w04

Yekta Amirkhalili

May 29th, 2023

Week 4 (Session 3) - May 29, 2023

material to cover:

0. …**Probability Distributions**…
All code references: https://cran.r-project.org/web/packages/Rlab/index.html (https://cran.r-project.org/web/packages/Rlab/index.html)
`r library("Rlab")`
`## Rlab 4.0 attached.`
`## ## Attaching package: 'Rlab'`

There are different types of random variables in Statistics. Here are some of the most well-known *Discrete* Random Variables (RV).

1. Bernouli Bernoulli RV takes in values of "Success" and "Failure" which can be represented mathematically as 1 and 0 respectively. Let's define probability of success as p, and following that probability of failure would be $1 - p$. The probability function would then be:

$$P(X = x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases}$$

```
set.seed(111)
N <- 10000

random_vars_bern <- rbern(N, prob = 0.8)
distro_bern <- dbern(random_vars_bern, 0.8, log = FALSE)
cdf_bern <- pbern(random_vars_bern, 0.8)

first_20 <- random_vars_bern[1:20]
print('The first 20 Bernoulli trials: ')
```

```
## [1] "The first 20 Bernoulli trials: "
```

```
print(first_20)
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1
```

```
print('The distribution')
```

```
## [1] "The distribution"
```

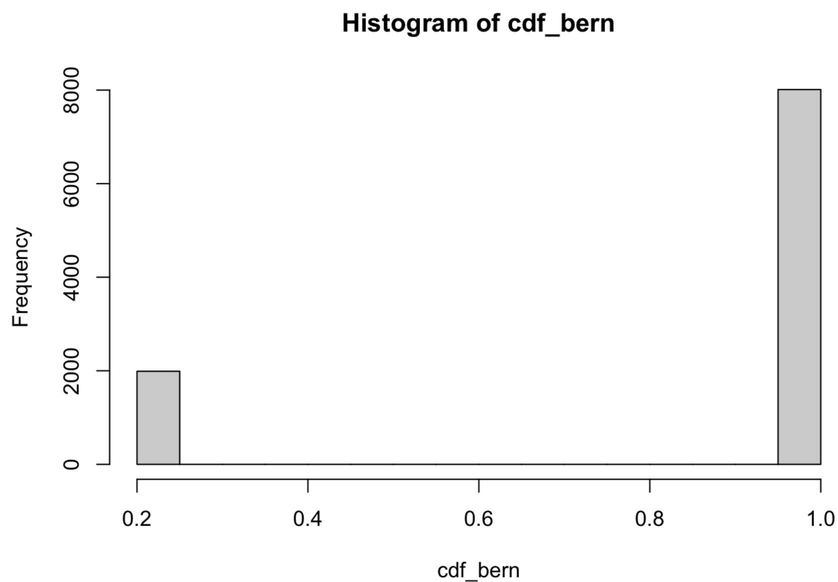
```
print(distro_bern[1:20])
```

```
##  [1] 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.2 0.8
## [20] 0.8
```

```
print('The CDF')
```

```
## [1] "The CDF"
```

```
hist(cdf_bern)
```



2. Binomial The number of "Success"es in n trials of independent Bernoulli experiments with success probability of p follows a Binomial Distribution. The probability function is:

$$P(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}$$

Where $i = 0, 1, 2, \ldots, n.$

```
set.seed(112)
N <- 10000

random_vars_binom <- rbinom(N, size = 10, prob = 0.7)
distro_binom <- dbinom(random_vars_binom, size = 10, prob = 0.7)
cdf_binom <- pbinom(random_vars_binom, size = 10, prob = 0.7)

first_20 <- random_vars_binom[1:20]
print('The first 20 Binomial: ')
```

```
## [1] "The first 20 Binomial: "
```

```
print(first_20)
```

```
##  [1] 8 5 3 4 8 9 9 6 9 7 8 8 5 7 6 4 7 7 6 8
```

```
print('The distribution')
```

```
## [1] "The distribution"
```

