Assignment 3 Stat 139

Callin Switzer

September 21, 2014

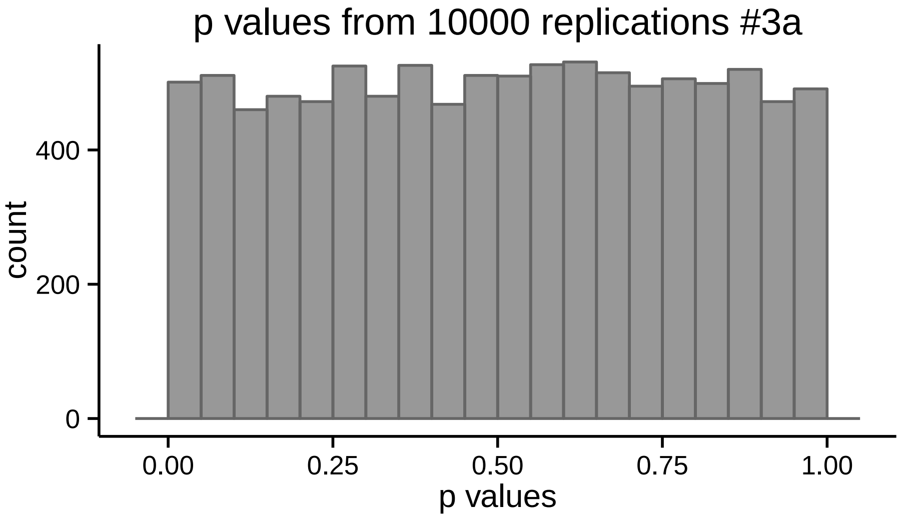
3.

From the test, 0.7207 is the p-value

	assumptions	met
1	Independence between units	yes
2	* Independent within populations	yes
3	* Independent between populations	yes
4	Homogeneity of units within each popoulation	yes
5	* equal means and variances within each population	yes
6	* equal variances between populations	yes
7	Populations are normally distributed	yes
8	Random sampling from populations	yes

Table 1: I. Which assumptions of the specified test are met and which are not, if any

3aI. All the assumptions are met. 3aII. I don't expect to see deviations from the actual type I error rate, because this test meets all the assumputions, and has a large sample size. See Table 1.



3aIII

3aIV. H0 was rejected 501 times out of 10000 – which corresponds to 0.0501 percent of the time. Though it is a little different from 0.05, I think the alpha level is still at the nominal level.

3b

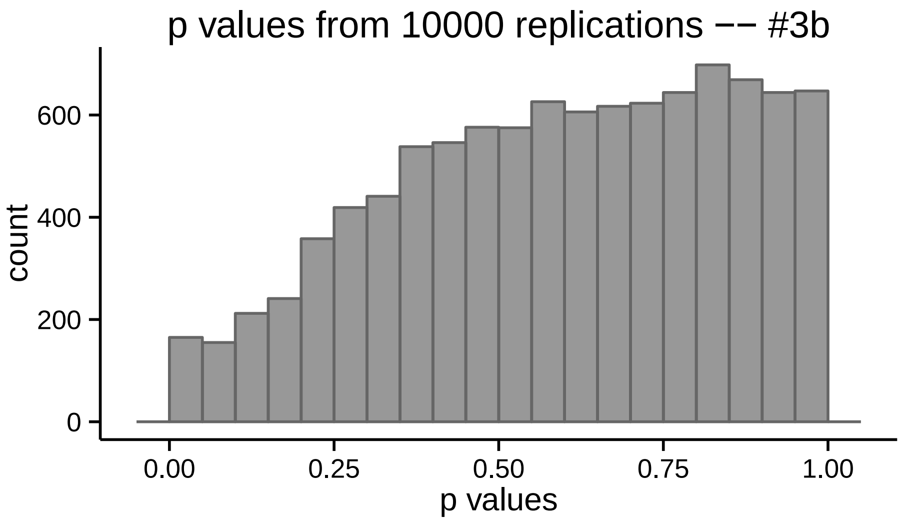
0.6749 is the p-value

3bI. There are equal means and variances within each population, but there are not equal variances between the populations. Thus the Homogeneity assumption is not met. Also, the populations are not normally distributed, and one sample size is small – this means that the central limit theorem will not come into play. See Table 2.

