Probability Distributions

Probability- Simulating Experiments

```
> x=sample(c("H","T"), 10, replace=T)# Tossing a coin 10 times
> x
 [1] "H" "T" "T" "T" "H" "H" "T" "T" "T" "T"
> table(x)
x
H T
3 7
> table(x)/10 # calculates the probabilities
x
  H   T
0.3 0.7
```

You can check and observe that the estimated probabilities approach the true value as the number of tosses increase.

```
> x=sample(c("H","T"), 1000, replace=T)# Tossing a coin 1000 times
> table(x)/1000 # calculates the probabilities
x
    H     T
0.486 0.514
```

Coin tossing experiment

```
sample.space <- c(0,1)
theta <- 0.5 # this is a fair coin
N <- 50 # we want to flip a coin 20 times

flips <- sample(sample.space,
                size = N,
                replace = TRUE,
                prob = c(theta, 1 - theta))
> flips
 [1] 0 1 1 1 0 0 1 0 0 1 1 1 1 1 0 1 0 1 0 1 1 0 0 1 1 0 0 0 0 1 1 0 0
[39] 0 1 1 1 0 0 1 0 1 1 1 0
>
```

Counting, Permutation and Combination

Factorial, permutation and combination are essentials to calculate the probability:

$$\begin{aligned}
\textbf{Factorial}: n! &= n \times (n-1) \times \cdots \times 3 \times 2 \times 1 \\
\textbf{Permutation}: P(n,k) &= \frac{n!}{(n-k)!} = n(n-1)(n-2)\cdots(n-(k-1)) \\
\textbf{Combination}: C(n,k) &= \frac{n!}{(n-k)!k!}
\end{aligned}$$

We can use the R code below to calculate the factorial, permutation and combination.

```
Factorial: factorial(n)
Permutation: prod(n:n-k+1)
Combination: choose(n,k)
```

Examples

```
> factorial(5)
[1] 120
> prod(10:(10-3+1))
[1] 720
> choose(10,3)
[1] 120
```

Birthday Problem

Suppose there are 25 randomly chosen people in a room. What is the probability two or more of them have the same birthday? Model simplifications for a simple combinatorial solution:

- Ignore leap years
- Assume all 365 days equally likely

Solution: Let X be the number of matches.

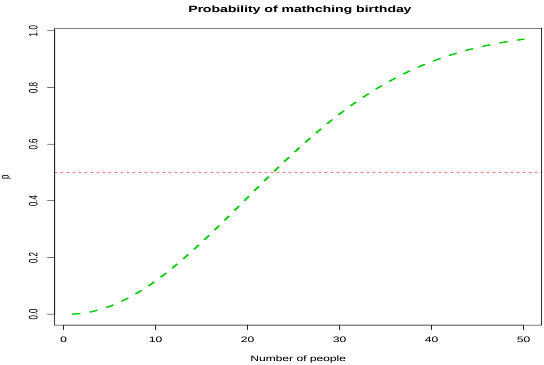
$$\begin{aligned} P(X \leq 1) &= 1 - P(X = 0) \\ &= 1 - \frac{365 \times 364 \times \cdots \times (365 - 24)}{365^{25}} \\ &= 1 - \prod_{i=0}^{24} \left(1 - \frac{i}{365}\right) \\ &= 1 - 0.4313 = 0.5687 \end{aligned}$$

In R

```
> 1-prod(1-(0:24)/365)
[1] 0.5686997
```

Birthday Problem

```
> p <- numeric(50)
> for (n in 1:50) {
+ q <- 1 - (0:(n - 1))/365
+ p[n] <- 1 - prod(q)}
> p
> plot(p, type="l", lty=2, lwd=3, col=3, xlab="Number of people",
 main=" Probability of mathching birthday")
```



Probability Distributions

R supports a large number of distributions. Usually, four types of functions are provided for each distribution:

