581-Assignment_5

GROUP PROJECT: SriKaavya Toodi(st938), Rithvik (ra830),Aravind reddy(ad1452),Sowmya(sa1607)

10/13/2020

Contents

Function to estimate

Method to estimate parameters for different distributions is below. Input parameters are "distribution_type" : Type of distribution. To be given as a string input "data_input" : Data generated using the specific distribution to be given as input

Mean and variance and given data is calculated and using Method of moments, calculations will be done to estimate specific parameter values. These will be printed after calculations.

We've to note that calculating variance for the data set implicitly has second moment in it. Using variance for estimating a parameter is almost similar to using second moment of the given data.

We've attempted the extra credit question. Method call and code are below. Derivation for hyper geometric parameters is available at the end of the document.

```r
library(purrr)
library(MASS)
 Parameter_finder <- function( distribution_type , data_input , k) {
  # This function is used to calculate parameters for
  #different distributions of data using mean and variance
  if(!distribution_type=="multinomial"){
   print(head(data_input))
  }
  mean_data = mean(data_input)
  var_data = var(data_input)

  if ( distribution_type == "Poisson") {
    print(paste0("Poisson distribution parameters - Lambda value is ", mean_data))
  }
  else if (  distribution_type == "Point mass at a") {
    print(paste0("Point mass at a distribution parameters - a value is", mean_data))
  }
  else if (  distribution_type == "Bernoulli") {
```

```r
  print(paste0("Bernoulli distribution parameters - p value is", mean_data))
}
else if (  distribution_type == "Binomial") {
  p = var_data/mean_data
  p = 1-p
  n = mean_data/p
  print(paste0("Binomial distribution parameters - p value is ",p))
  print(paste0("N value is ", n))
}
else if (  distribution_type == "Geometric") {
  p = 1/mean_data
  print(paste0("Geometric distribution parameters - p value is ", p))
}
else if (  distribution_type == "Uniform") {
  #sub_ba = (var_data/mean_data) * 6
  #sum_ab = mean_data * 2
  b = mean_data + (sqrt(3*var_data))
  a = mean_data - (sqrt(3*var_data))
  print(paste0("Uniform distribution parameters - a value is ", a,", b value is ", b))
}
else if (  distribution_type == "Normal") {
  print(paste0("Normal distribution parameters - Mean value is ", mean_data))
  print(paste0("Standard deviation is ", sqrt(var_data)))
}
else if (  distribution_type == "Exponential") {
  print(paste0("Exponential distribution parameters - rate value is ", 1/mean_data))
}
else if (  distribution_type == "Gamma") {
  beta = var_data/mean_data
  alpha = mean_data/beta
  print(paste0("Gamma distribution parameters - alpha value is", alpha))
  print(paste0("Beta value is ",beta, " and rate is ", 1/beta))
}
else if (  distribution_type == "Beta") {
  m <- mean_data
  v <- var_data
  alpha <- m*((m*(1-m)/v)-1)
  beta <- alpha*(1-m)/m
  print(paste0("Beta distribution parameters - alpha value is", alpha))
  print(paste0("Beta value is ",beta))
}
else if (  distribution_type == "tdist") {
  v = 2*var_data/(var_data-1)
  print(paste0("t distribution distribution parameters -v is ", v))
}
else if (  distribution_type == "Chi Square") {
  print(paste0("Chi Square distribution parameters - p value is ", mean_data))
}

else if (distribution_type=="multinomial")
{
  n_row = nrow(data_input)
  prob = c(0,0,0,0)
```

```
    for(i in 1:n_row)
      p[i]<-1-((var(data_input[i,]))/mean(data_input[i,]))
    n = sum(rowMeans(data_input))/sum(p[1:n_row])
    print(list(n=n, p=p))
  }
  else if (distribution_type=="multivariatenormal")
  {
    mean = colMeans(data_input)
    Sigma = var(data_input)
    print(list(mu_val=mean, summation=Sigma))
  }else if (distribution_type=="hypergeometric") {
    N <- length(data_input)
    p <- mean_data/k
    product <- var_data/(k * p * (1-p))
    m_plus_n <- (k-product)/(1-product)
    m <- p * m_plus_n
    print(paste0("N value is ", N, " Subclass 1(m) value is ", m, " Subclass 2(n) value is ", m_plus_n-
  }
  #print(paste0('Mean of data is ', mean_data))
  #print(paste0('Variation of data is ', var_data))
}
```

