Assignment4

Callin Switzer

September 29, 2014

2a

kv26 <- c(5.79, 1579.52, 2323.70) kv28 <- c(68.8, 108.29, 110.29, 426.07, 1067.60)

y1 <- log(kv26) y2 <- log(kv28)

2b-d on other paper

2c. The antilog of the output from b is 1.3426

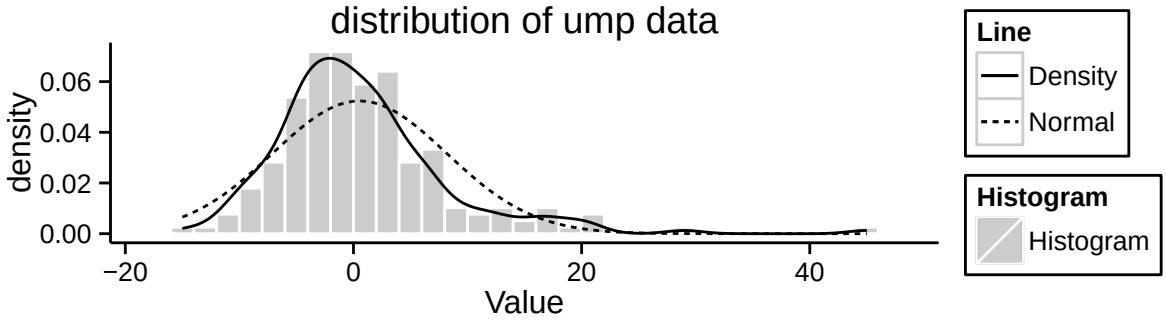
#2 According to the two-sided t-test, using log-transformed data, we fail to reject the null hypothesis that the two means are the same. We reject because the p-value is 0.8951

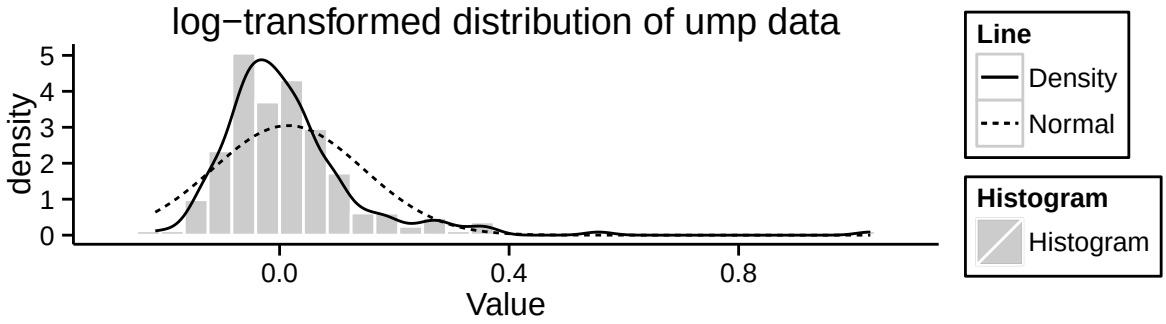
3

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

Table 1: Assumptions of t test

The following assumptions may not be met. 1. Independence might not be met, because the people who tend to be umps might not be fully independent. For instance, they umping might be more popular in certain parts of the US, where people are eating different diets – this could affect lifespan. 2. Homogeneity within the population might not be met, because it could similarly be affected by location. 3. The population isn't normally distributed – in this case, we're looking at the difference of the two groups (since we're using a paired t-test). See the distributions below. Also, we haven't randomly sampled from the population. Though these assumptions are not met, I will still go through with the analysis, using a paired t-test.





From the distributions, I don't think a log transformation is appropriate. I've also tried other transformations, and I cannot find one that makes the data more normal. However, I will run the ttest with and without the oulier and see what happens.

3a

H0: Expected - Lifelength = 0 (i.e. umps have shorter lives than expected) HA: Expected - Lifelength > 0

According to our ttest with untransformed data, the one-sided pvalue is 0.1624, which means we fail to reject the null hypothesis that the umpire life lengths are shorter than expected.

I also reran the t-test, excluding the outlier, and found that the pvalue was 0.2675. This pvalue is still not significant, so we still fail to reject the null hypothesis.

