Assignment-3

Mayank Raunak

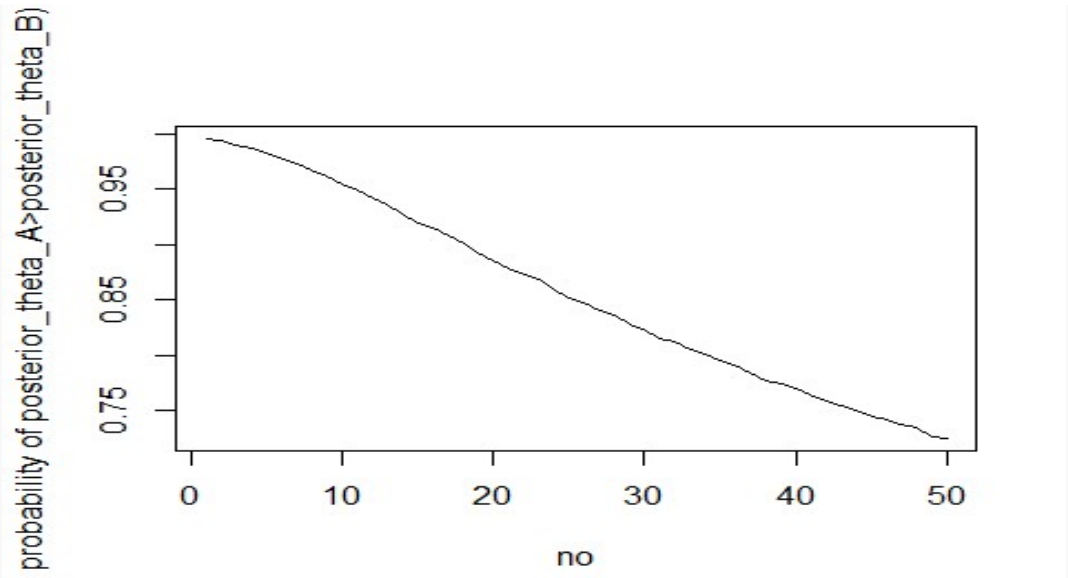
October 12, 2019

4.2 (Part a)

```
yA = c(12,9,12,14,13,13,15,8,15,6)
yB = c(11,11,10,9,9,8,7,10,6,8,8,9,7)
sum_yA=sum(yA)
sum_yB=sum(yB)
n_A=length(yA)
n_B=length(yB)
#prior_∂A =gamma(120,10), prior_∂B =gamma(12,1), p(∂A,∂B) = p(∂A)×p(∂B)
posterior_theta_A =rgamma(100000,120+sum_yA,10+n_A); posterior_theta_B =rgamma(100000,12+sum_y
B,1+n_B)

mean(posterior_theta_A>posterior_theta_B)

## [1] 0.99532
```

4.2(Part b)

```
no=seq(1,50)

#prior_∂A =gamma(120,10), prior_∂B =gamma(12*no,no), p(∂A,∂B) = p(∂A)×p(∂B)
posterior_theta_A =rgamma(100000,120+sum_yA,10+n_A); posterior_theta_B =rgamma(100000,12*no+su
m_yB,no+n_B)

mean(posterior_theta_A>posterior_theta_B)

for (no in seq(1,50)){
posterior_theta_A =rgamma(100000,120+sum_yA,10+n_A); posterior_theta_B =rgamma(100000,12*no+su
m_yB,no+n_B)

k=(mean(posterior_theta_A>posterior_theta_B))
k_1 <- c(k_1, k)
no_1<-c(no_1,no)
}


plot(no_1,k_1,type='l',ylab='probability of posterior_theta_A>posterior_theta_B)',xlab='no')
```