Testing distributions

Testing different distributions below

```
# Bernoulli is Binomial with size as 1
bernoulli_data <- rbinom(1000, 1, 0.75)
distribution_type = 'Bernoulli'
Parameter_finder(distribution_type,bernoulli_data)
```

```
## [1] 1 1 1 1 0 0
## [1] "Bernoulli distribution parameters - p value is0.768"
```

```
#Binomial distribution
binom_data <- rbinom(100, 1000, 0.75)
distribution_type = 'Binomial'
Parameter_finder(distribution_type,binom_data)
```

```
## [1] 751 760 756 771 743 737
## [1] "Binomial distribution parameters - p value is 0.733397046486386"
## [1] "N value is 1023.78377932824"
```

```
# Geometric distribution. Here parameter couln't be estimated even with really high value of N
geom_data <- rgeom(100000, 0.25)
distribution_type = 'Geometric'
Parameter_finder(distribution_type,geom_data)
```

```
## [1]  2 13  5  8  4  1
## [1] "Geometric distribution parameters - p value is 0.333044694598015"
```

```
# Poisson distribution. Lamdba is accurately estimated
poisson_data <- rpois(40000, lambda = 3)
distribution_type = 'Poisson'
Parameter_finder(distribution_type,poisson_data)
```

```
## [1] 7 3 1 3 1 1
## [1] "Poisson distribution parameters - Lambda value is 3.0032"
```

```r
#Uniform distribution data. Here a and B are accurately estimated if gap between them is large enough.
#If not,it seems inaccurate
 uniform_data <- runif(1000000, 1, 100)
distribution_type = 'Uniform'
Parameter_finder(distribution_type,uniform_data)
```

```
## [1] 50.228441 43.208789  5.444854 77.886212 98.108988  3.855769
## [1] "Uniform distribution parameters - a value is 1.06722685249467, b value is 100.024897200006"
```

```r
# Normal distribution
norm_data <- rnorm(100000, 20, 2)
distribution_type = 'Normal'
Parameter_finder(distribution_type,norm_data)
```

```
## [1] 21.87800 22.59300 24.46863 17.50718 18.10413 20.81067
## [1] "Normal distribution parameters - Mean value is 19.9910723775174"
## [1] "Standard deviation is 2.00178609383717"
```

```r
# Exponential distribution
exp_data <- rexp(100000, 5)
distribution_type = 'Exponential'
Parameter_finder(distribution_type,exp_data)
```

```
## [1] 0.14212729 0.10511083 0.01109111 0.10576077 0.33417595 0.43495792
## [1] "Exponential distribution parameters - rate value is 5.02496026245548"
```

```r
# Gamma distribution
gamma_data <- rgamma(10000, 2, 3)
distribution_type = 'Gamma'
Parameter_finder(distribution_type,gamma_data)
```

```
## [1] 0.06970137 0.24377793 0.11733750 0.26149109 0.76327512 0.96985317
## [1] "Gamma distribution parameters - alpha value is1.98672506742025"
## [1] "Beta value is 0.335259177933098 and rate is 2.98276696305553"
```

```r
# Beta distribution
beta_data <- rbeta(10000, 2, 8)
distribution_type = 'Beta'
Parameter_finder(distribution_type,beta_data)
```

```
## [1] 0.24574431 0.09636654 0.12864885 0.05188635 0.07096758 0.12087265
## [1] "Beta distribution parameters - alpha value is2.02712084406855"
## [1] "Beta value is 8.15829719857456"
```

```r
# Student T distribution
tdist_data = rt(100000, 6)
distribution_type = 'tdist'
Parameter_finder(distribution_type,tdist_data)
```

```
## [1]  1.6150679  0.4516914 -0.1660867 -1.8020881 -0.9687962  0.8854000
## [1] "t distribution distribution parameters -v is 5.89084014842776"
```

```r
# CHI-SQUARE dustribution
chi2_data <- rchisq(10000, 5)
distribution_type = 'Chi Square'
Parameter_finder(distribution_type,chi2_data)
```

```
## [1] 2.051394 1.101921 6.353598 3.864569 3.322056 2.678814
## [1] "Chi Square distribution parameters - p value is 4.98725135472518"
```

```r
#Multinomial Distribution
 p = c(0.20,0.40,0.05,0.30)
 data = rmultinom(1000,size=4,p)
 Parameter_finder("multinomial", data)
```

```
## $n
## [1] 4.200003
##
## $p
## [1] 0.15873016 0.42595736 0.04167325 0.32601953
```

```r
# Multi variate normal distribution
Sum = matrix(c(7,3,3,7),2,2)
data = mvrnorm(n = 100000, rep(0, 2), Sum)
distribution_type = 'multivariatenormal'
Parameter_finder(distribution_type, data)
```

```
##            [,1]       [,2]
## [1,] -0.5195364  0.5640343
## [2,]  1.7338070 -0.3226992
## [3,]  2.2374753  5.9403655
## [4,] -1.8997285 -0.9350794
## [5,]  1.9784065  1.3551868
## [6,]  3.5042940  2.0507666
## $mu_val
## [1] -9.143227e-05 -2.677354e-03
##
## $summation
##          [,1]     [,2]
## [1,] 6.984144 2.961936
## [2,] 2.961936 6.957802
```

```r
# Hyper Geometric Distribution
k <- 3
data <- rhyper(40000, 10, 15, 3)
distribution_type = 'hypergeometric'
Parameter_finder(distribution_type, data, k)
```

```
## [1] 0 1 1 2 0 2
## [1] "N value is 40000 Subclass 1(m) value is 9.54858929125295 Subclass 2(n) value is 14.1652854774789
```