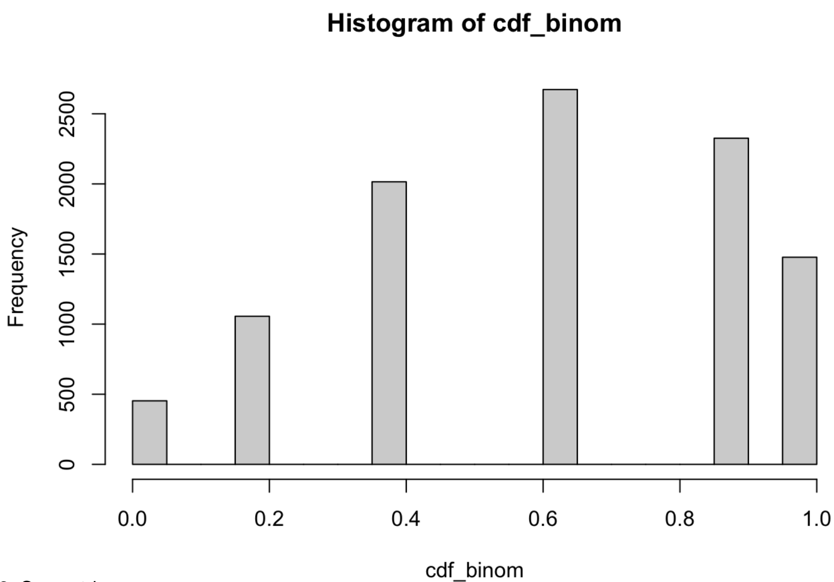
```
print(distro_binom[1:20])
```

```
##  [1] 0.233474440 0.102919345 0.009001692 0.036756909 0.233474440 0.121060821
##  [7] 0.121060821 0.200120949 0.121060821 0.266827932 0.233474440 0.233474440
## [13] 0.102919345 0.266827932 0.200120949 0.036756909 0.266827932 0.266827932
## [19] 0.200120949 0.233474440
```

```
print('The CDF')
```

```
## [1] "The CDF"
```

```
hist(cdf_binom)
```



3. Geometric

# of independent Bernoulli trials (with parameter p) to reach *the first* success follows a Geometric distribution. The probability function is:

$$P(X = i) = (1 - p)^{i-1} p \qquad i = 1, 2, 3, \ldots$$

```
set.seed(113)
N <- 10000

random_vars_geom <- rgeom(N, prob = 0.65)
distro_geom <- dgeom(random_vars_geom, prob = 0.65)
cdf_geom <- pgeom(random_vars_geom, prob = 0.65)

first_20 <- random_vars_geom[1:20]
print('The first 20 Geometric: ')
```

```
## [1] "The first 20 Geometric: "
```

```
print(first_20)
```

```
##  [1] 0 0 2 1 1 2 2 0 0 1 0 0 0 1 0 0 0 4 0 0
```

```
print('The distribution')
```

```
## [1] "The distribution"
```

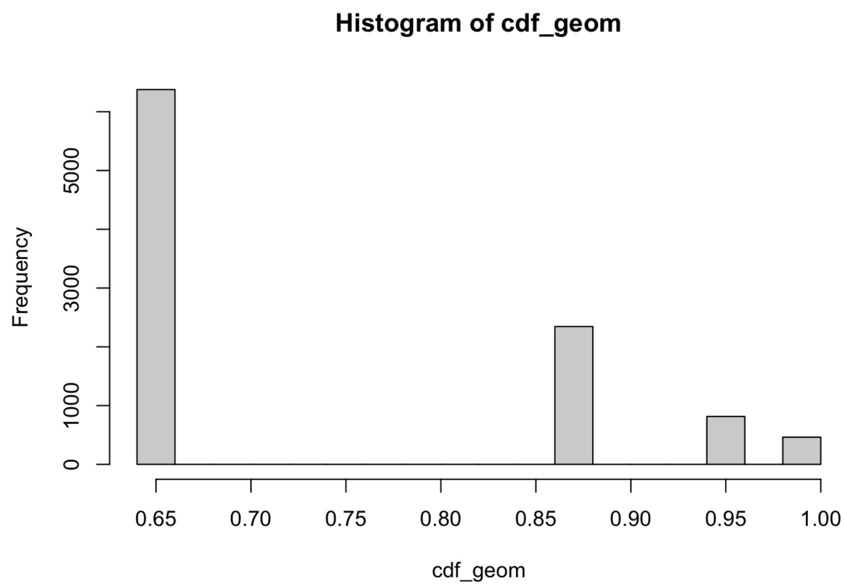
```
print(distro_geom[1:20])
```

```
##  [1] 0.650000000 0.650000000 0.079625000 0.227500000 0.227500000 0.079625000
##  [7] 0.079625000 0.650000000 0.650000000 0.227500000 0.650000000 0.650000000
## [13] 0.650000000 0.227500000 0.650000000 0.650000000 0.650000000 0.009754062
## [19] 0.650000000 0.650000000
```

```
print('The CDF')
```

```
## [1] "The CDF"
```

```
hist(cdf_geom)
```