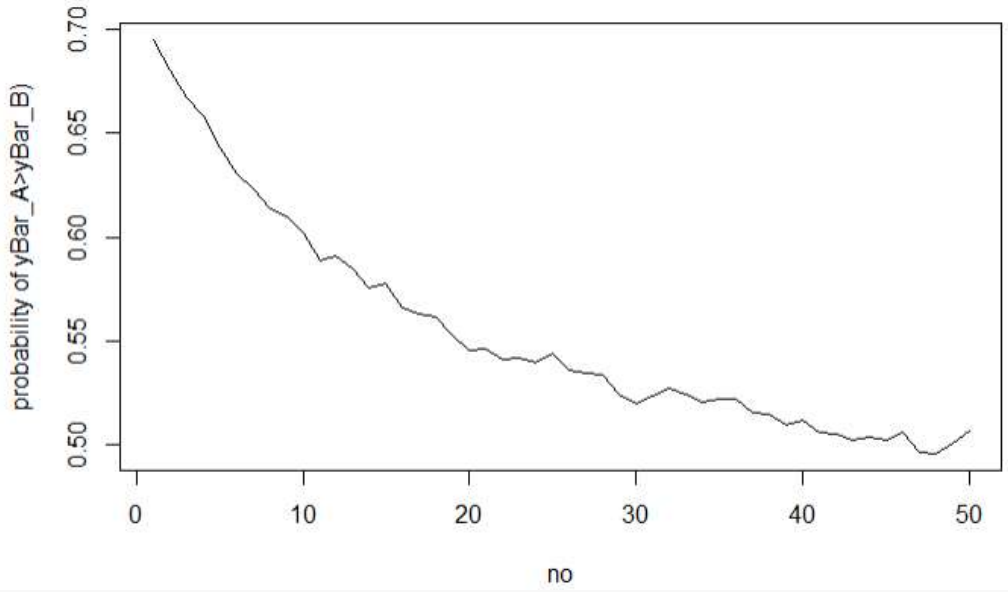
From the above graph we can observe that the probability of posterior_theta_A > posterior_theta_B decreases as we incre ase the value of no. The strong prior beliefs lead to nearer posterior distribution of theta_A and theta_B.

The conclusion about the event (theta_A>theta_B) is highly sensitive to the prior distribution, when the sample size is less. I observed that when I sampled only 10 values, I found in both the probability as1,which shows that it was highly sensitive to prior knowledge on the other hand as I increased the sample size to 100000, I observed the probability of the event (theta_A>theta_B) in part (a) was 0.99552 which shows that, the result has influence of the sampling and it is less sensitive to the prior beliefs as our sample size increases.

Part C(i)

```
yA = c(12,9,12,14,13,13,15,8,15,6)
yB = c(11,11,10,9,9,8,7,10,6,8,8,9,7)
sum_yA=sum(yA)
sum_yB=sum(yB)
n_A=length(yA)
n_B=length(yB)

a1=120;b1=10;a2=12;b2=1
#prior_∂A =gamma(120,10), prior_∂B =gamma(12,1), p(∂A,∂B) = p(∂A)×p(∂B)
for ( s in 1:10000) {

  theta_A<-rgamma(1 , a1+sum_yA , b1+n_A)

  yBar_A<-rpois (n_A , theta_A )

}

for ( s in 1:10000) {

  theta_B<-rgamma(1 , a2+sum_yB , b2+n_B)

  yBar_B<-rpois (n_B , theta_B)

}

mean(yBar_A>yBar_B)
```

```
[1] 0.6923077
```



we observed that the probability of yBar_A > yBar_B decreases as we increase the value of no. Here, the decrease is steeper as compared to the probability of theta_A >theta_B. In case of probability of theta_A >theta_B, the decrease was gradual.

4.5 part(a)

```
## the posterior distribution of θ given data (Y1,X1),...,(Yn,Xn) and a gamma(a,b) prior distribution
```

$\theta|(Y_1, X_1), \ldots (Y_n, X_n) = gamma(a + \sum_{i=1}^{n} Y_i, b + n)$

```
my_data_noreact <- read.table('http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/cancer_noreact.dat',header=T)


n1=sum(my_data_noreact$x)
sum_y1=sum(my_data_noreact$y)

my_data_react <- read.table('http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/cancer_react.dat',header=T)
n2=sum(my_data_react$x)
sum_y2=sum(my_data_react$y)
```