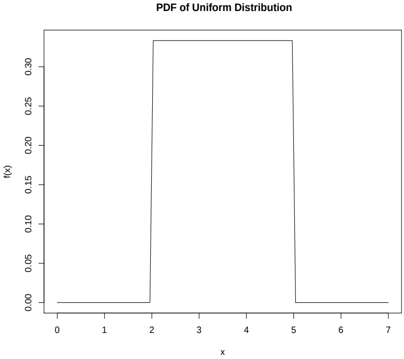
- d: density function
- p: cumulative distribution function, $P(X \leq x)$
- q: quantile function
- r: draw random numbers from the distribution

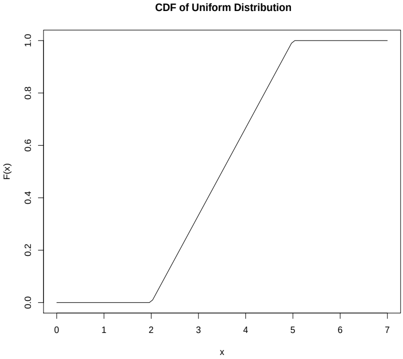
Uniform Distribution

If a random variable X has the probability density function

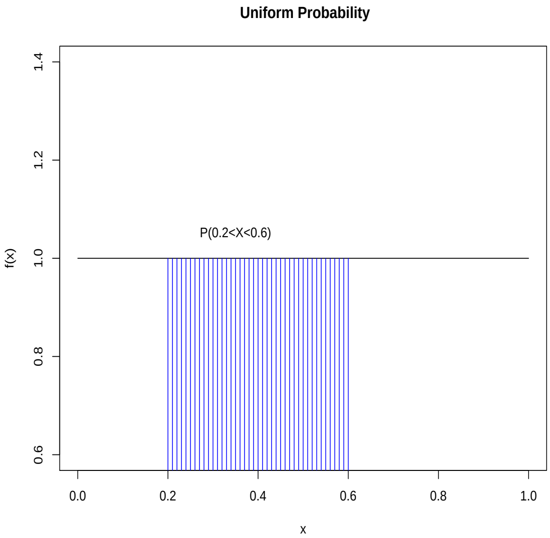
$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \alpha < x < \beta, \\ 0 & \text{otherwise,} \end{cases}$$

then it is said to have the uniform distribution on (α, β).





Uniform Distribution

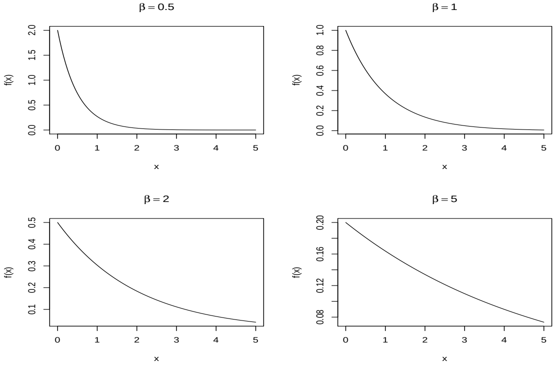


Exponential Distribution

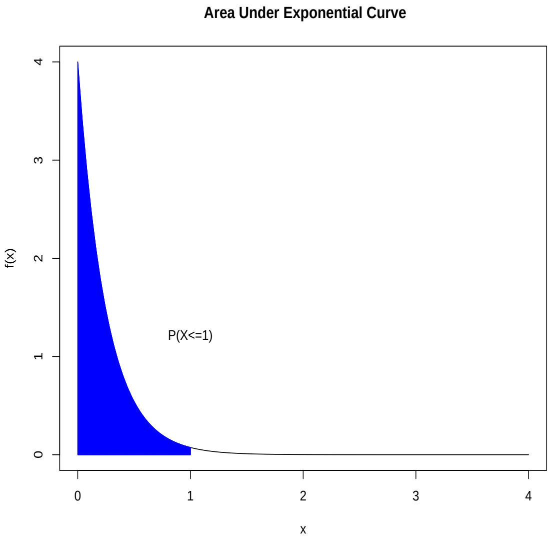
If a random variable X has the probability density function

$$f(x) = \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right), \qquad x \geq 0, \quad \beta > 0$$

then it is said to have the exponential distribution with parameter β.



