Statistical Inference

Aroogz

November 9, 2016

SYNOPSIS: The Analysis consists of two part

1. will investigate the exponential distribution in R and compare it with the Central Limit Theorem
2. will analyse the `ToothGrowth` dataset elucidating on the effect of `Supp` and `dose` variables on the `len`.

Preparing the Workspace

```
rm(list = ls())
if ("ggplot2" %in% row.names(installed.packages()) == FALSE){install.packages("ggplot2")}
if ("grid" %in% row.names(installed.packages()) == FALSE){install.packages("grid")}
if ("gridExtra" %in% row.names(installed.packages()) == FALSE){install.packages("gridExtra")}
library(ggplot2); library(grid); library(gridExtra)
```

Investigating the Exponential Distribution

get the cumulative means and variances

```
mns <- NULL; vars <- NULL; lambda <- 0.2
# simulating 1000 means and variances
for (i in 1:1000){
  mns <- c(mns, mean(rexp(40, lambda)))
  vars <- c(vars, var(rexp(40, lambda)))
}
# getting the mean of the means
cummeans <- cumsum(mns)/(1:1000); cumvar <- cumsum(vars)/(1:1000)
```

the theoretical values

```
mean.theor <- 1/lambda
var.theor <- (1/lambda)^2
print(paste("theoretical mean: ", mean.theor))
```

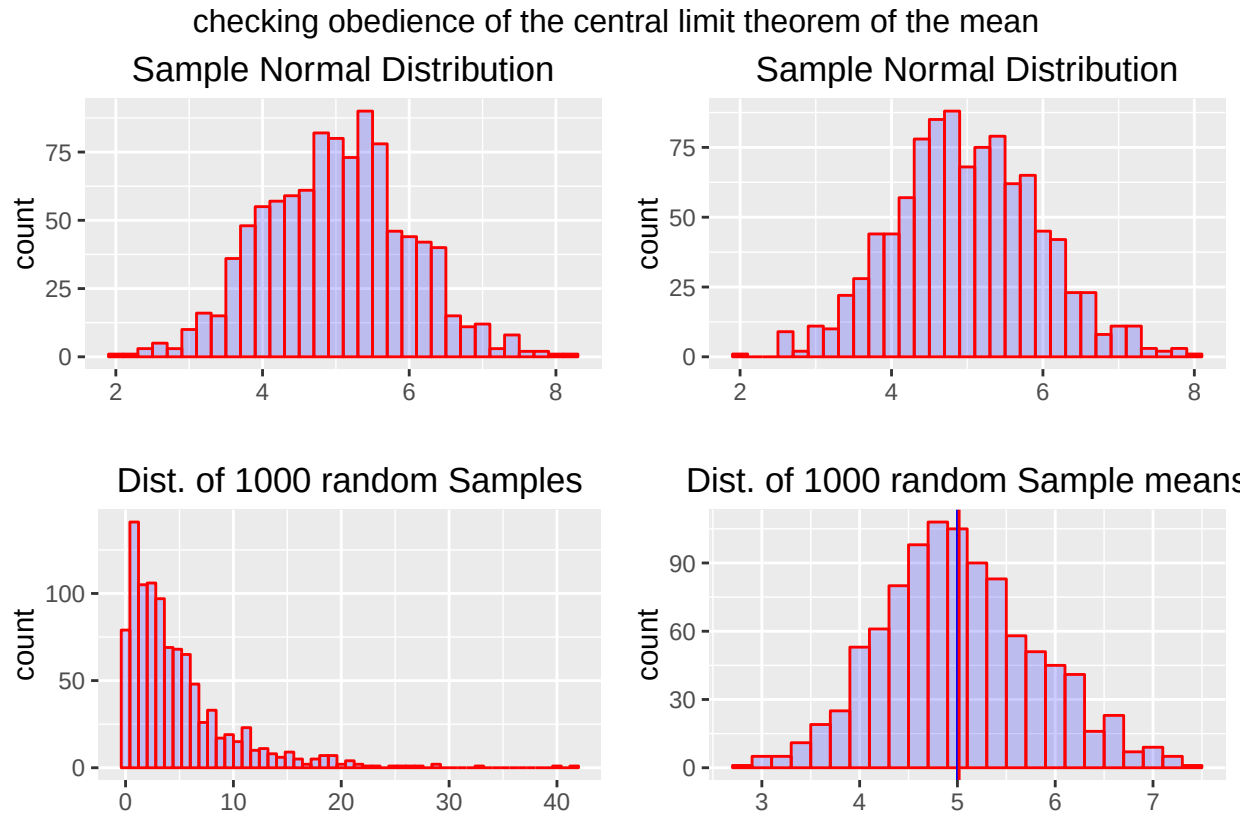
```
## [1] "theoretical mean:  5"
```

```
print(paste("theoretical variance: ", var.theor))
```

```
## [1] "theoretical variance:  25"
```