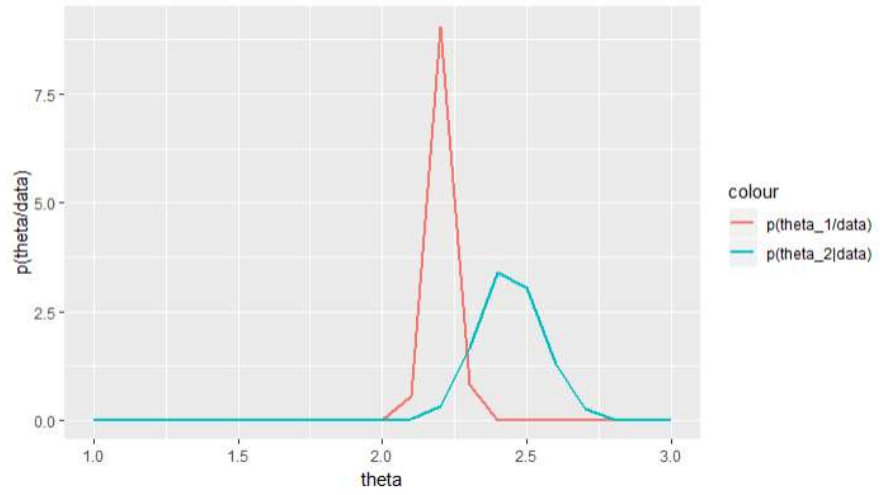
Part (b)(i) considering, a_1=2,b_1=1,a_2=3,b_2=2
$\theta_1|(Y_1, X_1), \ldots (Y_{n1}, X_{n1}) = gamma(a_1 + \sum_{i=1}^{n1} Y_{1,i}, b_1 + n1)$ gamma(2+sum_y1,1+n1)=gamma(2287,79)

$\theta_2|(Y_1, X_1), \ldots (Y_{n2}, X_{n2}) = gamma(a_2 + \sum_{i=1}^{n2} Y_{2,i}, b_2 + n2)$ gamma(2+sum_y2,1+n2)=gamma(259,12)

Part (c) ## cancer rates from previous years have been roughly ~ θ = 2.2 per 10,000

#i. Opinion 1: (a1 = a2 = 2.2×100,b1 = b2 = 100) $\theta_1 = gamma(a_1 + 2285, b_1 + n1)$ =gamma(2505,1137) $\theta_2 = gamma(a_2 + 256, b_2 + n2)$ =gamma(476,195)

```
 # E[ϑ1|data]
Expected_value_theta_1=2505/1137
Expected_value_theta_1

## [1] 2.203166

# E[ϑ2|data]
Expected_value_theta_2=476/195
Expected_value_theta_2

## [1] 2.441026

#95% quantile-based posterior intervals for ϑ1
intervals_theta_1= qgamma(c(0.025,0.975),2505,1137)
intervals_theta_1

## [1] 2.117726 2.290273

#95% quantile-based posterior intervals for ϑ2
intervals_theta_2= qgamma(c(0.025,0.975),476,195)
intervals_theta_2

## [1] 2.226633 2.665131

 #Pr(ϑ2 > ϑ1|data).
posterior_theta_1 =rgamma(100000,2505,1137); posterior_theta_2 =rgamma(100000,476,195)
mean(posterior_theta_1<posterior_theta_2)

## [1] 0.97818

#plot
theta=seq(1,3,0.1);
library(ggplot2)
data <- data.frame( theta=theta, noraect = dgamma(theta,2505,1137), react = dgamma(theta,476,1
95))
ggplot(data, aes(theta)) + geom_line(aes(y=noraect, color='p(theta_1/data)'),size=1) + geom_li
ne(aes(y=react, color='p(theta_2|data)'), size=1)+ylab('p(theta/data)')
```