Extra Credit - Hyper Geometric distribution

The hypergeometric distribution is used for sampling without replacement.

p(x) = choose(m, x) choose(n, k-x) / choose(m+n, k)

for x = 0, …, k.

Note that p(x) is non-zero only for max(0, k-n) <= x <= min(k, m).

E[X] = μ = k p

and variance

Var(X) = k p (1 - p) * (m+n-k)/(m+n-1)

N → sample size of distribution

m → subclass -1 variable

n → subclass -2 variable

k → sample of element drawn

p → probability.

Proof:

$$p(x=i) = \frac{\binom{m}{i}\binom{n}{k-i}}{\binom{m+n}{k}}$$

The i^{th} selection has an equal likelihood of being in any trial, so the fraction of acceptable selection p is

$$p = \frac{m}{m+n} \quad \text{i.e.,} \quad P(x_i = 1) = \frac{m}{n+m}$$

$$E(x) = \mu = \left(\sum_{i=1}^{k} x_i\right) = \sum_{i=1}^{k} \langle x_i \rangle$$

$$= \sum_{i=1}^{k} \frac{m}{m+n} = \frac{km}{m+n} = k*p$$

$$var(x) = \sum_{i=1}^{k} var(x_i) + \sum_{i=1}^{k} \sum_{j=1}^{k} cov(x_i, x_j).$$

$\because$ x_i is a Bernoulli variable,

$$var(x_i) = p(1-p)$$

$$= \frac{m}{m+n}\left(1 - \frac{m}{m+n}\right)$$

$$= \frac{m}{m+n}\left(\frac{m+n-m}{m+n}\right) = \frac{nm}{(m+n)^2}$$

so, $\sum_{i=1}^{k} var(x_i) = \frac{knm}{(m+n)^2}$

for $i<j$, the covariance is

$$\text{cov}(x_i, x_j) = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle.$$