3bII. I expect to see deviations from the nominal Type-I error rate, because Dr. Liublinsak showed us a similar example in class, with small sample sizes and skewed distributions. I think the number of p-values below 0.05 will be lower than nominal, because the sample size of one group is so small.

	assumptions	met
1	Independence between units	yes
2	* Independent within populations	yes
3	* Independent between populations	yes
4	Homogeneity of units within each popoulation	no
5	* equal means and variances within each population	yes
6	* equal variances between populations	no
7	Populations are normally distributed	no
8	Random sampling from populations	yes

Table 2: I. Which assumptions of the specified test are met and which are not, if any



3bIII.

3bIV. H0 was rejected 165 times out of 10000 – which corresponds to 0.0165 percent of the time. This is much lower than the nominal level (deflated).

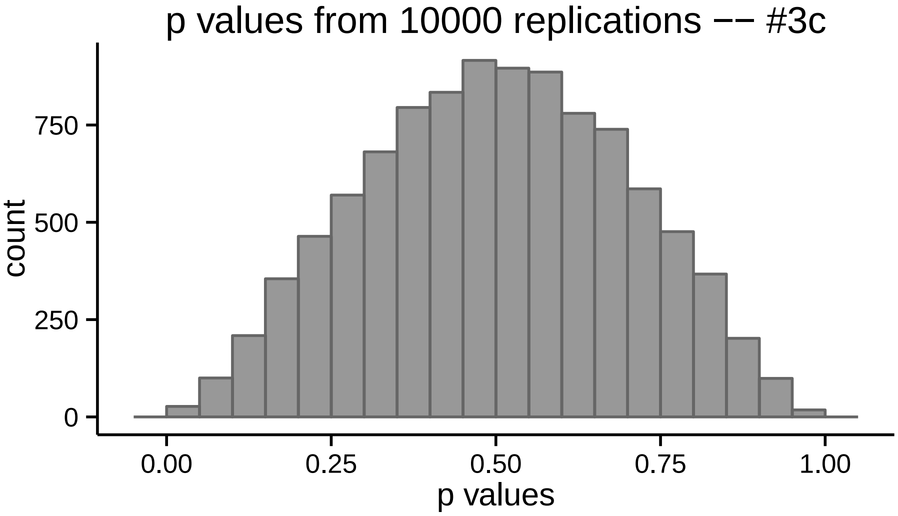
3c

	assumptions	met
1	Independence between units	no
2	* Independent within populations	yes
3	* Independent between populations	no
4	Homogeneity of units within each popoulation	no
5	* equal means and variances within each population	yes
6	* equal variances between populations	no
7	Populations are normally distributed	no
8	Random sampling from populations	yes

Table 3: I. Which assumptions of the specified test are met and which are not, if any

3c. The p-value is 0.5257. 3cI. The assumptions that are violated are independence between populations, and equal variances between populations. I expect these to influence the Type-I error rate. See Table 3.