4. Negative Binomial The number of independent Benoulli trials (with parameter p) to reach the rth success follows a Negative Binomial distribution. The probability function is:

$$P(X = i) = \binom{i-1}{r-1} p^r (1-p)^{i-r} \qquad i = r, r+1, r+2, \ldots$$

```
set.seed(114)
N <- 10000

random_vars_nb <- rnbinom(N, size = 9, prob = 0.7)
distro_nb <- dnbinom(random_vars_nb, size = 9, prob = 0.7)
cdf_nb <- pnbinom(random_vars_nb, size = 9, prob = 0.7)


first_20 <- random_vars_nb[1:20]
print('The first 20 Negative Binomial: ')
```

```
## [1] "The first 20 Negative Binomial: "
```

```
print(first_20)
```

```
##  [1] 6 6 7 1 5 3 4 2 4 4 1 1 2 2 4 0 2 4 3 2
```

```
print('The distribution')
```

```
## [1] "The distribution"
```

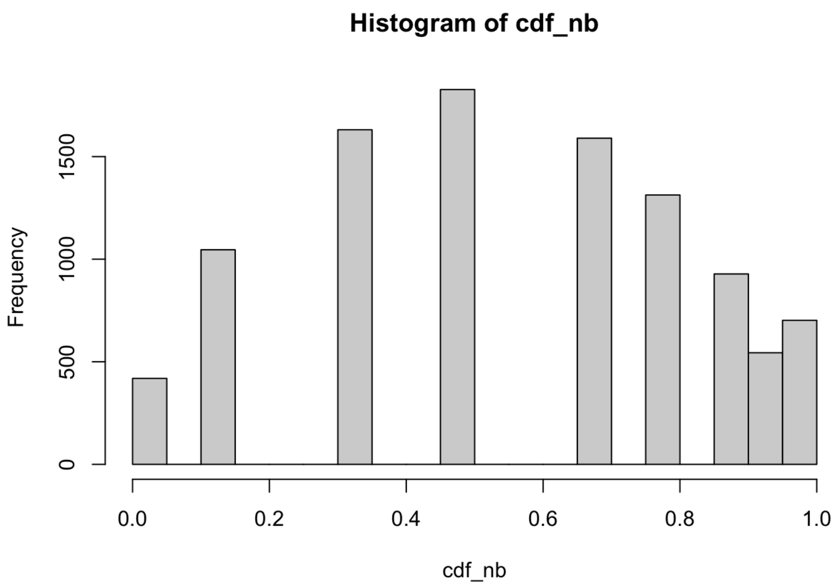
```
print(distro_nb[1:20])
```

```
##  [1] 0.08834159 0.08834159 0.05679102 0.10895474 0.12620227 0.17977532
##  [7] 0.16179779 0.16343211 0.16179779 0.16179779 0.10895474 0.10895474
## [13] 0.16343211 0.16343211 0.16179779 0.04035361 0.16343211 0.16179779
## [19] 0.17977532 0.16343211
```

```
print('The CDF')
```

```
## [1] "The CDF"
```

```
hist(cdf_nb)
```



5. Poisson It is basically an approximation on Binomial Distribution when n is a large number ($n \geq 20$) and the probability of success is small. The probability function is:

$$P(X = i) = \frac{e^{-\lambda}\lambda^{i}}{i!}$$

```
set.seed(115)
N <- 100000

rv_pos <- rpois(N, 10)
dis_pos <- dpois(rv_pos, 10)
cdf_pos <- ppois(rv_pos, 10)

first_20 <- rv_pos[1:20]
print('The first 20 Poisson: ')
```

```
## [1] "The first 20 Poisson: "
```

```
print(first_20)
```

```
##  [1] 11 11 11  9 10 12 10  9 10  8  9 10  7  7 11 14  7  8 10  5
```

```
print('The distribution')
```

```
## [1] "The distribution"
```