the distribution of means converging to a normal

```
set.seed(100)
#sample normal distribution
p1 <- ggplot(mapping = aes(rnorm(1000, mean = mean.theor)))+
  geom_histogram(binwidth = 0.2, col="red", fill= "blue", alpha=0.2)+
  labs(x = "", title= "Sample Normal Distribution")
#distribution of 1000 random exps; not gaussian
p2 <- ggplot(data.frame(val = rexp(1000, 0.2)), aes(val))+
  geom_histogram(binwidth = 0.8, col="red", fill= "blue", alpha=0.2)+
  labs(x="", title= "Dist. of 1000 random Samples")
# distribution of means
mns.data <- data.frame(mns)
p3 <- ggplot(mns.data, aes(mns))+
  geom_histogram(binwidth = 0.2, col="red", fill= "blue", alpha=0.2)+
  geom_vline(xintercept = c(mean.theor, mean(mns)), col= c("blue", "red"),
                            size = 0.5)+
  labs(x="", title= "Dist. of 1000 random Sample means")
# arranging plots
set.seed(100)
lay <- rbind(c(1,2),c(3,4))
grid.arrange(p1, p1, p2, p3, layout_matrix= lay,
             top = "checking obedience of the central limit theorem of the mean")
```

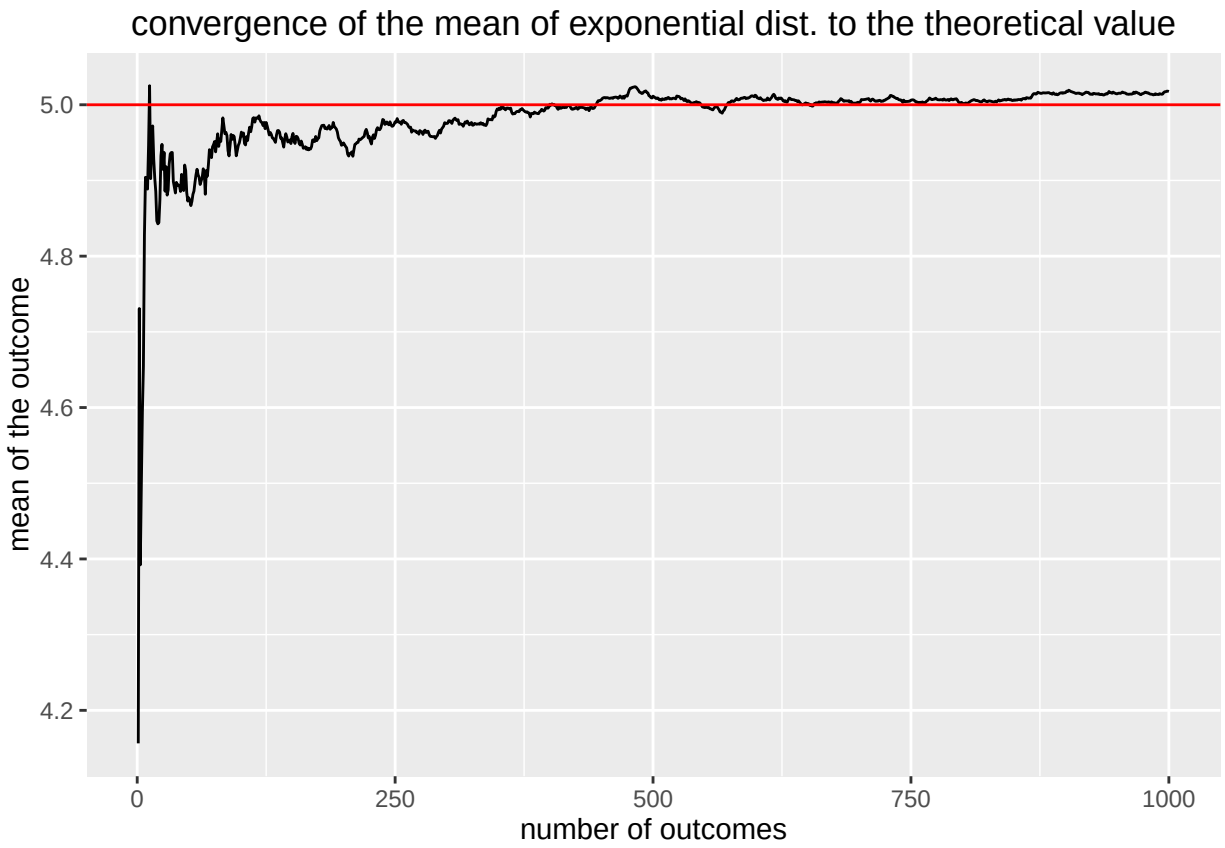


it is observable that the distribution of the means of the exponential distribution (bottom right) converges to

a Normal distibution as compared the distribution of a 1000 exponentials on the bottom left. Thus showing the Central Limit Theorem at work.

showing the convergence of the mean distribution

```
ggplot(mapping = aes(1:1000, cummeans))+
  geom_line()+
  geom_hline(yintercept = mean.theor, col= "red")+
  labs(x = "number of outcomes", y = "mean of the outcome",
       title= "convergence of the mean of exponential dist. to the theoretical value")
```



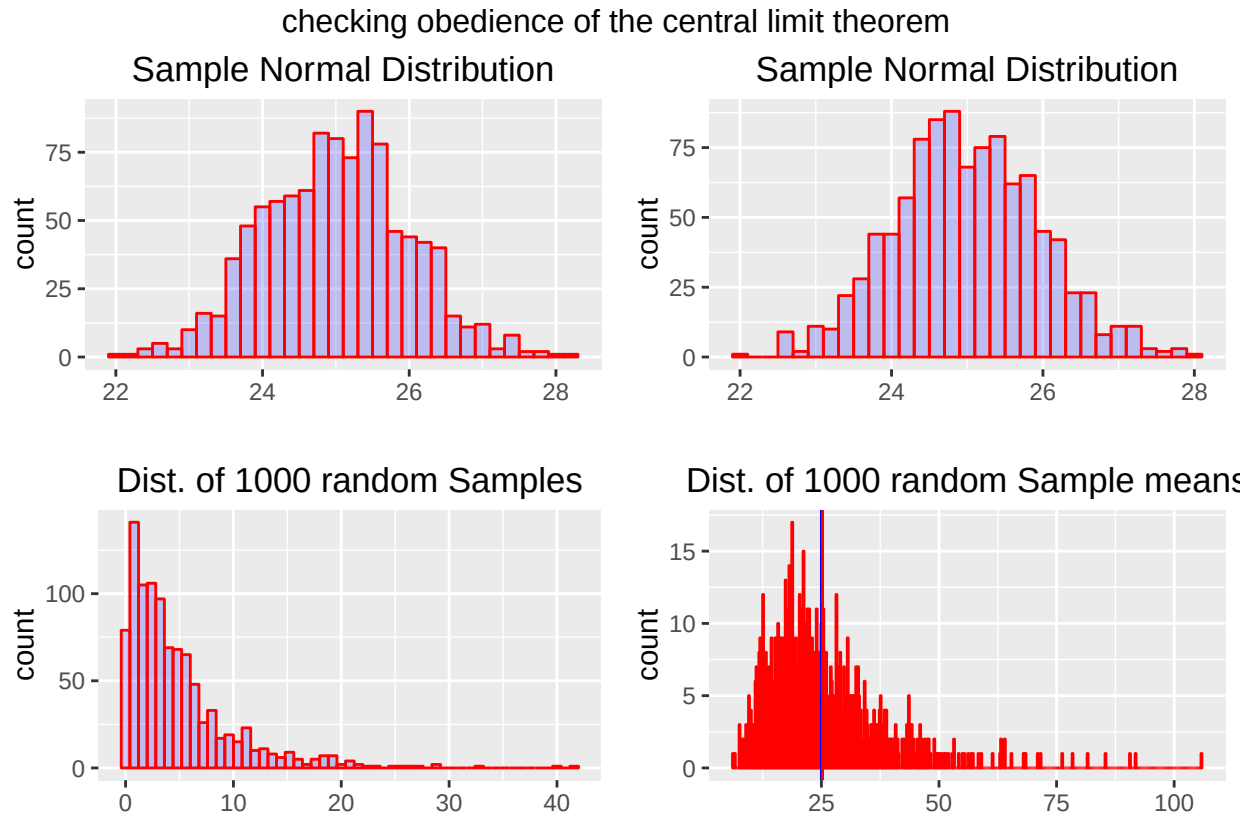
we see in the plot above as the mean distribution converges to the theoretical mean, `theo.mean` (marked by the horizontal line) as the outcomes increase