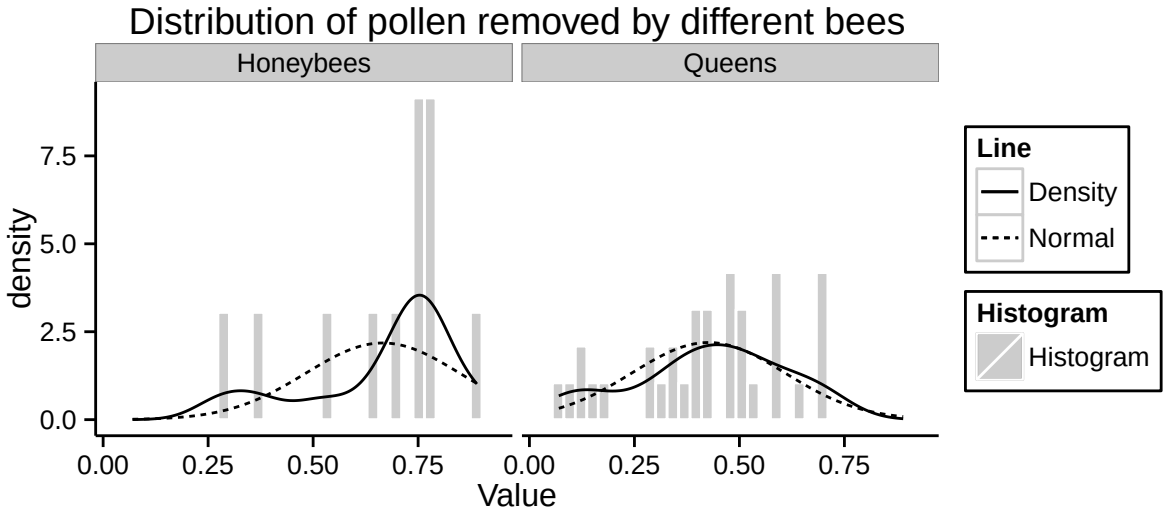
3b

The potential consequences of ignoring the 212 of the 441 original umpires could have consequences. If we reanalyzed, using a different method, we might find a signifcant result. Ignoring half of the data could greatly throw off our conclusions. It probably gives us a sample that is not random, and probably not representative.

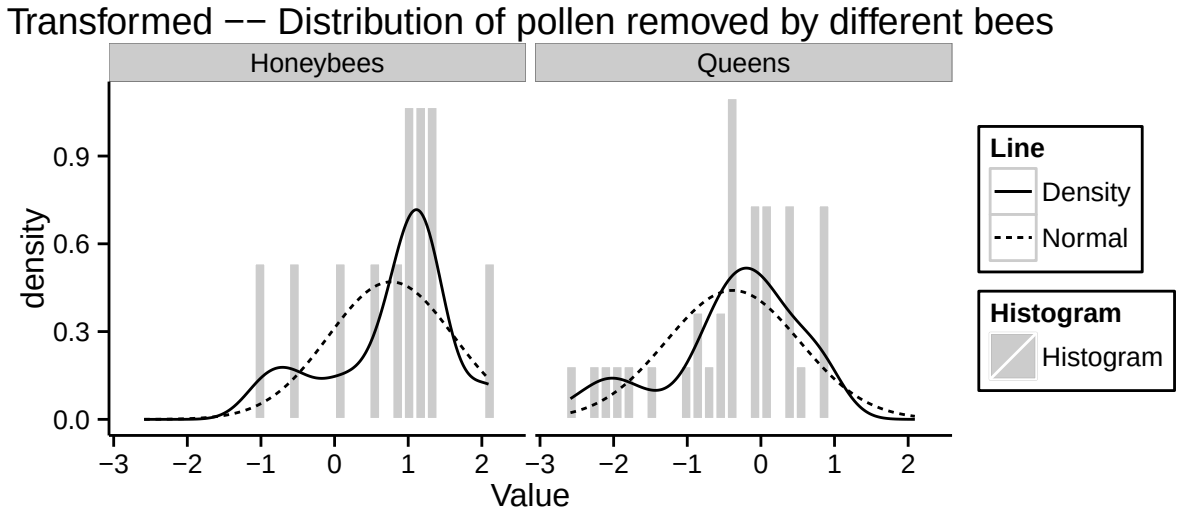
3c

Consequences of ignoring the censored data exist. We could analyze these data using a nonparametric test, and that may give us better insight. If we ignore the censored data, we may be biasing our sample – making it not representative of the population.