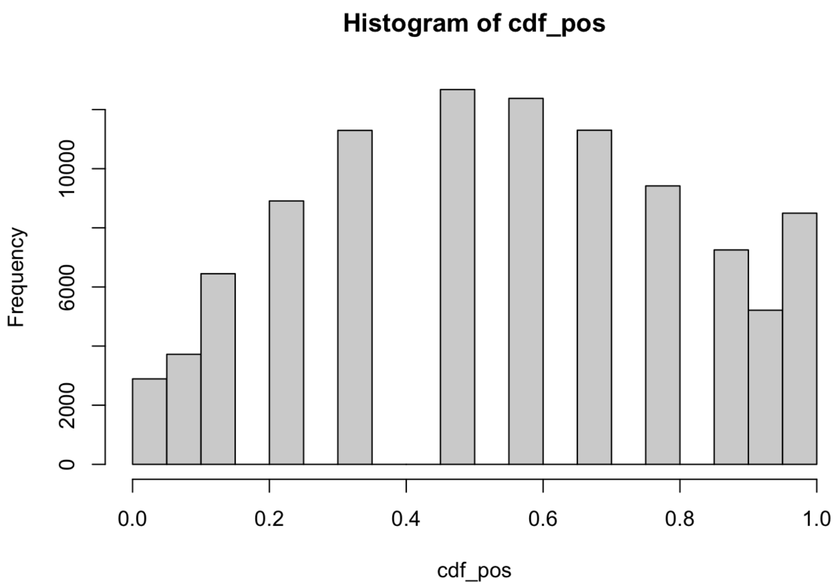
```
print(dis_pos[1:20])
```

```
##  [1] 0.11373640 0.11373640 0.11373640 0.12511004 0.12511004 0.09478033
##  [7] 0.12511004 0.12511004 0.12511004 0.11259903 0.12511004 0.12511004
## [13] 0.09007923 0.09007923 0.11373640 0.05207710 0.09007923 0.11259903
## [19] 0.12511004 0.03783327
```

```
print('The CDF')
```

```
## [1] "The CDF"
```

```
hist(cdf_pos)
```



6. Hyper Geometric Suppose there exists a box with m defective and $N - m$ working parts. We randomly take out n parts without replacement. If X represents the number of defective parts taken out, we have:

$$P(X = i) = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$$

```
set.seed(116)
N <- 10000
m <- 480
n <- N - m

rv_hyper <- rhyper(N, m, n, 5)
dis_hyper <- dhyper(rv_hyper, m, n, 5)
cdf_hyper <- phyper(rv_hyper, m, n, 5)

first_20 <- rv_hyper[1:20]
print('The first 20 Hyper Geometric: ')
```

```
## [1] "The first 20 Hyper Geometric: "
```

```
print(first_20)
```

```
##  [1] 0 0 0 0 2 0 1 0 0 0 0 1 0 0 1 0 0 0 0 2
```

```
print('The distribution')
```

```
## [1] "The distribution"
```

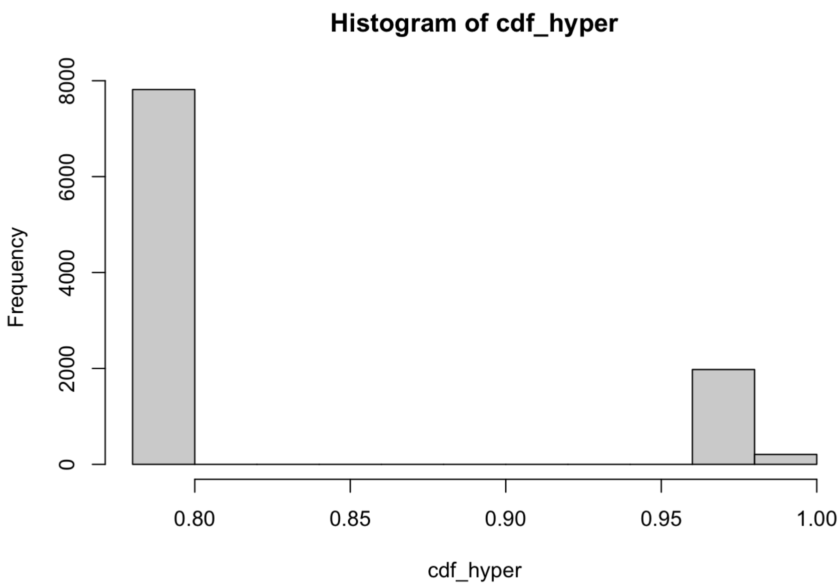
```
print(dis_hyper[1:20])
```

```
##  [1] 0.78192093 0.78192093 0.78192093 0.78192093 0.01985112 0.78192093
##  [7] 0.19720578 0.78192093 0.78192093 0.78192093 0.78192093 0.19720578
## [13] 0.78192093 0.78192093 0.19720578 0.78192093 0.78192093 0.78192093
## [19] 0.78192093 0.01985112
```

```
print('The CDF')
```

```
## [1] "The CDF"
```

```
hist(cdf_hyper)
```