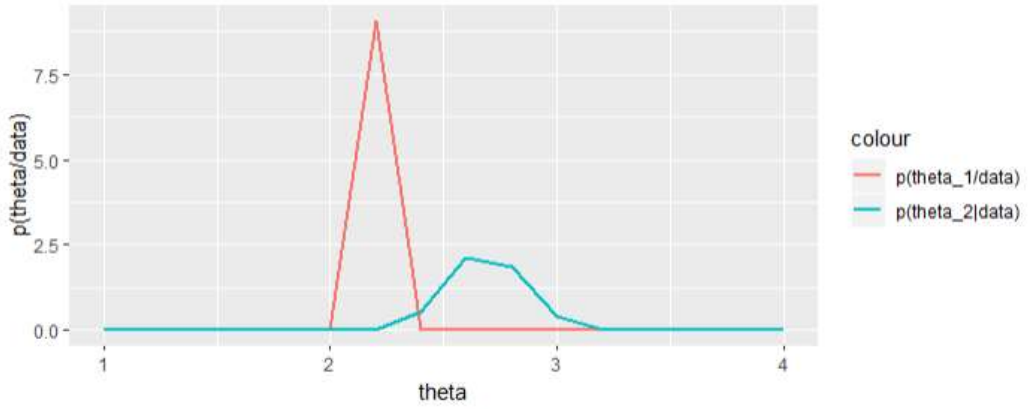


In opinon1: we had prior belief that cancer rates for both types of counties are similar to average rate across all counties from previous year, this belief changed after observing the data, for not near nuclear reactor counties, the posterior expectation was same(2.2) but for the near nuclear reactor counties, the value was slightly higher, and the probability of fatalities of being higher in near nuclear reactor counties is 0.9782,

#ii. Opinion 2: (a1 = 2.2×100,b1= 100,a2=2.2,b2=1) $\theta_1 = gamma(a_1 + 2285, b_1 + n1)$ =gamma(2505,1137) $\theta_2 = gamma(a_2 + 256, b_2 + n2)$ =gamma(258.2,96)

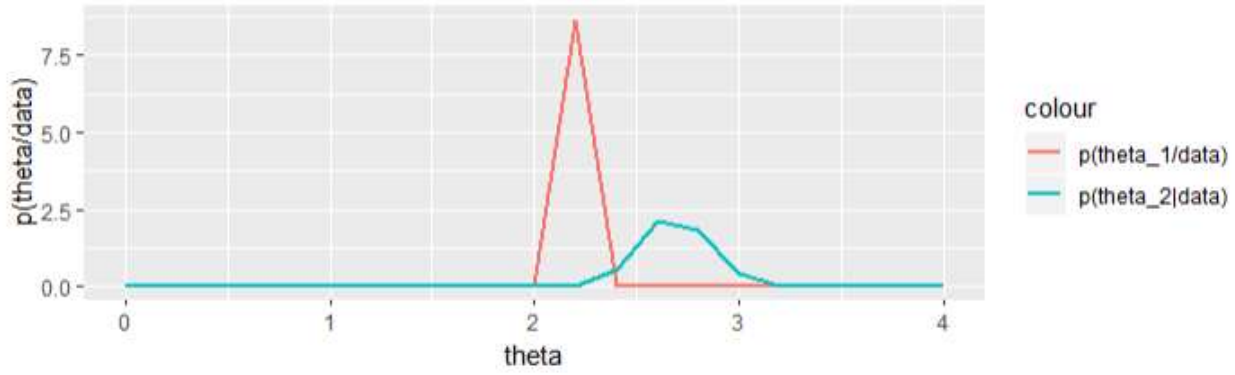
```
# E[ϑ1|data]
Expected_value_theta_1=2505/1137
Expected_value_theta_1

## [1] 2.203166

# E[ϑ2|data]
Expected_value_theta_2=258.2/96
Expected_value_theta_2

## [1] 2.689583

#95% quantile-based posterior intervals for ϑ1
intervals_theta_1= qgamma(c(0.025,0.975),2505,1137)
intervals_theta_1

## [1] 2.117726 2.290273

#95% quantile-based posterior intervals for ϑ2

intervals_theta_2= qgamma(c(0.025,0.975),258.2,96)
intervals_theta_2

## [1] 2.371497 3.027397

#Pr(ϑ2 > ϑ1|data).
posterior_theta_1 =rgamma(100000,2505,1137); posterior_theta_2 =rgamma(100000,258.2,96)
mean(posterior_theta_1<posterior_theta_2)

## [1] 0.99847

#plot
theta=seq(1,4,0.2);
library(ggplot2)
data <- data.frame( theta=theta, noraect = dgamma(theta,2505,1137), react = dgamma(theta,258.2
,96))
ggplot(data, aes(theta)) + geom_line(aes(y=noraect, color='p(theta_1/data)'),size=1) + geom_li
ne(aes(y=react, color='p(theta_2|data)'), size=1)+ylab('p(theta/data)')
```

In opinon2: Cancer rates in this year for nonreactor counties are similar to rates in previous years in nonreactor counties, the posterior expectation for nonreactor counties remains the same(2.2). This means that fatalities rate for nonreactor counties was similar to that of previous year.