4ai



4aii

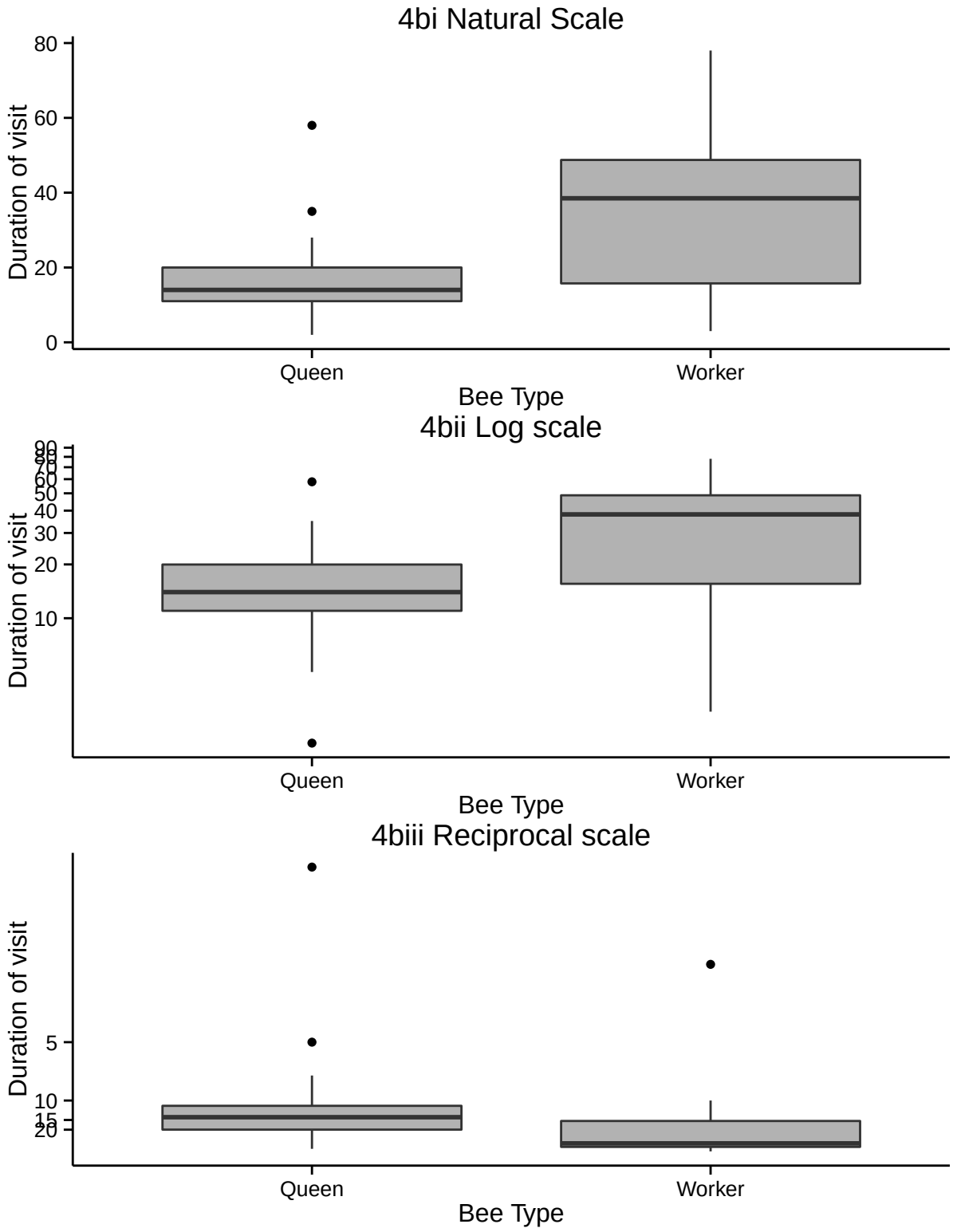


#4aiii

H0: The distributions of proportion removed is equal among the two bee types HA: The distributions of proportion removed is not equal among the two bee types

Since the ratio of variances is 1.1366, I will use a two-sample ttest with pooled variances. According to the ttest, the pvalue is 3.7151 × 10-4, which suggests that we reject H0.