7. Uniform If a RV X takes the values of $x_1, x_2, x_3, \ldots, x_n$ with equal chance of $\frac{1}{n}$, then X has a uniform distribution with the following probability function:

$$P(X = x_i) = \frac{1}{n}$$

```
N <- 100

rv_unif <- rep(0.01, 100) #repeat 1/100 100 times
dis_unif <- function(nums){
    u <- 1/nums
    u_ <- rep(u, nums)

    return(u_)
} #number of values : 1/nums = probability

dis_uniform <- dis_unif(N)

cdf_unif <- function(pr, l){
    cdf <- c()
    le <- length(l)

    for(i in le:1){
        cdf <- c(cdf, sum(l[1:i]))
    }

    return(rev(cdf))
}

cdf_uniform <- cdf_unif((1/N), rv_unif)

first_20 <- rv_unif[1:20]
print('The first 20 Uniform: ')
```

```
## [1] "The first 20 Uniform: "
```

```
print(first_20)
```

```
##  [1] 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01
## [16] 0.01 0.01 0.01 0.01 0.01
```

```
print('The distribution')
```

```
## [1] "The distribution"
```

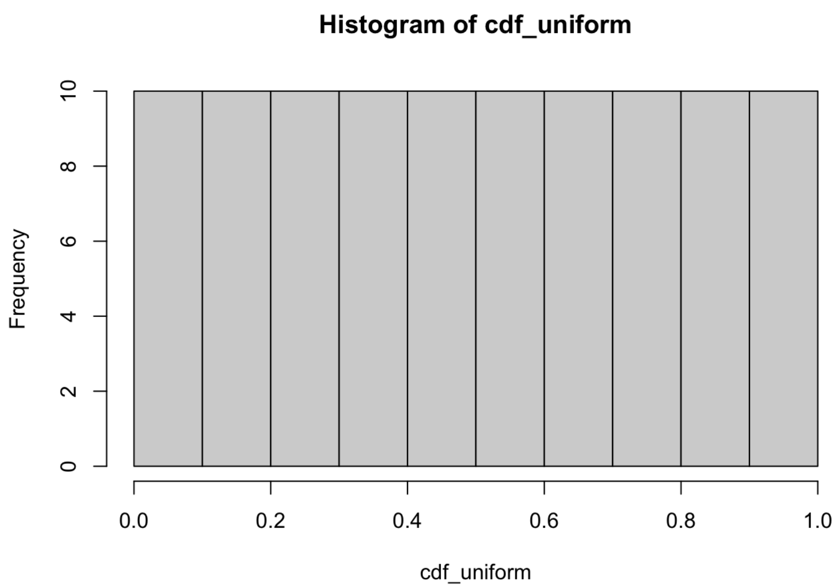
```
print(dis_uniform[1:20])
```

```
##  [1] 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01
## [16] 0.01 0.01 0.01 0.01 0.01
```

```
print('The CDF')
```

```
## [1] "The CDF"
```

```
hist(cdf_uniform)
```



More: Read up on packages that do this for you!

Here are some of the most well-known *Continous* Random Variables (RV).

1. Continuous Uniform The RV X has a Uniform distribution in interval (a, b) if the probability density function is:

$$f(x) = k = \frac{1}{b - a}, \qquad a \leq x \leq b$$

```
set.seed(100)
N <- 1000

uni <- runif(N, min = 0, max = 100)
uni_dist <- dunif(uni, min = 0, max = 100, log = FALSE)
uni_cdf <- punif(uni, min = 0, max = 100)


first_20 <- uni[1:20]
print('The first 20 Uniform: ')
```

```
## [1] "The first 20 Uniform: "
```

```
print(first_20)
```

```
##  [1] 30.776611 25.767250 55.232243  5.638315 46.854928 48.377074 81.240262
##  [8] 37.032054 54.655860 17.026205 62.499648 88.216552 28.035384 39.848790
## [15] 76.255108 66.902171 20.461216 35.752485 35.947511 69.029053
```