The probability that both i, j are succesful for $i \neq j$ is

$$P(x_i = 1, x_j = 1) = P(x_i = 1) * P(x_j = 1 \mid x_i = 1)$$

$$= \frac{m}{m+n} * \frac{m-1}{m+n-1}$$

$$= \frac{m(m-1)}{(m+n)(m+n-1)}$$

$\because$ x_i, x_j are Bernoulli variables, their product is also a Bernoulli variable.

$$\langle x_i, x_j \rangle = P(x_i = 1, x_j = 1) = \frac{m(m-1)}{(m+n)(m+n-1)}$$

$$\langle x_i \rangle \langle x_j \rangle = \frac{m}{m+n} * \frac{m}{m+n} = \frac{m^2}{(m+n)^2}$$

$$\therefore \operatorname{cor}(x_i, x_j) = \frac{m(m-1)}{(m+n)(m+n-1)} - \frac{m^2}{(m+n)^2}$$

$$= \frac{(m+n)(m^2-m) - m^2(m+n-1)}{(m+n)^2(m+n-1)}$$

$$= \frac{\cancel{m^2(m+n)} - \cancel{m^2} - mn - \cancel{m^2(m+n)} + \cancel{m^2}}{(m+n)^2(m+n-1)}$$

$$= \frac{-mn}{(m+n)^2(m+n-1)}$$

$$\sum_{i=1}^{K} \sum_{j=1}^{K} \operatorname{cor}(x_i, x_j) = \frac{-k(k-1)mn}{(m+n)^2(n+m-1)}$$

$$\text{var}(x) = \frac{kmn}{(m+n)^2} - \frac{k(k-1)mn}{(n+m)^2(n+m-1)}$$

$$= \frac{kmn(m+n-k)}{(n+m)^2(n+m-1)}$$

$$= kp(1-p)\left(\frac{m+n-k}{n+m-1}\right)$$

$$\because \; p = \frac{m}{n+m} \qquad 1-p = \frac{n}{n+m}$$

In the given problem, we need to find N, m(or)n given a distribution & k value.

$N \rightarrow$ length of distribution.

$P = \mu / k$ so we can find p here.

substituting p in variance we get m+n

$$m+n = \frac{k - \left(\frac{\text{var}(x)}{k * p(1-p)}\right)}{1 - \left(\frac{\text{var}(x)}{k * p(1-p)}\right)}$$

we know,

$$P = \frac{m}{n+n} \Rightarrow m = P * (m+n)$$

Appendix - Derivations

DERIVATIONS:

Estimating the Method of Moments Estimators for different Distributions.

① Binomial Distribution :-

We know that for Binomial (n, p)

$$P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $x = 1, \ldots, n$.

$$\text{Mean} = np$$

$$\text{Variance} = np(1-p)$$

Sample mean [S.Mean] $= \frac{1}{N}\sum_{i=1}^{N} x(i)$, Sample Variance [S.variance] $= \frac{1}{N}\sum_{i=1}^{N}(x(i) - \bar{x})^2$

So,

$$np = \frac{1}{N}\sum_{i=1}^{N} x(i) \quad \text{— ①}$$

$$np(1-p) = \frac{1}{N}\sum_{i=1}^{N}(x(i) - \bar{x})^2 \quad \text{— ②}$$

② divided by ①

$$\frac{②}{①} = \frac{np(1-p)}{np} \Rightarrow 1 - p = \frac{\sum_{i=1}^{N}(x(i) - \bar{x})^2}{\sum_{i=1}^{N} x(i)}$$

$$\hat{p} = 1 - \frac{\sum_{i=1}^{N}(x(i) - \bar{x})^2}{\sum x(i)} = \left[1 - \frac{\text{S.Variance}}{\text{S.mean}}\right]$$

therefore, Binomial distribution parameters -

p value is $1 - \frac{\text{Var\_data}}{\text{mean\_data}}$ [In terms of R code Implementation].

$$1 - \frac{\text{S.Variance}}{\text{S.Mean}}$$

$$n = \frac{\text{S.Mean}}{1 - \frac{\text{S.Variance}}{\text{S.mean}}} = \frac{\text{S.mean}^2}{\text{S.Variance} - \text{S.Mean}}.$$

Scanned with CamScanner

② **Bernoulli Distribution :-**

we know that for Bernoulli (p)

Mean $= P$

Variance $= P(1-p)$

probability density function is given as

$p^x (1-p)^{1-x} \quad x \in \{0,1\}$.

clearly, Method of moments Estimator of P is Mean 'M'.