the distribution of variances converging to normal

```
#sample normal distribution
p5 <- ggplot(mapping = aes(rnorm(1000, mean = var.theor)))+
  geom_histogram(binwidth = 0.2, col="red", fill= "blue", alpha=0.2)+
  labs(x = "", title= "Sample Normal Distribution")

# distribution of variances
p6 <- ggplot(mapping= aes(vars))+
```

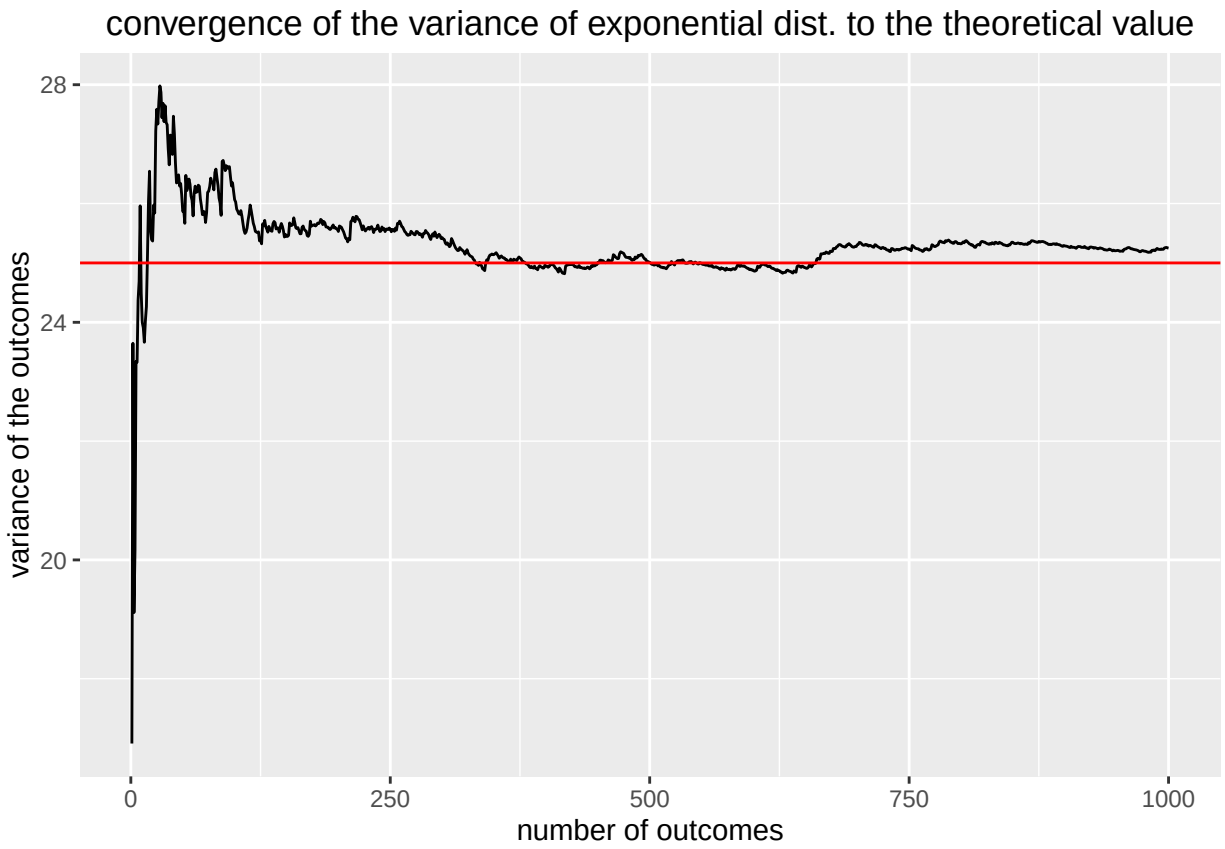
```
  geom_histogram(binwidth = 0.2, col="red", fill= "blue", alpha=0.2)+
  geom_vline(xintercept = c(var.theor, mean(vars)), col= c("blue", "red"),
             size = 0.5)+
  labs(title= "Dist. of 1000 random Sample means", x= "")
set.seed(100)
lay2 <- rbind(c(1, 2),
              c(3, 4))
grid.arrange(p5,p5, p2, p6, layout_matrix= lay2,
             top= "checking obedience of the central limit theorem")
```