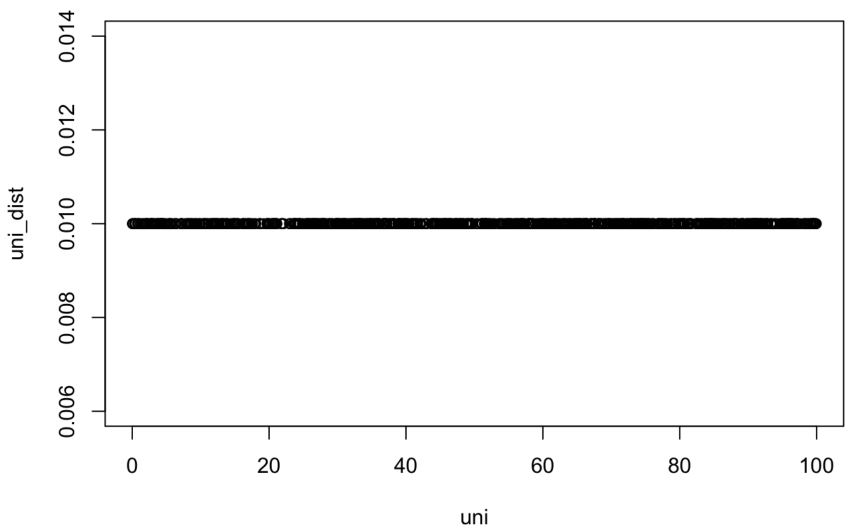
```
print('The distribution')
```

```
## [1] "The distribution"
```

```
print(uni_dist[1:20])
```

```
##  [1] 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01
## [16] 0.01 0.01 0.01 0.01 0.01
```

```
plot(uni, uni_dist)
```



2. Normal It was first introduced as an approximation of Binomial (with parameters n and $p = 0.5$ where n was large and p was not too large and not too small). The probability density function is as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < +\infty$$

```
set.seed(101)
N <- 1000

x <- rnorm(N, 0, 1)
y <- dnorm(x, mean = 0, sd = 1)

norm_cdf <- pnorm(x, 0, 1)

first_20 <- x[1:20]
print('The first 20 Normal: ')
```

```
## [1] "The first 20 Normal: "
```

```
print(first_20)
```

```
##  [1] -0.3260365  0.5524619 -0.6749438  0.2143595  0.3107692  1.1739663
##  [7]  0.6187899 -0.1127343  0.9170283 -0.2232594  0.5264481 -0.7948444
## [13]  1.4277555 -1.4668197 -0.2366834 -0.1933380 -0.8497547  0.0584655
## [19] -0.8176704 -2.0503078
```

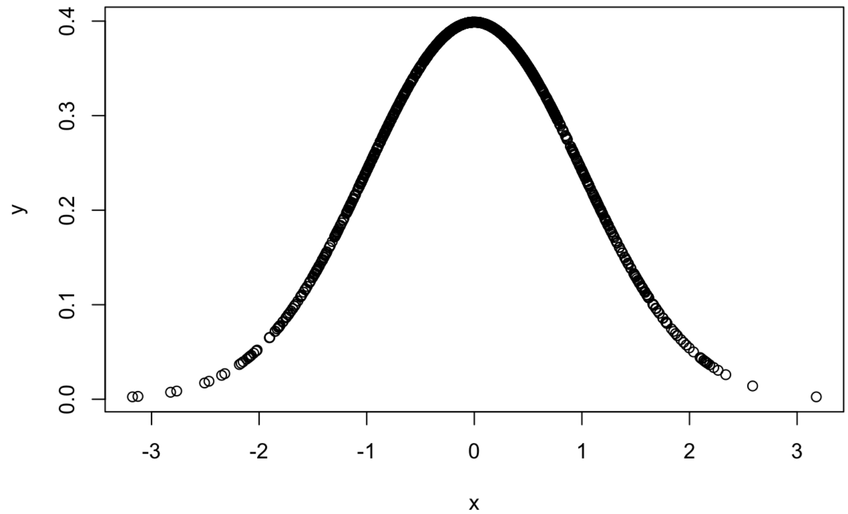
```
print('The distribution')
```

```
## [1] "The distribution"
```

```
print(y[1:20])
```

```
##  [1] 0.37829218 0.34247878 0.31767923 0.38988108 0.38013559 0.20028039
##  [7] 0.32943080 0.39641523 0.26200047 0.38912257 0.34731875 0.29088497
## [13] 0.14396552 0.13605194 0.38792314 0.39155538 0.27804284 0.39826103
## [19] 0.28558061 0.04876124
```

```
plot(x, y)
```