#iii. Opinion 3: (a1 =a2=2.2,b1=b2=1) $\theta_1 = gamma(a_1 + 2285, b_1 + n1)$ =gamma(2287.2,1038) $\theta_2 = gamma(a_2 + 256, b_2 + n2)$ =gamma(258.2,96)

```
# E[ϑ1|data]
Expected_value_theta_1=2287.2/1038
Expected_value_theta_1

## [1] 2.203468

# E[ϑ2|data]
Expected_value_theta_2=258.2/96
Expected_value_theta_2

## [1] 2.689583

#95% quantile-based posterior intervals for ϑ1
intervals_theta_1= qgamma(c(0.025,0.975),2287.2,1038)
intervals_theta_1

## [1] 2.114081 2.294680

#95% quantile-based posterior intervals for ϑ2
intervals_theta_2= qgamma(c(0.025,0.975),258.2,96)
intervals_theta_2

## [1] 2.371497 3.027397

#Pr(ϑ2 > ϑ1|data).
posterior_theta_1 =rgamma(100000,2287.2,1038); posterior_theta_2 =rgamma(100000,258.2,96)
mean(posterior_theta_1<posterior_theta_2)

## [1] 0.99861

#plot
theta=seq(0,4,0.2);
library(ggplot2)
data <- data.frame( theta=theta, noraect = dgamma(theta,2287.2,1038), react = dgamma(theta,258
.2,96))
ggplot(data, aes(theta)) + geom_line(aes(y=noraect, color='p(theta_1/data)'),size=1) + geom_li
ne(aes(y=react, color='p(theta_2|data)'), size=1)+ylab('p(theta/data)')
```



We observed that Pr(θ2 > θ1|data) for different opinions and found that the for the first opinion it is 0.9782 and for 2nd and the 3rd opinion it is 0.99, higher than the first opinion. This may be due to the reason that, we had stronger prior belief that both the counties have same fatality rate in first opinion

After observing the data and the different opinions, we concluded that the expected value of fatality rate for nonreactor counties was similar to that of the previous year, but the expected value of the fatality rate was higher for the reactor counties as compared to the fatality rate of previous year as well as fatality rate of nonreactor in the current year. However, data for the reactor counties was less, hence, it is difficult to make a strong statement about the fatality rate regarding reactor counties.

Overall, from the above information and graphs, we can conclude that the fatality rate in the reactor counties are higher than the fatality rate in the nonreactor counties.

```
5

# Downloaded the Data from the site
data <- list()
data[1] <- read.table('http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/school1.dat',header=T)
data[2] <-read.table('http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/school2.dat',header=T)
data[3] <- read.table('http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/school3.dat',header=T)

# Given Data for the conjugate prior
mu0 <- 5
k0 <- 1
s20 <- 4
nu0 <- 2

# Sample size
n <- sapply(data, length)

# Sample mean
ybar <- sapply(data, mean)

#Sample variance
s2 <- sapply(data, var)

## Posrerior Inference

# Total sample size for estimating mu
kn <- k0 + n

nun <- nu0 + n
#mean of current and prior observation
mun <- (k0 * mu0 + n * ybar)/kn

s2n <- (nu0 * s20 + (n - 1) * s2 + k0 * n * (ybar - mu0)^2/kn)/nun


#created three matrices containing 3 columns for each schools
#sampled variance from the distribution and eventually standard deviation and then sampled the
ta from its posterior distribution, here s2.posterior is variance s.posterior is standard devi
ation and theta.posterior is (theta\data)


s.posterior <- s2.posterior <- theta.posterior <- matrix(0, 10000, 3, dimnames = list(NULL, c(
"school1", "school2", "school3")))
```

```r
for (i in c(1, 2, 3)) {
    s2.posterior[, i] <- 1/rgamma(10000, nun[i]/2, s2n[i] * nun[i]/2)
    s.posterior[, i] <- sqrt(s2.posterior[, i])
    theta.posterior[, i] <- rnorm(10000, mun[i], s.posterior[, i]/sqrt(kn[i]))

}

# posterior mean
colMeans(theta.posterior)
```

```
##  school1  school2  school3
## 9.574555 7.230911 7.971956
```

```r
# 95% CI for Theta(mean)
apply(theta.posterior, 2, function(x) {
    quantile(x, c(0.025, 0.975))
})
```

```
##         school1  school2  school3
## 2.5%   8.122896 5.472430 6.264617
## 97.5% 11.084760 9.015137 9.658561
```

```r
# posterior mean and 95% CI for the Standard Deviations
colMeans(s.posterior)
```

```
##  school1  school2  school3
## 3.703196 4.255858 3.756389
```

```r
apply(s.posterior, 2, function(x) {
    quantile(x, c(0.025, 0.975))
})
```

```
##        school1  school2  school3
## 2.5%  2.831445 3.228098 2.785707
## 97.5% 4.940331 5.717917 5.136450
```