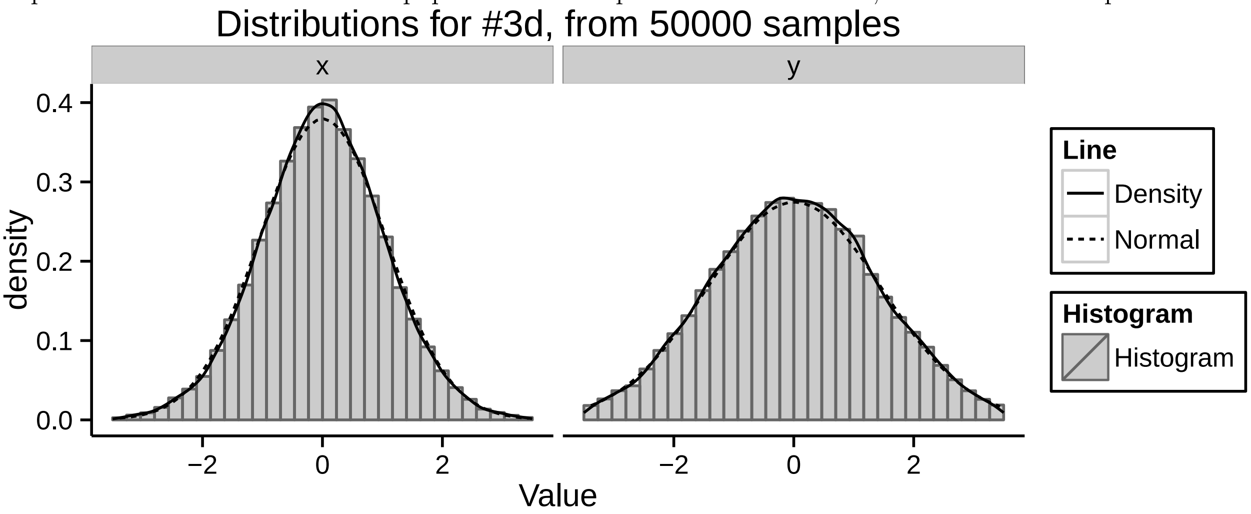
3cII. I expect to see huge deviations in the actual Type-I error from nominal levels for several reasons. First, the samples are dependent – if I know y1, then I also know something about x1. Second, the variances aren't equal. These distributions probably don't differ from normality enough that the CLT won't come into play (we don't have to worry about non-normal distributions) – see distributions below.



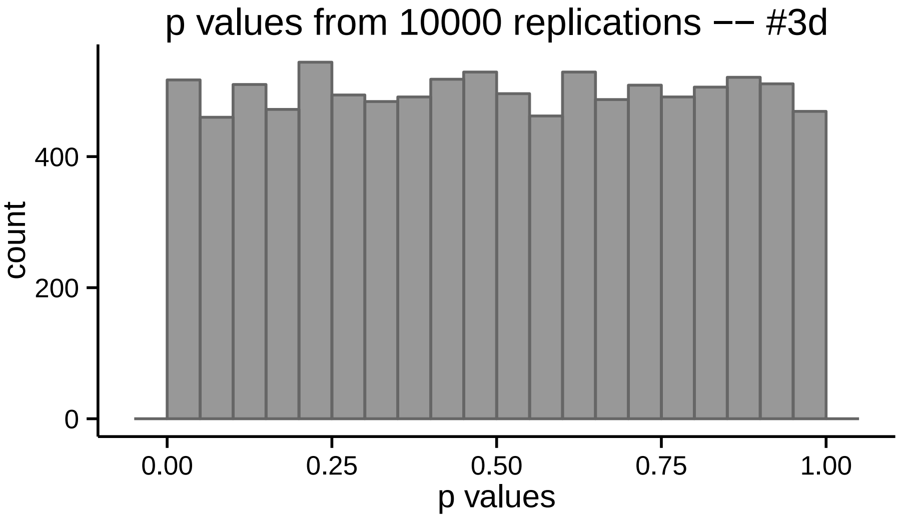
3cIII

3cIV. H0 was rejected 27 times out of 10000 – which corresponds to 0.0027 percent of the time. This is much lower than the nominal level (deflated).

3d. I first decided to look at the distributions of the two populations, to make sure that the variances were not equal. They were not, and the populations are non-normally distributed. I will use a t-test with unpooled variances. Since the two populations are dependent on each other, I decided to run a paired t-test.



Here are the results of the unpooled variance, paired t-test.



With the paired, unpooled variance ttest, H0 was rejected 517 times out of 10000 – which corresponds to 0.0517 percent of the time. This is type I error rate greatly improved.

4.