Exponential Distribution



Empirical Distribution

The empirical cumulative distribution function F_n (written ECDF) is the probability distribution that places probability mass $1/n$ on each of the values $x_1, x_2, ..., x_n$. The empirical PMF takes the form

$$f_X(x) = \frac{1}{n}, \quad x \in \{x_1, x_2, \ldots, x_n\}$$

If the value x_i is repeated k times, the mass at x_i is accumulated to k/n.

Probability Distributions

The distributions supported include continuous distributions:

- *unif*: Uniform distribution
- *norm*: Normal distribution
- t: Student's t- distribution
- *chisq:* Chi-square distribution
- f: Fisher's F distribution
- *gamma:* Gamma distribution
- *exp:* Exponential distribution
- *beta:* Beta distribution
- *lnorm:* Log-normal distribution
- *cauchy:* Cauchy distribution
- *logis:* Logistic distribution
- *weibull:* Weibull distribution

Probability Distributions

The distributions supported include discrete distributions:

- *binom*: Binomial distribution
- *geom* : Geometric distribution
- *hyper:* Hypergeometric distribution
- *nbinom:* Negative Binomial distribution
- *pois:* Poisson distribution

Using help: ?runif

```
Uniform {stats} R Documentation
The Uniform Distribution
Description


dunif(x, min=0, max=1, log = FALSE)
punif(q, min=0, max=1, lower.tail = TRUE, log.p = FALSE)
qunif(p, min=0, max=1, lower.tail = TRUE, log.p = FALSE)
runif(n, min=0, max=1)

Arguments
x,q         vector of quantiles.
p             vector of probabilities.
n             number of observations. If length(n) > 1, the length
                is taken to be the number required.
min,max        lower and upper limits of the distribution.
log, log.p     logical;if TRUE, probabilities p are given as log(p).
lower.tail     logical;if TRUE (default),probabilities are P[X <= x],
                otherwise, P[X > x].
```

set.seed

Computer-generated random numbers are not really random; they are "pseudo-random." This means that the computer generates a fixed sequence of random-looking numbers. The sequence is very, very long, so the pseudo-random numbers are not repeated within any typical computation. Therefore when we generate a set of random numbers from any distribution every time we will get different set. But if we want to generate the same set of "random" numbers at later we need to use the **set.seed()** function. The **set.seed()** function puts the random number generator in a reproducible state. To avoid using the same random numbers we have to set the seed to different numbers each time.

```
> set.seed(20)
> rnorm(5)
[1]  1.1626853 -0.5859245  1.7854650 -1.3325937 -0.4465668
> set.seed(20)
> rnorm(5)
[1]  1.1626853 -0.5859245  1.7854650 -1.3325937 -0.4465668
```