it is observable that the distribution of the variances of the exponential distribution (bottom right) converges to a Normal distibution as compared the distribution of a 1000 exponentials on the bottom left. Thus showing the Central Limit Theorem at work.

showing the convergence of the mean variances

```
ggplot(mapping = aes(1:1000, cumvar))+
  geom_line()+
  geom_hline(yintercept = var.theor, col= "red")+
  labs(x = "number of outcomes", y = "variance of the outcomes",
       title= "convergence of the variance of exponential dist. to the theoretical value")
```



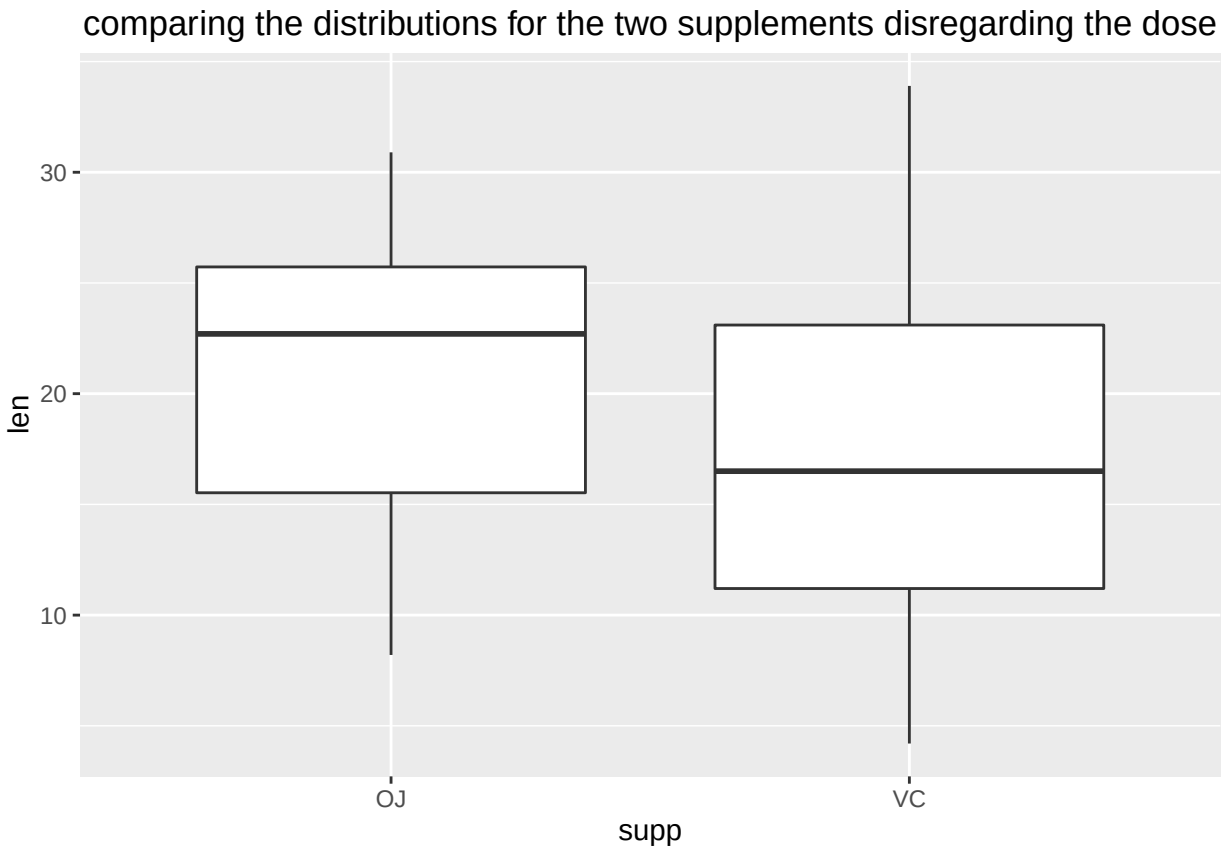