5.1(b)

```r
# determine the ranks of each posterior sample for theta.
require(combinat)
```

```
## Loading required package: combinat
```

```
## Warning: package 'combinat' was built under R version 3.5.2
```

```
##
## Attaching package: 'combinat'
```

```
## The following object is masked from 'package:utils':
##
##     combn
```

```r
theta.ranks <- c(apply(theta.posterior, 1, rank))
theta.probs <- list()
for (i in permn(3)) {
    index <- apply(theta.ranks, 1, function(row) {
        all(row == i)
    })
    theta.probs[[paste(i, collapse = ",")]] <- length(theta.ranks[index, 1])/10000
    }

theta.probs[["1,2,3"]]
```

```
## [1] 0.0047
```

```
theta.probs[["1,3,2"]]
```

```
## [1] 0.0033
```

```
theta.probs[["2,1,3"]]
```

```
## [1] 0.07
```

```
theta.probs[["2,3,1"]]
```

```
## [1] 0.0155
```

```
theta.probs[["3,1,2"]]
```

```
## [1] 0.6539
```

```
theta.probs[["3,2,1"]]
```

```
## [1] 0.2526
```

5.1.(c)

```
# the posterior probability that ~ Yi < ~ Yj < ~ Yk for all six permutations

predicted.postsample <- matrix(0, 10000, 3, dimnames = list(NULL, c("school1", "school2", "sch
ool3")))
for (i in c(1, 2, 3)) {
    predicted.postsample[, i] <- rnorm(10000, mun[i], sqrt(s2.posterior[, i] * (1 + 1/kn[i])))
}
# determine the ranks of each posterior sample for a prediction.
pred.ranks <- c(apply(predicted.postsample, 1, rank))
predicted_sample.probs <- list()
for (i in permn(3)) {
    index <- apply(pred.ranks, 1, function(row) {all(row == i)
    })
    predicted_sample.probs[[paste(i, collapse = ",")]] <- length(pred.ranks[index, 1])/10000
}
```

```
predicted_sample.probs[["1,2,3"]]
```

```
## [1] 0.1006
```

```
predicted_sample.probs[["1,3,2"]]
```

```
## [1] 0.1025
```

```
predicted_sample.probs[["2,1,3"]]
```

```
## [1] 0.178
```

```
predicted_sample.probs[["2,3,1"]]
```

```
## [1] 0.1458
```

```
predicted_sample.probs[["3,1,2"]]
```

```
## [1] 0.2672
```

```
predicted_sample.probs[["3,2,1"]]
```

```
## [1] 0.2059
```

5.1(d)

```
# the posterior probability that ϑ1 is bigger than both ϑ2 and ϑ3, from above part

theta.probs[["3,1,2"]] + theta.probs[["3,2,1"]]

0.6539 + 0.2526

## [1] 0.9065

# the posterior probability that ~ Y1 is bigger than both ~ Y2 and ~ Y3
predicted_sample.probs[["3,1,2"]] + predicted_sample.probs[["3,2,1"]]

=0.2672+0.2059

## [1] 0.4731
```

Reference was taken from

https://www.tutorialspoint.com/r/index.htm,

http://www.r-tutor.com/r-introduction ,

http://rstudio-pubsstatic.s3.amazonaws.com/4952_7af63ededf804cbc9dbc85dda50d07e3.html,
https://stackoverflow.com/questions/2871763/is-there-an-r-function-to-get-the-number-of-permutations-of-n-objects-take-k-pn