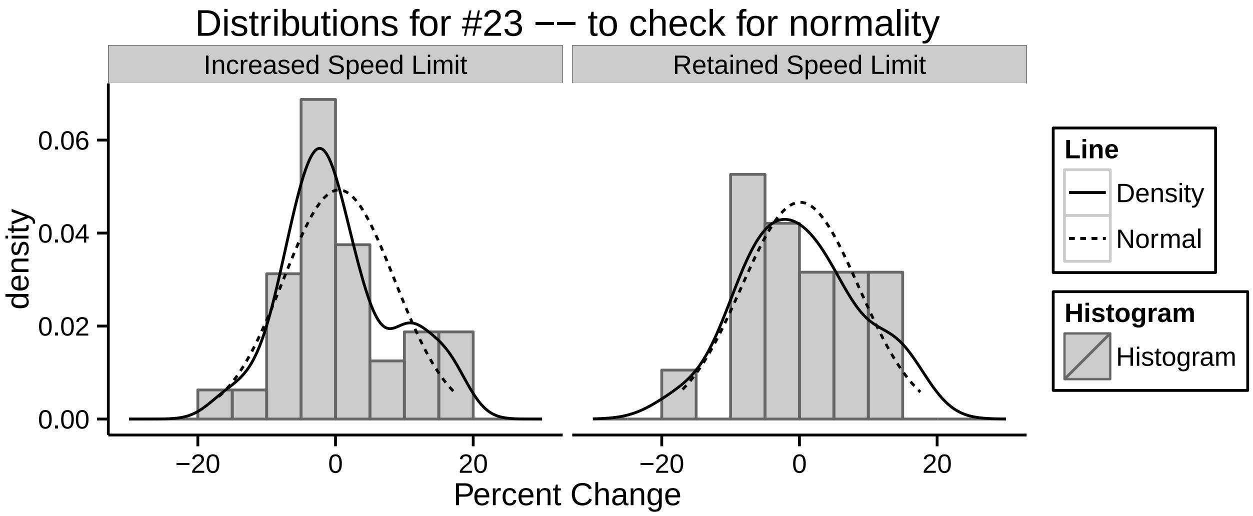
The evidence that there percentage change was greater in states that increased their speed limits is the fact that the mean of the PctIncrease in traffic accidents was higher in the group of states that increased their speed limits than in the group of states that retained their speed limit. However, using a t-test with pooled variances, we fail to reject the H0 that there is no difference between the means of the two groups.

The ratio of variances is 0.8949, which is between 0.5, and 2. Going on this, I used a t-test with pooled variance.The p-value is 0.4349. The difference of the (mean of states that increased speed limit) - 0.3943%.

4a. Normality: I've graphed histograms for both distributions below. I've shown a density line and a normal curve plotted over each histogram. If they were perfectly normal, then the density lines and normal lines would match up. However, they're not quite normal. They are both sligthly skewed, though the "Increased Speed Limits" states have a less normal distribution than the other states.

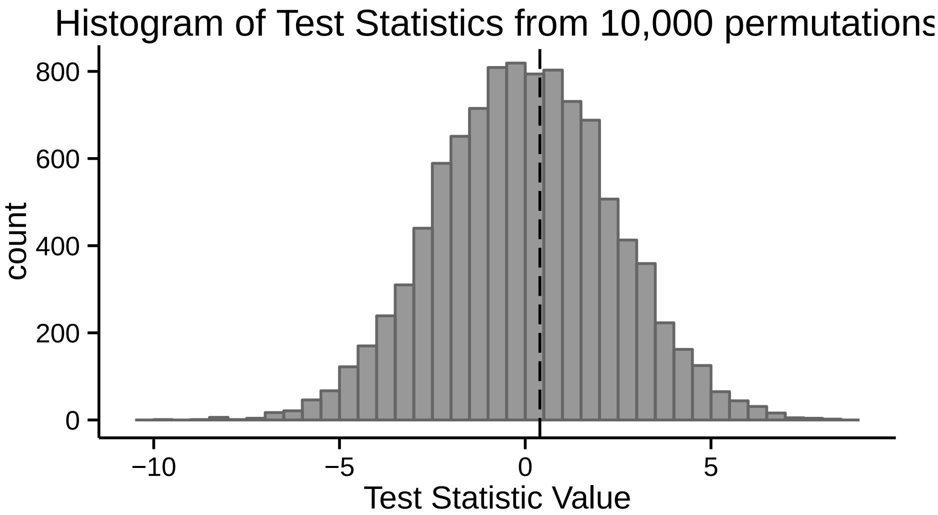
Independence: I don't think the states could be considered completely independent. Geographic location can determine states' politics, and therefore their likelihood of increasing their speed limit. If we know the location, we already have information about that state.

```
##
## Increased Speed Limit  Retained Speed Limit
##                    32                    19
```