we see in the plot above as the mean distribution converges to the theoretical variance `theor.var` (marked by the horizontal line) as the outcomes increase.

2. The analysis of the `ToothGrowth` dataset

The dataset consists of 60 rows and 3 columns. It looks intuitive that the observations are paired by the `supp` variable so we would consider this case.

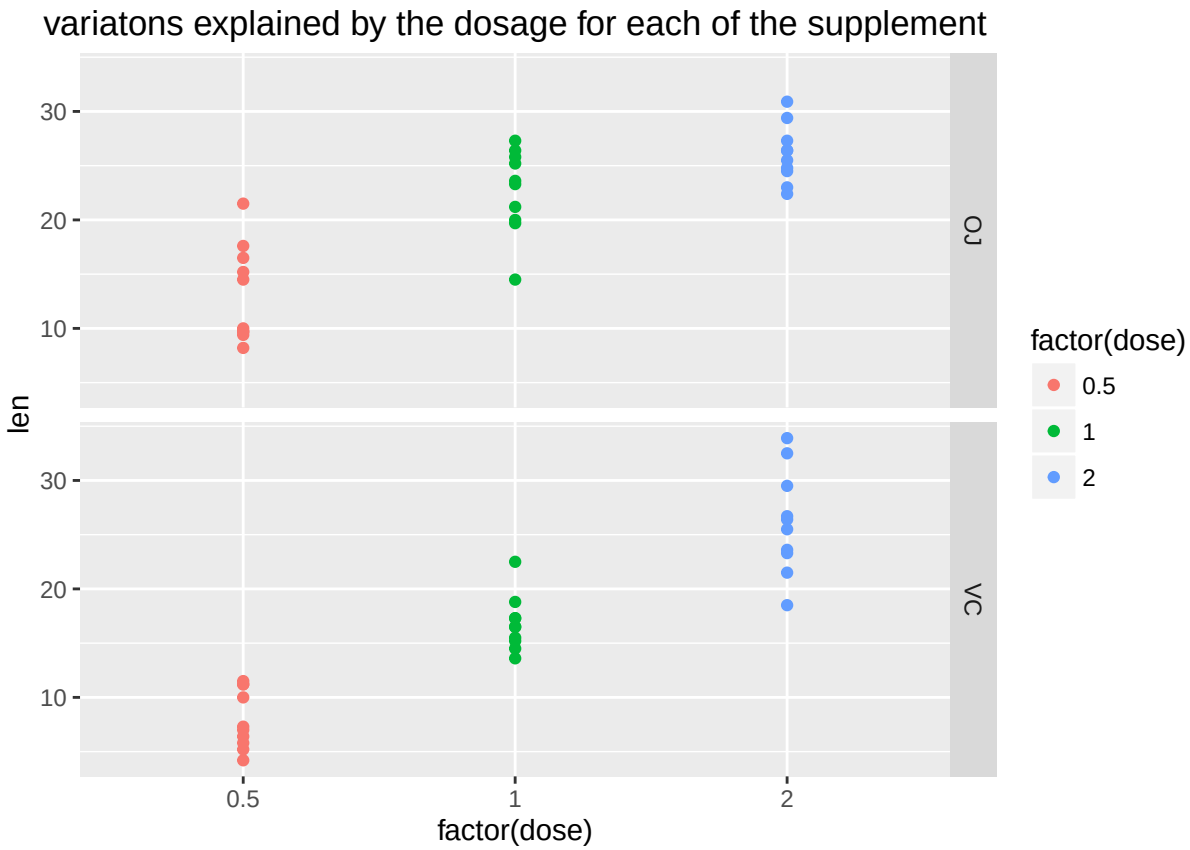
Visualising the data showing some plots

```
ggplot(data = ToothGrowth)+
  geom_boxplot(mapping = aes(x=supp, y= len) ) +
  labs(title= "comparing the distributions for the two supplements disregarding the dose")
```