Continuous Distributions

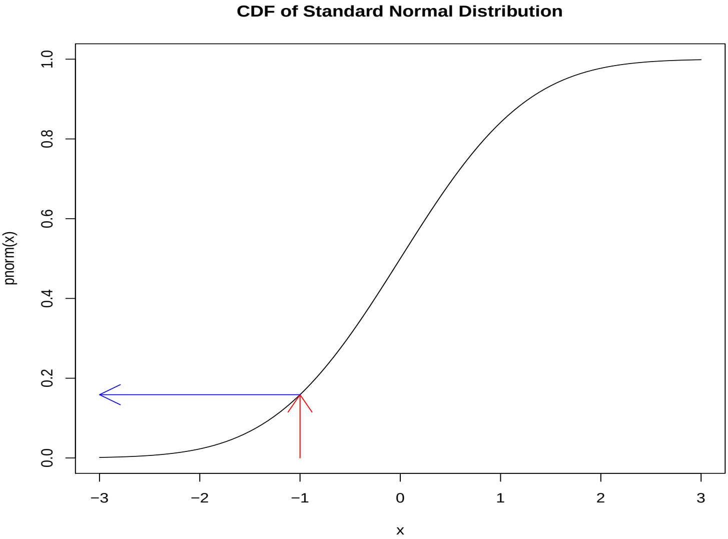
The cumulative probability function is a straightforward notion: it is an S-shaped curve showing, for any value of x, the probability of obtaining a sample value that is less than or equal to x. Here is what it looks like for the normal distribution:

```
>curve(pnorm(x),-3,3, main="CDF of Standard Normal Distribution")
>arrows(-1,0,-1,pnorm(-1),col="red")
>arrows(-1,pnorm(-1),-3,pnorm(-1),col="blue")
```

The value of $x(-1)$ leads up to the cumulative probability (red arrow) and the probability associated with obtaining a value of this size (-1) or smaller is on the y axis (blue arrow). The value on the y axis is 0.1586553:

```
>pnorm(-1)
[1] 0.1586553
```

CDF- Normal Distribution



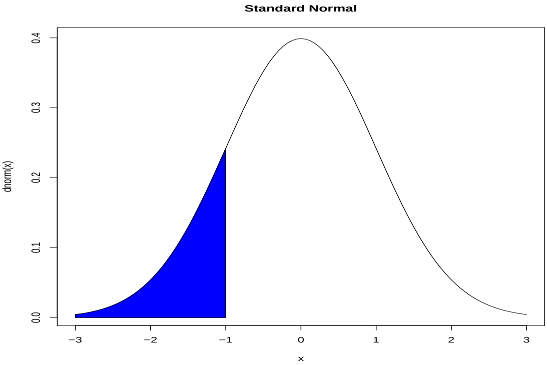
Continuous Distributions

For a continuous random variable X , the function $f(x)$ is said to be a probability density function if it satisfies the following properties:
a) $f(x) \geq 0$ for all values of x.
b) $P(a < x < b) = \int_a^b f(x)dx$ for all a and b.
c) $\int_{-\infty}^{\infty} f(x)dx = 1$.
For a normal distribution the slope starts out very shallow up to about x=-2, increases up to a peak (at x = 0 in this example) then gets shallower, and becomes very small indeed above about x=2. Here is what the density function of the normal (dnorm) looks like:

```
>curve(dnorm(x),-3,3, main="pdf of standard Normal Distribution")
```

```
>dnorm(-1)
[1] 0.2419707
```

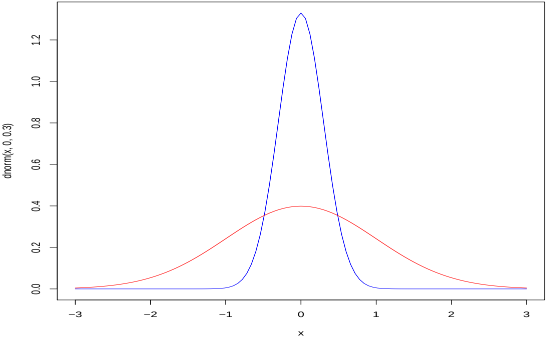
PDF- Normal Distribution



```
cord.x <- c(-3,seq(-3,-1,0.01),-1)
cord.y <- c(0,dnorm(seq(-3,-1,0.01)),0)
curve(dnorm(x),xlim=c(-3,3),main='Standard Normal')
polygon(cord.x,cord.y,col='blue')
```

PDF- Two Normal Distribution

Superimpose two PDFs:



```
curve(dnorm(x,0,0.3),from=-3,to=3,col="blue")
curve(dnorm(x,0,1),from=-3, to=3, col="red", add=T)
```

The "from" and "to" can be omitted.