4b


4bi Natural Scale
Duration of visit
80
60
40
20
0
Queen
Worker
Bee Type
4bii Log scale
Duration of visit
90
80
75
60
50
40
30
20
10
Queen
Worker
Bee Type
4biii Reciprocal scale
Duration of visit
5
10
15
20
Queen
Worker
Bee Type

4biv

The scale that seems most appropriate is the log scale, because it looks like the two groups distributions are most symmetrical on that scale. Since the ratio of variances is 0.386, I will use an unpooled ttest.

4bv

The 95% confidence interval for the log-transformed data is (-1.275, -0.0274). Since the study was observational, when we use exp(conf. int) we're getting an estimate of median(worker)/median(queen). exp(conf. int) = (0.2794, 0.973), which is the confidence interval for the ratio above. This means that we expect the median duration of time spent by workers to be between 0.28 and 0.97 times the median duration of time spent by queens.

4bvi

- Natural scale requires very little interpretation – your confidence interval and test can be described on the original scale. It is the easiest for a reader to understand. *Log transformation can be easily converted to an interpretation on the original scale (see #4bv), so it is easy to describe and easy for the reader to understand.
- Reciprocal transformation can be interpreted as a rate – which is also easy for the reader to understand.

4bvii

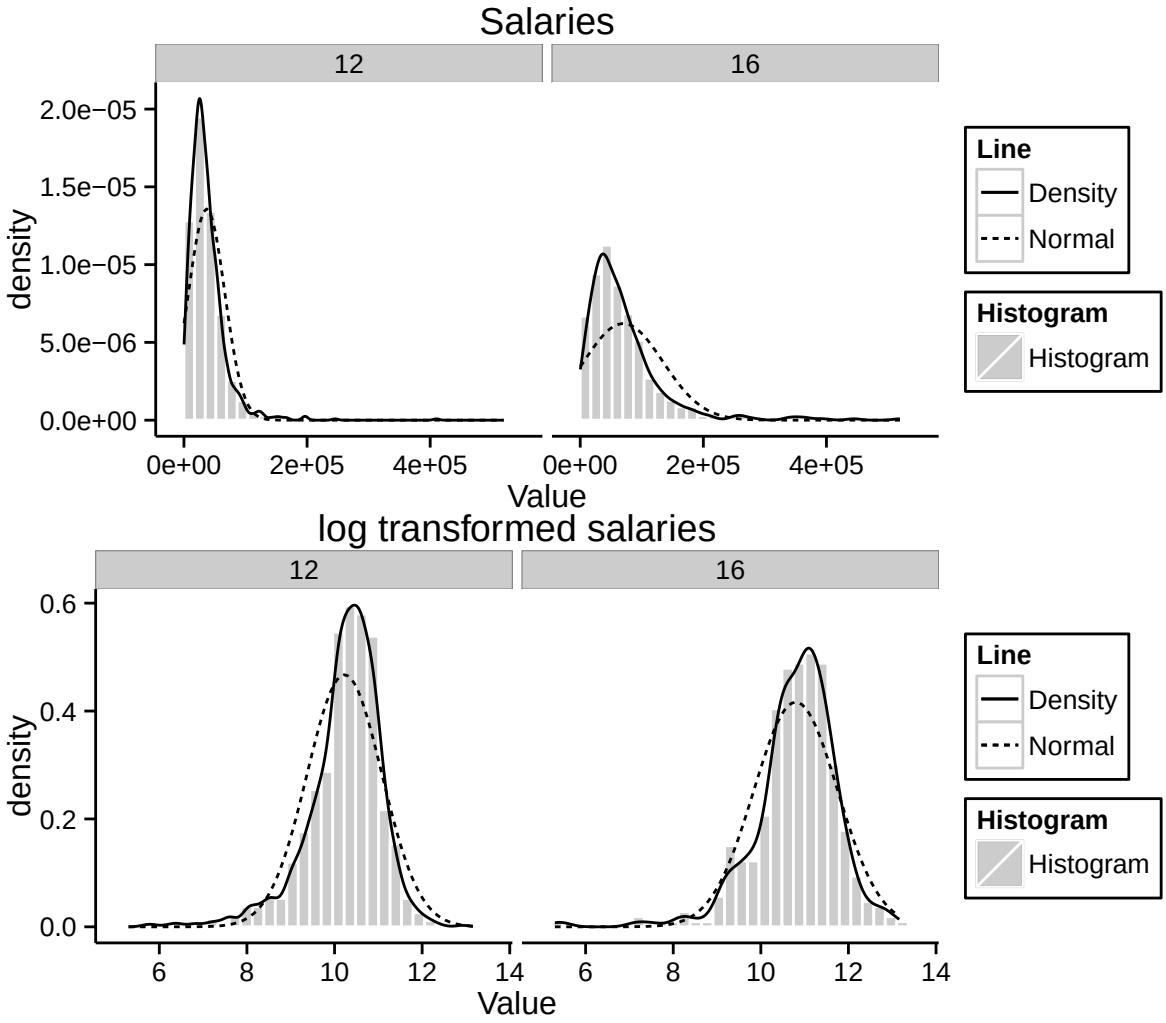
Assessing equality of population standard deviation is very difficult. With small samples, the estimate of std dev can be quite different from the actual std dev.