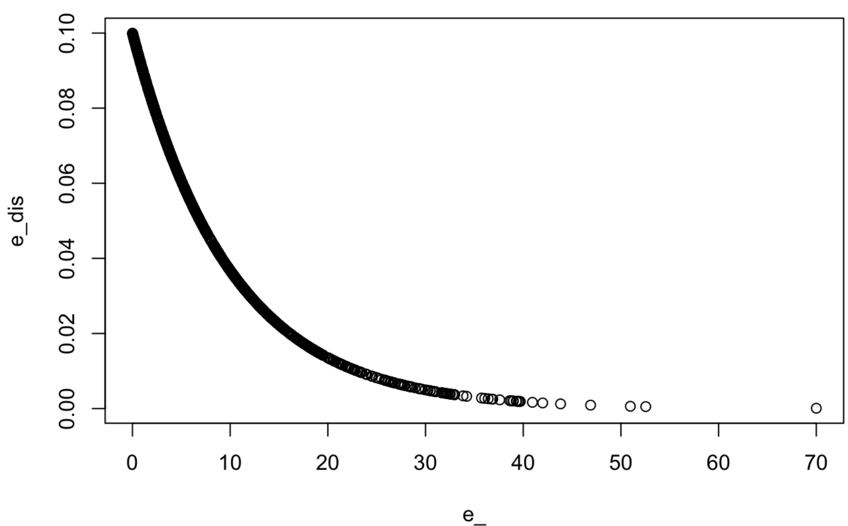
3. Normal Approximation to the Binomial If X is a Binomial RV with parameters n and p, and n is large while p is neither large nor small, then we can approximate this distribution with new parameters: $\mu = np$ and $\sigma^2 = np(1-p)$ and that would be a normal distribution.
4. Exponential RV X follows the exponential distribution with parameter $\lambda > 0$ when it has the following probability density function:

$$f(x) = \lambda e^{-\lambda x} \quad x \geq 0$$

```
set.seed(105)
N <- 1000

e_ <- rexp(N, rate = 0.1)
e_dis <- dexp(e_, rate = 0.1)
e_cdf <- pexp(e_, rate = 0.1)

plot(e_, e_dis)
```



5. Gamma RV X has a gamma distribution with parameters $\alpha > 0$ and $\lambda > 0$ if the probability density function is as follows:

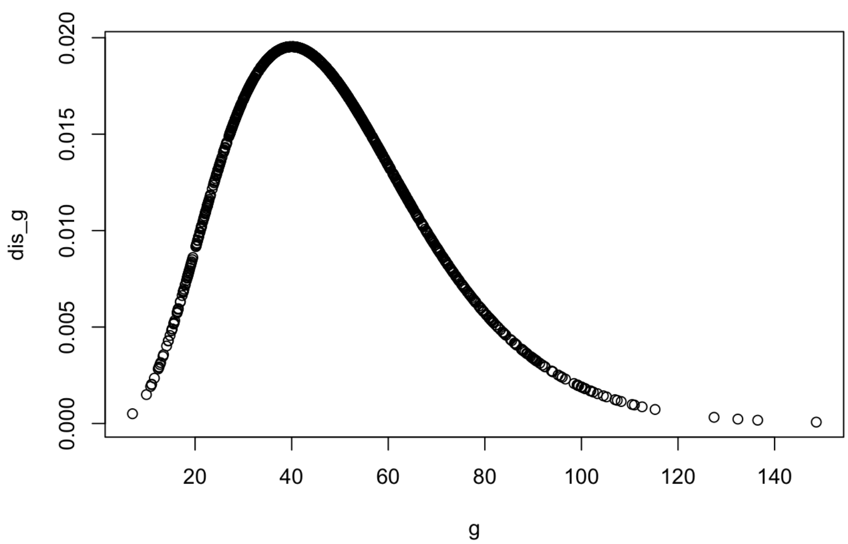
$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \quad x \geq 0$$

Read more about Gamma function: https://www.statlect.com/mathematical-tools/gamma-function (https://www.statlect.com/mathematical-tools/gamma-function)

```
set.seed(106)
N <- 1000

#shape is alpha, rate is lambda and scale is 1/lambda
g <- rgamma(N, shape = 5, rate = 0.1)
dis_g <- dgamma(g, shape = 5, rate = 0.1)

plot(g, dis_g)
```