Sample Mean $= \binom{n}{k} p^k (1-p)^{n-k}$

Therefore, Bernoulli distribution parameters. P value is Mean_data. [i.e Sample Mean]

③ **Geometric Distribution (P) :-**

Here, Mean $= \frac{1}{P}$

Variance $= \frac{(1-P)}{P^2}$, $P_X(x) = p(1-p)^{x-1}$

clearly, Method of Moment Estimator P is $\frac{1}{\text{Mean}}$. $\left[\because \text{Mean} = \frac{1}{P} \quad P = \frac{1}{\text{Mean}} \right]$

④ **Normal Distribution (μ, σ^2) :-**

Here, Mean $= \mu$

Variance $= \sigma^2$

Probability density function $= f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

Sample Mean $= \frac{1}{N} \sum_{i=1}^{N} x_i$.

Scanned with CamScanner

Therefore, It is clear that
Method of Moment estimater P for Normal distribution is ~~Mean~~ Sample Mean.

⑤ Poisson distribution

$$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \lambda \geq 0$$

Mean = λ

Variance = λ.

Therefore, Method of Moment estimator 'λ' for Poisson distribution is Mean (i.e Sample mean).

⑥ Point mass at a :

Here, Mean = a

Variance = 0

$$\text{Sample mean} = \frac{1}{N}\sum_{i=1}^{N} x_i \Rightarrow a = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Therefore Method of Moment Estimator a is Sample Mean.

a

⑦ Uniform distribution.

Here, Mean = $\frac{a+b}{2}$

Variance = $\frac{(b-a)^2}{12}$.

$E[x] = \frac{a+b}{2}$

$$E[x^2] = \int_a^b x^2 f(x)\,dx = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx$$

$$= \frac{a^2+b^2+ab}{3}$$

Scanned with CamScanner

Sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample Second moment $\bar{x}_{(2)} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

$$\bar{x} = \frac{a+b}{2}, \quad \bar{x}_{(2)} = \frac{a^2+b^2+ab}{3}$$

$a+b = 2\bar{x}$

$a = 2\bar{x} - b \qquad \bar{x}_{(2)} = \frac{(2\bar{x}-b)^2 + b^2 + (2\bar{x}-b)b}{3}$

$$b = \bar{x} \pm \frac{1}{2}\sqrt{4\bar{x}^2 - 4 \cdot 4\bar{x}^2 + 4 \cdot 3\bar{x}_{(2)}}$$

$$b = \bar{x} + \sqrt{3(\bar{x}_{(2)} - \bar{x}^2)}$$

$$b = \bar{x} + \sqrt{3(\text{Variance})}$$

$$b = \text{Mean} + \sqrt{3(\text{Variance})}$$

$$= \bar{x} + \sqrt{3s^2}$$

$$a = 2\bar{x} - b = \bar{x} - \sqrt{3(\text{Variance})}$$

$$= \text{mean} - \sqrt{3(\text{Variance})}$$

$$= \bar{x} - \sqrt{3s^2}$$

Therefore, Method of Momement Estimators

$a = \text{Mean} - \sqrt{3(\text{Variance})}$, $b = \text{Mean} + \sqrt{3(\text{Variance})}$;

$a = \bar{x} - \sqrt{3s^2}$, $b = \bar{x} + \sqrt{3s^2}$.

Scanned with CamScanner

Normal distribution

1) First moment about origin is

$$E(x_i) = \mu$$

2) Second moment about Mean is

$$Var(x_i) = E\left((x_i - \mu)^2\right) = \sigma^2$$

Equating

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = E(x)$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

So $\hat{\mu} = \bar{x}$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

↳ verience of data

$$sd = \sqrt{\sigma^2}$$

So we diretly return the Mean and veriance of data as the estimated Mean and verience Perameters

Exponential distribution

It is already given in the question,

Mean of data = β

But R uses $1/\beta$ as the input.