4b. 1. This was an observational study. 2. We cannot make causal inferences, based on these data. We can call any of our findings "associations", but not causes. 3. As mentioned above, the location and political party of the stat could be confounding variables that are related to both the predictor and the responding variable.

4c. Here is a histogram of the possible test statistics. The dotted line is our observed test statistic.



Hypotheses and assumptions:

H0: *the outcome is not related to the group status – increased or retained speed limit*

HA: *the outcome is associated with group status*

Assumptions: Independence of study units

Conclusions: The probability of getting a test statistic >= the actual calculated test statistic of 0.3943 is 0.4343. This means we fail to reject our H0, and conclude that the outcome (pct increase in accidents) is not associated with group status. It is consistent with the t-test above.

Code

```
##################
## Setup
##################
options(xtable.comment = FALSE)
require(xtable)
require(ggplot2)
require(plyr)
set.seed(12345)
pp <- t.test(rnorm(50), rnorm(50), alternative = "greater", var.equal = T)$p.value

##################
## 3aI
##################

assumptions <- c("Independence between units",
                 "*     Independent within populations",
                 "*     Independent between populations",
                 "Homogeneity of units within each popoulation",
                 "*     equal means and variances within each population",
                 "*     equal variances between populations",
                 "Populations are normally distributed",
                 "Random sampling from populations")

#assumptions
met <- c(rep("yes", 8))

print(xtable(data.frame(assumptions, met), caption = "I. Which assumptions of the specified test are me


##################
## 3aIII
##################

nsim <- 10000
foo <- data.frame(rep = replicate(nsim, t.test(rnorm(50), rnorm(50), alternative = "greater", var.equal
ggplot(foo) +
     geom_histogram(aes(x = rep), col = "grey40", alpha = 0.5, binwidth = 0.05)+
     theme_bw() +
     theme(
          plot.background = element_blank()
          ,panel.grid.major = element_blank()
          ,panel.grid.minor = element_blank()
          ,panel.border = element_blank()
           ) +

     #draws x and y axis line
     theme(axis.line = element_line(color = 'black')) +
     labs(x = "p values", title = "p values from 10000 replications #3a")

alpha <- length(foo$rep[foo$rep <= 0.05])/length(foo$rep)
```

```
##################
## 3b
##################

set.seed(12345)
pp <- t.test(rchisq(n = 10, df = 1), rchisq(100, 2) - 1, alternative = "two.sided", var.equal = T)$p.val

##################
## 3bI
##################

#assumptions
met <- c("yes", "yes", "yes", "no", "yes", "no", "no", "yes")
print(xtable(data.frame(assumptions, met), caption = "I. Which assumptions of the specified test are me


##################
## 3bIII
##################

nsim <- 10000
foo <- data.frame(rep = replicate(nsim, t.test(rchisq(n = 10, df = 1), rchisq(100, 2) - 1, alternative
ggplot(foo) +
     geom_histogram(aes(x = rep), col = "grey40", alpha = 0.5, binwidth = 0.05)+
     theme_bw() +
     theme(
          plot.background = element_blank()
          ,panel.grid.major = element_blank()
          ,panel.grid.minor = element_blank()
          ,panel.border = element_blank()
           ) +

     #draws x and y axis line
     theme(axis.line = element_line(color = 'black')) +