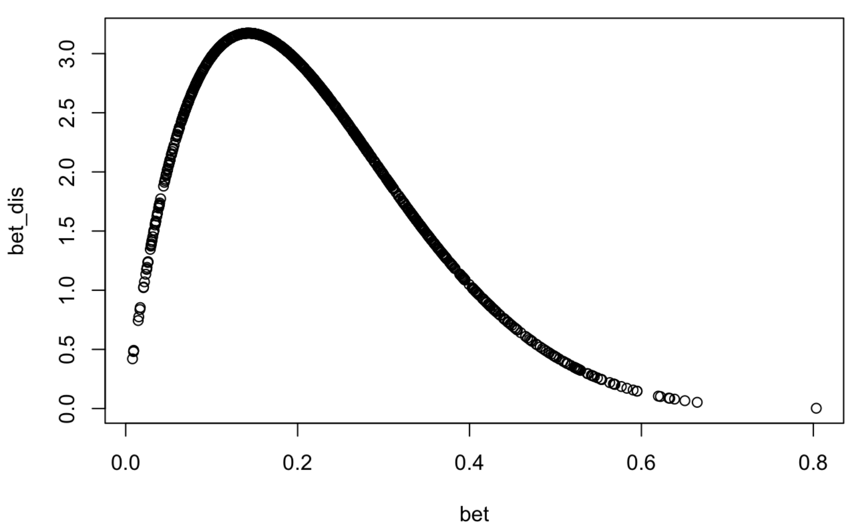
6. Beta RV X has a beta distribution with positive parameters a and b if the probability density function is:

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} \qquad 0 < x < 1$$

```
set.seed(107)
N <- 1000

bet <- rbeta(N, shape1 = 2, shape2 = 7)
bet_dis <- dbeta(bet, shape1 = 2, shape2 = 7)

plot(bet, bet_dis)
```



Thought: What happens if $a = b = 1$?

7. Chi-sqaured

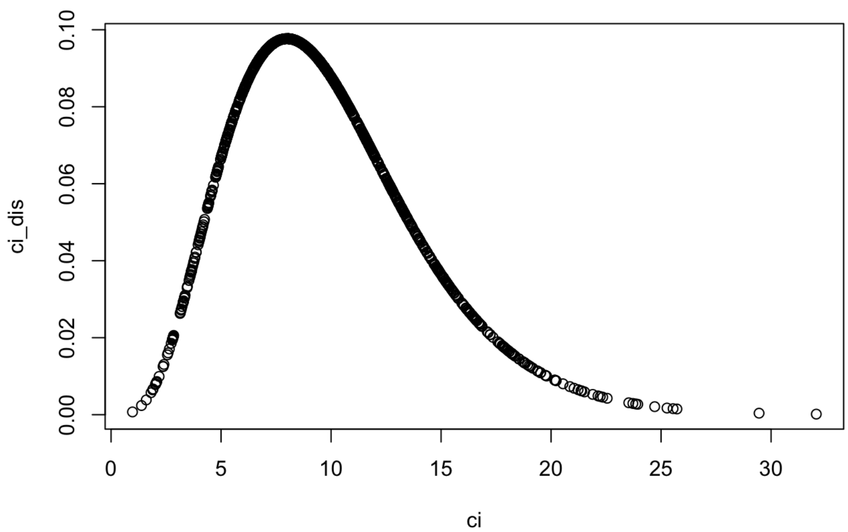
NOTEIf a random variable Z follows the Standard Normal distribution, then $Y = Z^2$ follows a gamma distribution with parameters $\lambda = 0.5$ and $\alpha = 0.5$.

If $Z_1, Z_2, \ldots, Z_n$ are independent random variables following Standard Normal Distribution, then the RV $Y = Z_1^2 + Z_2^2 + \ldots + Z_n^2$ follows a gamma distribution with parameters $\lambda = 0.5$ and $\alpha = \frac{n}{2}$. This special case of Gamma is called a Chi-squared distribution with n degrees of freedom, denoted by χ_n^2.

```
set.seed(108)
N <- 1000

ci <- rchisq(N, 10)
ci_dis <- dchisq(ci, 10)

plot(ci, ci_dis)
```



Additional: LogNormal Read: https://www.itl.nist.gov/div898/handbook/eda/section3/eda3669.htm (https://www.itl.nist.gov/div898/handbook/eda/section3/eda3669.htm)