5

	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	unclear, by maybe

Table 2: Assumptions of T Test

I will assume that the units are independent, though I cannot be sure without reading more about the methods. Also, I'll assume that the sample is random, though I can't tell from the description. The main violations of assumptions are that the variances are not equal vetween populations and the populations are not normally distributed. A log transformation makes the data fit the assumptions of equal variances between populations and normal distributions much better – see histograms below.



H0: exp("salary at education level 12" - "salary at education level 16") = 1

HA: exp("salary at education level 12" - "salary at education level 16") not = 1

In other words, we're estimating the ratio of salary for education level 12 / salary for eduation level 16 (since the distributions are symmetric).

On the log-transformed data, I used a two-sample pooled t-test, because the variances were approximately equal. The resulting test had a p-value of <2.2e-16, which suggests we should reject our null hypothesis. The confidence interval for the log-transformed data is (0.468, 0.6717). When we take the antilog of that interval we get (1.5968, 1.9576). This is interpreted as the following: the median of the salary at education level 16 is 1.6 to 1.96 times as large as the median salary of the people with education level 12.

Code

```
######################
###### Functions
######################

tAssumpChecker <- function(group1,
                           group2,
                           group1title = "group1",
                           group2title = "group2",
                           graphTitle = "Histograms for Groups") {

    mdat <- data.frame(dat = c(group1, group2),
                       trt = c(rep(group1title, length(group1)),
                               rep(group2title, length(group2))))
    sp <- data.frame(row.names = 1:nrow(mdat))
    sp$PctChange <- mdat$dat
    sp$SpeedLimit <- mdat$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 = "white",
                         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 = graphTitle) +
         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"))
     ## the distribution looks like we need to log-transform the data
}



histNormKern <- function(diffs = diffs, title = "Distribution"){
     colnames(diffs) <- "diffs"
     ggplot(diffs, aes(x = diffs)) +
     geom_histogram(aes(y = ..density.., fill = "Histogram"), color = "white",
                    alpha = 0.2)+
     # density line
     geom_line(aes(y = ..density..,  lty = "Density"), stat = 'density')+
     #geom_line(aes(y = ..density..,  lty = "Normal"), stat = 'normal')+
     # normal approx
     stat_function(aes(lty = "Normal"), fun=dnorm,
                   args=list(mean=mean(diffs$diffs), sd=sd(diffs$diffs)))+

     labs(x = "Value", title = title) +
     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"))
}

########################
###### Setup
########################

options(xtable.comment = FALSE)
require(xtable)
require(ggplot2)
require(plyr)
library("scales")
set.seed(12345)
########################
###### Q2
########################

kv26 <- c(5.79, 1579.52, 2323.70)
kv28 <- c(68.8, 108.29, 110.29, 426.07, 1067.60)

y1 <- log(kv26)
y2 <- log(kv28)
```

```r