Few Examples- Marking and shading

1) Plot the standard normal PDF and mark the 90th percentile:

```
curve(dnorm,-3,3)
lines(qnorm(0.9),dnorm(qnorm(0.9)),type="h", col="red")
```

2) Shade the area under the N(0,1) pdf to the right of the 90th percentile:

```
x1=seq(qnorm(0.9),3,0.01);
y1=dnorm(x1)
curve(dnorm,-3,3); lines(x1,y1,type="h",col="red")
```

3) More options:

```
curve(dnorm,-3,3)
polygon(c(rep(1,201),rev(seq(1,3,.01))),c(dnorm(seq(1,3,.01)),
dnorm(rev(seq(1,3,.01)))),col="orange", lty=2, lwd=2,
border="red")
```

Example

Generate 5 random numbers from $N(2, 2^2)$

```
> rnorm(5, mean=2, sd=2)
[1]  3.1985264  0.4473159 -0.8669186 -1.1561986  3.0654766

> rnorm(5) # This generates with mean 0 and sd=1
[1]  0.6998394  1.0070251 -1.0634292 -0.1712023  0.9539368
```

Obtain 95% quantile for the standard normal distribution

```
> qnorm(0.95)
[1] 1.644854
```

Binomial Distribution

A Binomial experiment consists of n independent Bernoulli trials - with the probability of success p remaining constant throughout the trials. Let X be the number of successes recorded. Then X is said to have the binomial distribution with parameters n and p.

Binomial Properties

a) The experiment consists of a fixed number, n, of Bernoulli trials.

b) The trials are identical and independent with probability of success, p.

c) The random variable X denotes the the number of successes obtained in the n trials.

The probability mass function of the binomial random variable X is:

$$P(X = x) = \binom{n}{x} p^{x} (1 - p)^{n-x}, \qquad x = 0, 1, \ldots, n;$$

When $n = 1$ the binomial distribution reduces to the bernoulli distribution.

Binomial Distribution

- dbinom(x, size, prob)
- pbinom(q, size, prob)
- qbinom(p, size, prob)
- rbinom(n, size, prob)

In Binomial distribution where size is the sample size and prob is the probability of success. Example:
What is the probability of 0 to 5 heads when a fair coin is tossed 10 times
dbinom(0:5, 10, .5)
Probability of 5 or less heads of fair coin out of 10 flips
pbinom(5, 10, .5)
Suppose you have a biased coin that has a probability of 0.8 of coming up heads. The probability of getting 5 heads in 16 tosses of this coin is
dbinom(5,16, .8)
The 0.25 quantile is
qbinom(.25,16,.8)

Poisson Distribution

Given a continuous interval (in time, length, etc), assume discrete events occur randomly throughout the interval. If the interval can be partitioned into subintervals of small enough length such that

(i) the probability of more than one occurrence in a subinterval is zero;

(ii) the probability of one occurrence in a subinterval is the same for all subintervals and proportional to the length of the subinterval;

(iii) the occurrence of an event in one subinterval has no effect on the occurrence or non-occurrence in another non-overlapping subinterval,

If the mean number of occurrences in the interval is λ, the random variable X that equals the number of occurrences in the interval is said to have the Poisson distribution with parameter λ.

The probability mass function of the poisson random variable X is:

$$P(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!}, \qquad x = 0, 1, 2, \ldots;$$

Poisson Distributions

- dpois(x, lamda)
- ppois(q, lamda
- qpois(p, lamda)
- rpois(n, lamda)

In Poisson distribution with mean=λ probability of 0,1, or 2 events with $\lambda = 4$
dpois(0:2, 4)
probability of at least 3 events with $\lambda = 4$
1- ppois(2,4)
Suppose a certain region of California experiences about 5 earthquakes a year. Assume occurrences follow a Poisson distribution. What is the probability of 3 earthquakes in a given year?
Here $\lambda = 5$.
dpois(3,5)