     labs(x = "p values", title = "p values from 10000 replications -- #3b")

alpha <- length(foo$rep[foo$rep <= 0.05])/length(foo$rep)


##################
## 3cI
##################

yCalc <- function(){
     tsamp <- rt(50,df = 20)
     zsamp <- rnorm(50)
     ff <- data.frame(y = tsamp - zsamp, x = tsamp)
     return(t.test(ff$x, ff$y, alternative = "less", var.equal = T)$p.value)
}

#var(tsamp)
#var(ff$y)
pv <-yCalc()
```

```
#assumptions
met <- c("no","yes", "no", "no", "yes", "no", "no", "yes" )
print(xtable(data.frame(assumptions, met), caption = "I. Which assumptions of the specified test are me



##################
## 3cIII
##################

yCalcW <- function(){
     tsamp <- rt(50,df = 20)
     zsamp <- rnorm(50)
     ff <- data.frame(y = tsamp - zsamp, x = tsamp)
     return(t.test(ff$x, ff$y, alternative = "less", var.equal = T)$p.value)
}

nsim <- 10000
foo <- data.frame(rep = replicate(nsim, yCalcW()))

ggplot(foo) +
     geom_histogram(aes(x = rep), col = "grey40", alpha = 0.5, binwidth = 0.05)+
     theme_bw() +
     theme(
          plot.background = element_blank()
          ,panel.grid.major = element_blank()
          ,panel.grid.minor = element_blank()
          ,panel.border = element_blank()
           ) +

     #draws x and y axis line
     theme(axis.line = element_line(color = 'black')) +
     labs(x = "p values", title = "p values from 10000 replications -- #3c")

alpha <- length(foo$rep[foo$rep <= 0.05])/length(foo$rep)


##################
## 3d
##################

dds <- function(nsamp = 50){
     tsamp <- rt(nsamp,df = 20)
     zsamp <- rnorm(nsamp)
     ff <- data.frame(dat = c((tsamp - zsamp), tsamp), trt = c(rep("y", nsamp), rep("x", nsamp)))
     return(ff)
}


sp <- dds(50000)

sp$PctChange <- sp$dat
sp$SpeedLimit <- sp$trt
```

```
grid <- with(sp, seq(min(PctChange), max(PctChange), length = 100))

normaldens <- ddply(sp, "SpeedLimit",
                    function(df) {
                         data.frame(
                                   predicted = grid,
                                   density = dnorm(grid, mean(df$PctChange), sd(df$PctChange)))
                    })

# look at distributions of data
ggplot(sp, aes(x = PctChange)) +
     # histogram
     geom_histogram(aes(y = ..density.., fill = "Histogram"), color = "grey40",
                     alpha = 0.2) +
     # kernel density line
     geom_line(aes(y = ..density..,  lty = "Density"), stat = 'density')+
     # normal line
     geom_line(aes(y = density, x = predicted, lty = "Normal"), data = normaldens) +
     # facet
     facet_grid(~SpeedLimit)  +

     # labels and theme
     xlim(c(-3.5, 3.5))+
     labs(x = "Value", title = "Distributions for #3d, from 50000 samples") +
     theme_bw() +
     theme(legend.background = element_rect(colour = "black"),
          plot.background = element_blank()
          ,panel.grid.major = element_blank()
          ,panel.grid.minor = element_blank()
          ,panel.border = element_blank()
          ,axis.line = element_line(color = 'black')) +
     # Names for the legend
     scale_linetype(name = "Line")+
     scale_fill_manual(name = "Histogram", values = c("black"))


##################
## 3dII
##################
nsim <- 10000
tdat <- function() {
     foo <- dds()
     t.test(foo$dat ~ foo $trt, var.equal = F, alternative = "less", paired = T)
}
#tdat()$p.value

foo <- data.frame(rep = replicate(nsim, tdat()$p.value))
ggplot(foo) +
