2
```

Question 1

```
ggplot(tips, aes(sample = (tips$tip/tips$total_bill)*100)) + stat_qq()
```

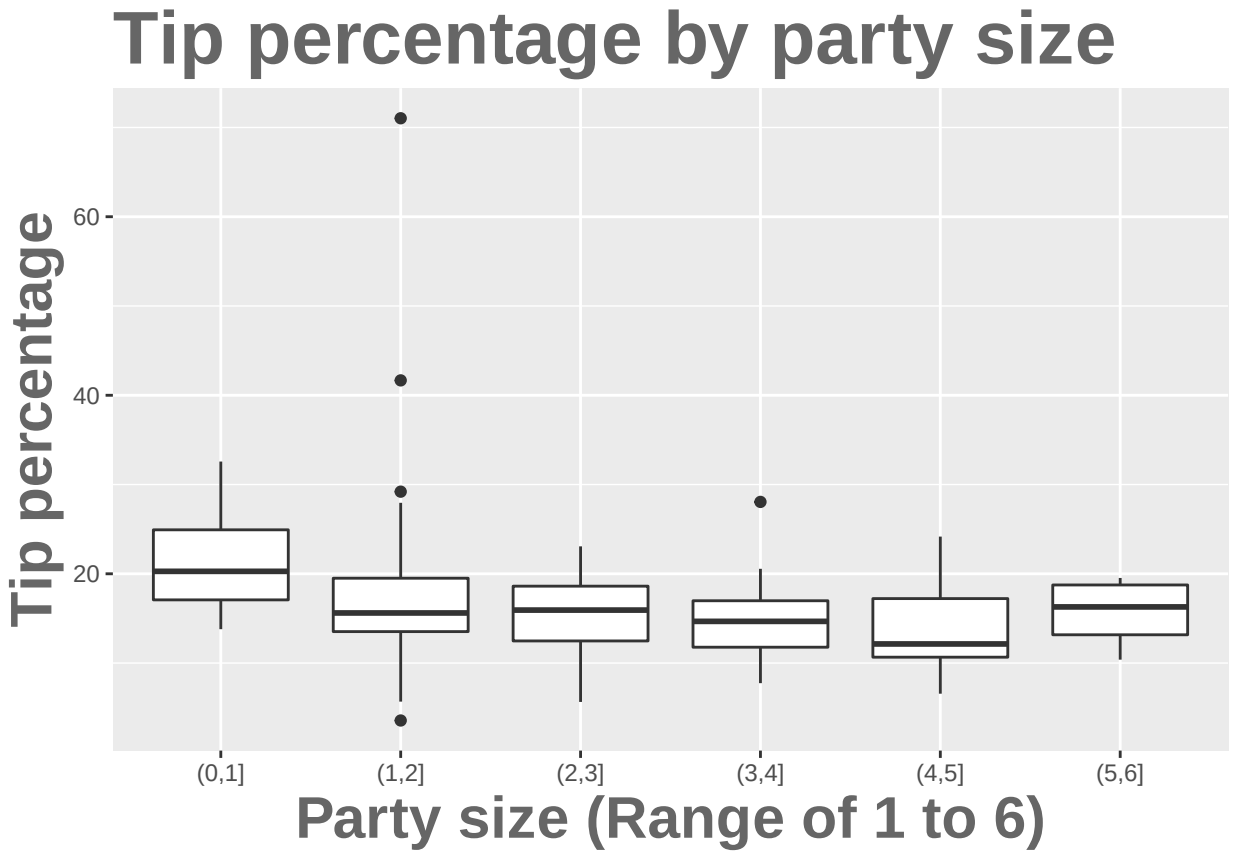


Percentage tipped seems to follow a pretty normal distribution overall. The clear problem here are the two outlier datapoints (big tippers!) Aside from these points though, the line returned by the qqplot is quite straight.

Question 2

```
tips$sizebin <- cut(tips$size, seq(0,6,1))