So dividing Mean by one

Estimate $\hat{\beta} = \frac{1}{\bar{x}}$

Gamma distribution

As explaind in Mini lecture,

Mean is given as $\alpha\beta$

Variance $= \alpha\beta^2$

So $\beta = \frac{\sigma^2}{\bar{x}}$

$\alpha = \frac{\bar{x}}{\beta}$ or $\frac{(\bar{x})^2}{\sigma^2}$ $\left[\because \frac{\alpha^2\beta^2}{\alpha\beta^2} = \alpha\right]$

Beta distribution

For this,

Mean $= \alpha/(\alpha+\beta) = \bar{x}$

Var $= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = s^2$

Solving,

$$\frac{\alpha}{(\alpha+\beta)} = \bar{x}$$

$$\alpha+\beta = \frac{\alpha}{\bar{x}}$$

x

$$\beta = \frac{\alpha}{\bar{x}} - \alpha$$

$$= \frac{\alpha(1-\bar{x})}{\bar{x}}$$

Substituting this in s^2 formula which is second moment.

$$s^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$= \frac{\alpha\left(\frac{\alpha(1-\bar{x})}{\bar{x}}\right)}{\left(\alpha + \frac{\alpha(1-\bar{x})}{\bar{x}}\right)^2\left(\alpha + \frac{\alpha(1-\bar{x})}{\bar{x}} + 1\right)}$$

$$= \frac{\left(\alpha^2(1-\bar{x})\right)/\cancel{\bar{x}}}{\frac{(\cancel{\alpha\bar{x}} + \alpha - \cancel{\alpha\bar{x}})^2}{\bar{x}^2}\left(\frac{\cancel{\alpha\bar{x}} + \alpha - \cancel{\alpha\bar{x}} + \bar{x}}{\cancel{\bar{x}}}\right)}$$

$$= \frac{\cancel{\alpha^2}(1-\bar{x})\,\bar{x}^2}{(\cancel{\alpha^2})(\alpha+\bar{x})}$$

$(1-\bar{x})\;\bar{x}^2$

$$s^2 = \frac{(1-\bar{x})\,\bar{x}^2}{(\alpha+n)}$$

$$(\alpha+n)s^2 = (1-\bar{x})\,\bar{x}^2$$

$$\alpha s^2 = (1-x)\,\bar{x}^2 - n s^2$$

$$\alpha = \frac{(1-x)\,\bar{x}^2 - n s^2}{s^2}$$

$$= \left(\left(\frac{\bar{x}^2}{s^2}\right)(1-\bar{x}) - \bar{x}\right)$$

$$\boxed{\alpha = \bar{x}\left(\frac{\bar{x}}{s^2}(1-\bar{x}) - 1\right)}$$

using this we get β

$$\boxed{\beta = \frac{\alpha(1-\bar{x})}{\bar{x}}}$$

T - distribution

For this we've in book

$$s^2 = \frac{v}{v-2}$$

$$(v-2)s^2 = v$$

$$\frac{v-2}{v} = \frac{1}{s^2}$$

$$1 - \frac{2}{v} = \frac{1}{s^2}$$

$$\frac{2}{v} = 1 - \frac{1}{s^2} = \frac{s^2-1}{s^2}$$

$$v = \left(\frac{2s^2}{s^2-1}\right)$$

Chi-Square distribution

For Chi-Square, it is directly given

then Mean of data = P

So we directly return the Mean value

Multinomial distribution

Here, we calculate the values for each row

For each row, Mean is given by np

$$s^2 = np(1-p)$$

$$np \to \text{mean}$$

$$\frac{s^2}{\bar{x}} = 1-p$$

$$p_i = \left(1 - \frac{s_i^2}{\bar{x}_i}\right)$$

n can be deduced from this

$$n = \Sigma\left(\frac{\bar{x}}{p}\right)$$

Multivariate normal

For this we just find Mean and Variance of each row by using colMeans and summation