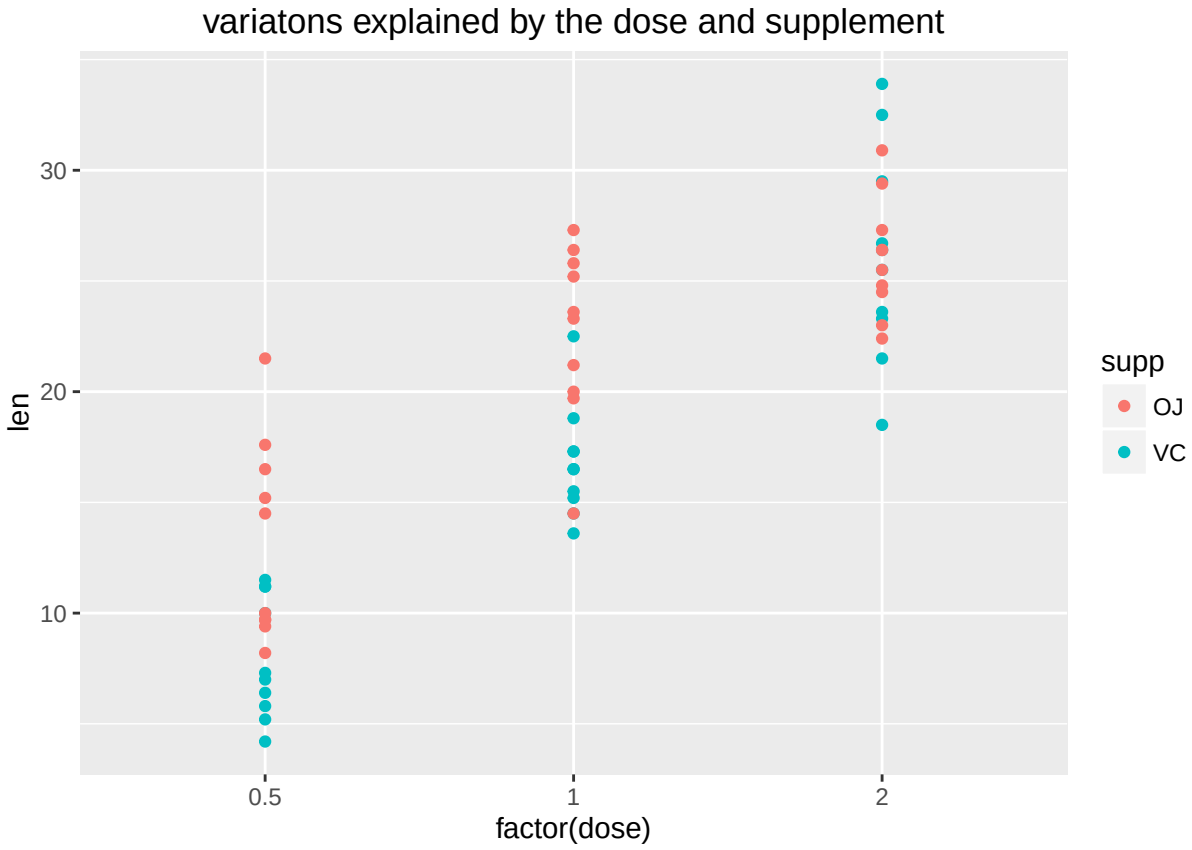
Here we observe the distribution of the `len` variable with respect to the `supp` variable

```
# var explained by dose, grid by supp
ggplot(ToothGrowth)+
  geom_point(mapping = aes(x=factor(dose), y=len, col= factor(dose)))+
  facet_grid(supp~.)+
  labs(title= "variatons explained by the dosage for each of the supplement")
```