ggplot(tips, aes(x = tips$sizebin, y=(tips$tip/tips$total_bill)*100)) + geom_boxplot() + labs(x="Party s
theme(axis.title = element_text(color="#666666", face="bold", size=22))
```



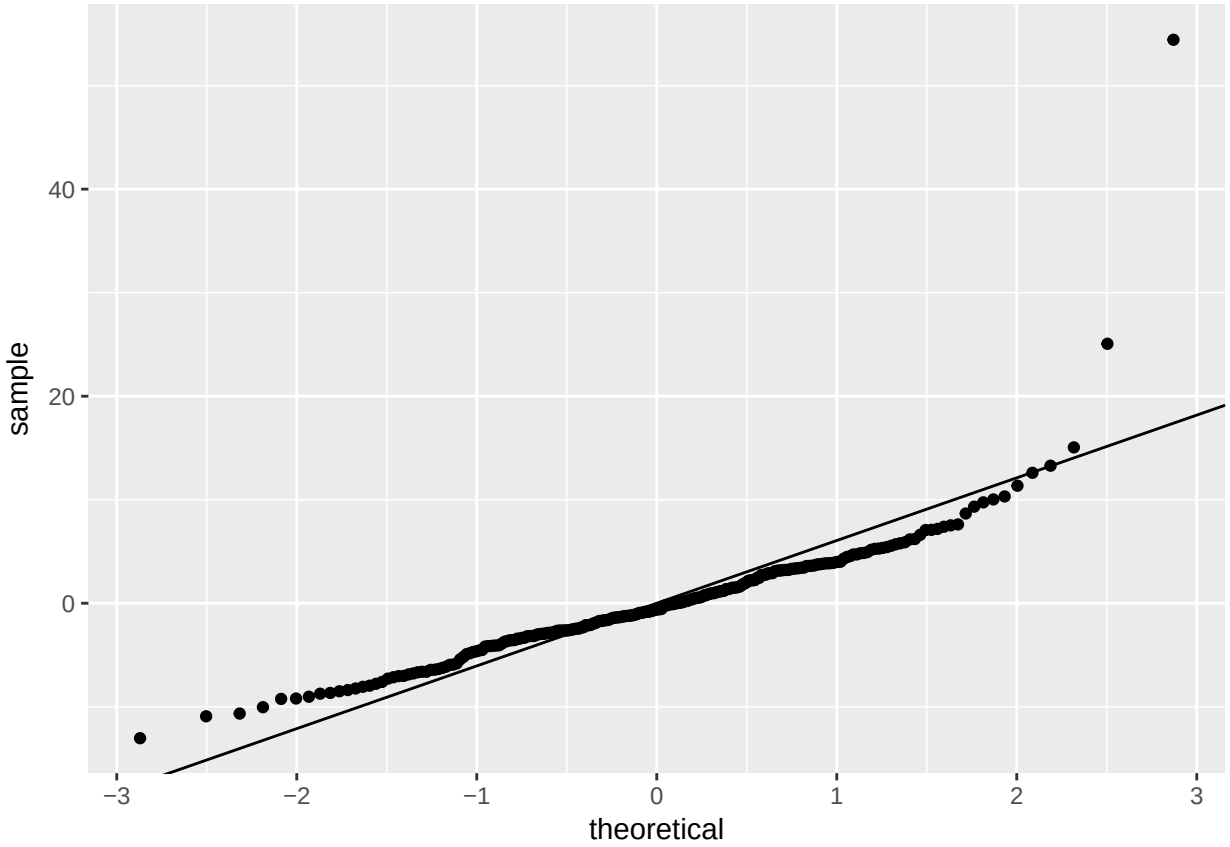
The distribution of tip percentage seems to be highest, on average, for single diners (around 20%) and lowest for parties of 5 (around 12%), again on average. Interestingly, groups of two tip both the highest percentage, and the lowest out of all the group sizes yet still manage to have a very "average" mean amount tipped despite the presence of a few outlying points. Overall though, the average tip percentage for all groups, regardless of size seems to fall within the very American 10-20% range, with four of the six groups coming in between around 15 and 18%.

Question 3

```
tips.lm <- lm((tips$tip/tips$total_bill)*100 ~ tips$size, data = tips)
tips.lm
```

```
##
## Call:
## lm(formula = (tips$tip/tips$total_bill) * 100 ~ tips$size, data = tips)
##
## Coefficients:
## (Intercept)    tips$size
##     18.4375      -0.9173
```

```r
tips.res <- data.frame(party.size=tips$sizebin, residual=residuals(tips.lm))
#ggplot(tips.res, aes(sample = residual)) + stat_qq() + facet_wrap(~party.size, ncol = 2)
#normalqq plot of residuals of tip percentage broken up by party size
#ggplot(tips.res, aes(x= residual)) + stat_ecdf()
# supposed to be S shaped if residuals normal, not looking great...
#ggplot(tips.res, aes(sample = residual)) +stat_qq()
#round(mean(tips.res$residual), 3)
#round(sd(tips.res$residual), 3)
ggplot(tips.res, aes(sample = residual)) +stat_qq() + geom_abline(intercept = 0, slope = summary(tips.lm
```



This final qqplot showing the residuals overlayed with a line of y-int = mean of the residuals and slope = sd of the residuals shows that the residuals are relatively normal, though there is one (two?) problem: the extreme outlier in the top right corner. It seems as though the presence of this outlier is doing a lot to skew the data, and bend the slope of the line plotted towards it. The linear model developed for this data would do a better job at predicting percent tipped from party size if this point was omitted and the analysis run again, as I believe that would lead to a much straighter qqplot, and so more normal residuals better in line with the assumption of linear regression of normally distributed residuals.