     geom_histogram(aes(x = rep), col = "grey40", alpha = 0.5, binwidth = 0.05)+
     theme_bw() +
     theme(
          plot.background = element_blank()
          ,panel.grid.major = element_blank()
          ,panel.grid.minor = element_blank()
          ,panel.border = element_blank()
           ) +

     #draws x and y axis line
     theme(axis.line = element_line(color = 'black')) +
     labs(x = "p values", title = "p values from 10000 replications -- #3d")

alpha <- length(foo$rep[foo$rep <= 0.05])/length(foo$rep)

##################
## 4a
##################
sp <- read.csv("data/ex0223.csv")
#levels(sp$SpeedLimit)

vv <- var(sp$PctChange[sp$SpeedLimit == "Inc"])/var(sp$PctChange[sp$SpeedLimit == "Ret"])

pv <- t.test(sp$PctChange~sp$SpeedLimit, var.equal = T, alternative = "greater")
#pv$estimate[1] - pv$estimate[2]
# this tests mean(Inc) - mean(Ret) > 0

##################
## 4b
##################

sp <- read.csv("data/ex0223.csv")
sp$SpeedLimit <- ifelse(sp$SpeedLimit == "Inc", yes = sp$SpeedLimit <- "Increased Speed Limit", no = sp$

grid <- with(sp, seq(min(PctChange), max(PctChange), length = 100))

normaldens <- ddply(sp, "SpeedLimit",
                    function(df) {
                         data.frame(
                                   predicted = grid,
                                   density = dnorm(grid, mean(df$PctChange), sd(df$PctChange)))
                    })

# look at distributions of data
ggplot(sp, aes(x = PctChange)) +
     # histogram
     geom_histogram(aes(y = ..density.., fill = "Histogram"), color = "grey40",
                     alpha = 0.2, binwidth = 5) +
     # kernel density line
     geom_line(aes(y = ..density..,  lty = "Density"), stat = 'density')+
     # normal line
     geom_line(aes(y = density, x = predicted, lty = "Normal"), data = normaldens) +
     # facet
     facet_grid(~SpeedLimit)  +

     # labels and theme
     xlim(c(-30, 30))+
     labs(x = "Percent Change", title = "Distributions for #23 -- to check for normality") +
     theme_bw() +
     theme(legend.background = element_rect(colour = "black"),
          plot.background = element_blank()
          ,panel.grid.major = element_blank()
          ,panel.grid.minor = element_blank()
          ,panel.border = element_blank()
          ,axis.line = element_line(color = 'black')) +
     # Names for the legend
     scale_linetype(name = "Line")+
     scale_fill_manual(name = "Histogram", values = c("black"))

##################
## 4c
##################

sp <- read.csv("data/ex0223.csv")
numInc <- sum(sp$SpeedLimit == "Inc")
numRet <- sum(sp$SpeedLimit == "Ret")

tstat <- with(sp, mean(PctChange[SpeedLimit == "Inc"]) - mean(PctChange[SpeedLimit == "Ret"]))

perr <- function() {
     ind <- sample(1:length(sp$PctChange), numInc)
     return(mean(sp$PctChange[ind]) - mean(sp$PctChange[-ind]))
}


nsim <- 10000
foo <- data.frame(rep = replicate(nsim, perr()))

ggplot(foo) +
     geom_histogram(aes(x = rep), col = "grey40", alpha = 0.5, binwidth = 0.5)+
     theme_bw() +
     theme(
          plot.background = element_blank()
          ,panel.grid.major = element_blank()
          ,panel.grid.minor = element_blank()
          ,panel.border = element_blank()
           ) +

     #draws x and y axis line
     theme(axis.line = element_line(color = 'black')) +
     labs(x = "Test Statistic Value",
          title = "Histogram of Test Statistics from 10,000 permutations") +
     geom_vline(xintercept=tstat, lty = 5)


alpha <- length(foo$rep[foo$rep >= tstat])/length(foo$rep)
```