It is obvious from the above, the effect of the `dose` on the outcome of the `len` for both categories of `supp` variable

```
#exploring variations due to all factors
ggplot(ToothGrowth)+
  geom_point(mapping = aes(x=factor(dose), y=len, col= supp))+
  labs(title= "variatons explained by the dose and supplement")
```



from the above plot, we observe an increase in the OJ `supp` for the first two dose (0.5 and 1). It becomes unclear the effect of the different supplements ### Assuming paired observation

```
data("ToothGrowth"); n <- nrow(ToothGrowth)/2

#divison alongn supp variable
g1 <- ToothGrowth$len[1:30]; g2 <- ToothGrowth$len[31:60]

# showing the confidence interval (Assuming paired observations)
mn <- mean(g2-g1); s <- sd(g2-g1)
mn + c(-1, 1)*qt(0.975, n-1)*s/sqrt(n)
```

```
## [1] 1.408659 5.991341
```

we see the interval (95% confidence interval) here does not contain zero which implies that we could confidently reject the hypothesis of a zero mean difference. This implies that the variability due to the `supp` variable is indeed significant and could not have happened by chance. ### Assuming unpaired observation

```
t.test(x= g2, y = g1, paired = FALSE)$conf
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

the unpaired test however shows an interval that contains zero and so we do not have enough evidence to reject the hypothesis of a 0 mean difference

implementing the permutation method

```
#implement the permutation test
cal.diff <- function(value, group){
  # function to get the mean different for a list
  # differentiated by groups (another column)
  divs <- levels(group)
  meandiff <- mean(value[group == divs[1] ]) - mean(value[group == divs[2]])
  meandiff
}
observation <- cal.diff(ToothGrowth$len, ToothGrowth$supp)
permutations <- NULL
for (i in 1:1000){
  permutations <- c(permutations, cal.diff(ToothGrowth$len,
                                           sample(ToothGrowth$supp)))
}
print(paste("the observed mean difference: ", observation))
```

```
## [1] "the observed mean difference:  3.7"
```

```
### check fraction of permutation greater than the observation
mean(permutations > observation)
```

```
## [1] 0.028
```

we see a low percentage of the permutations greater than the observation

```
quantile(permutations, c(0.025,0.975))
```

```
##      2.5%     97.5%
## -3.848167  3.746833
```

we see it is highly unlikely to have gotten this value if it were left to randomness.

```
quantile(permutations, 0.95)
```

```
##   95%
## 3.195
```

the 95% quantile is also less than the observation indication our observation is very unlikely due to chance. hence there is a difference in the `len` is associated with the `supp`. Particularly, the mean difference is of the `len` is not zero in favour of the `OJ` `supp` category.