mean(y1, na.rm=T)
mean(y2, na.rm = T)

#data.frame(y1,y2)

ts <- mean(y1, na.rm=T) - mean(y2, na.rm = T)

antil <- exp(mean(y1, na.rm=T) - mean(y2, na.rm = T))

liqT <- t.test(y1,y2, var.equal = F)

var(y1, na.rm = T)/var(y2)

SE = sqrt(var(y1)/3 + var(y2)/5)

sx1 <- var(y1)
sx2 <- var(y2)

df <-  (sx1 / 3 + sx2/5)^2 / (1/2 * (sx1/3)^2 + 1/4 * (sx2/5)^2)
df

t = qt(p = 0.975, df = df)
t

ts - t*SE
ts + t*SE
exp(ts - t*SE)
exp(ts + t*SE)

########################
###### Q3
########################

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

#assumptions
print(xtable(data.frame(assumptions), caption = "Assumptions of t test",
             type = "latex"))

ump <- read.csv("data/ex0321.csv")
ump <- ump[ump$Censored == "0",]
ump$Censored <- NULL

diffs <- data.frame(diffs = ump$Expected - ump$Lifelength)
histNormKern(diffs, "distribution of ump data")

ump1 <- log(ump)
diffsLog <- data.frame(diffs = ump1$Expected - ump1$Lifelength)
```

```r
histNormKern(diffsLog, "log-transformed distribution of ump data")

diffsO <- diffs[diffs$diffs != max(diffs$diffs),]

td <- t.test(diffs$diffs, alternative = "greater")$p.value # t test with outlier
to <- t.test(diffsO, alternative = "greater")$p.value # t test without outlier
########################
###### Q4
########################
pol <- read.csv("data/ex0327.csv")

qu <- pol$PollenRemoved[ pol$BeeType == "Queen"]
hb <- pol$PollenRemoved[ pol$BeeType == "Worker"]

tAssumpChecker(group1 = qu, group2 = hb,
               group1title = "Queens", group2title = "Honeybees",
               "Distribution of pollen removed by different bees")


qu1 <- log(qu/(1-qu))
hb1 <- log(hb/(1-hb))

tAssumpChecker(group1 = qu1, group2 = hb1,
               group1title = "Queens", group2title = "Honeybees",
               "Transformed -- Distribution of pollen removed by different bees")

# ttest
vv <- var(qu1)/var(hb1)
ts <- t.test(qu1, hb1,alternative = "two.sided", var.equal = T )$p.value

mm <- ggplot(pol, aes(x = BeeType, y = DurationOfVisit)) +
     geom_boxplot(fill = "grey40", alpha = 0.5) +
     labs(x = "Bee Type", y = "Duration of visit",
          title = "4bi Natural Scale") +
          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'))
mm

mm + scale_y_continuous(trans=log_trans(), breaks = c((1:10)*10)) +
     labs(title = "4bii Log scale")

mm + scale_y_continuous(trans=reciprocal_trans(), breaks = c((1:4)*5)) +
     labs(title = "4biii Reciprocal scale")

pol$DurationOfVisitLog <- log(pol$DurationOfVisit)
rv <- var(pol$DurationOfVisitLog[pol$BeeType == "Queen"]) /
     var(pol$DurationOfVisitLog[pol$BeeType == "Worker"])
```

```r
lt <- t.test(pol$DurationOfVisitLog~pol$BeeType,alternative = "two.sided",
             var.equal = F )

##################
## Q5
##################

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("yes", "yes", "yes", "no", "yes", "no", "no", "unclear, by maybe")

print(xtable(data.frame(assumptions, met), caption = "Assumptions of T Test",
             type = "latex"))

work <- read.csv("data/ex0330.csv")

tw <- work$Income2005[work$Educ == 12]
six <- work$Income2005[work$Educ == 16]
tAssumpChecker(tw, six, "12", "16", "Salaries")
tAssumpChecker(log(tw), log(six), "12", "16", "log transformed salaries")

twl <- log(tw)
sixl <- log(six)

#var(twl)/var(sixl)

tw <- t.test(sixl, twl, alternative = "two.sided", var.equal = T)
```