Continuous random variables

Ignacio Cascos

2018

Outline

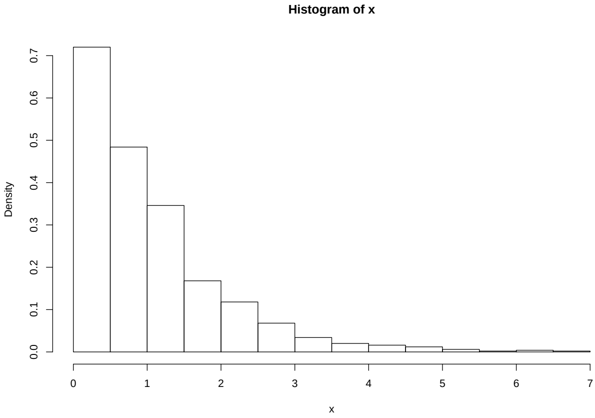
Introduction

A **random variable** $X : S \mapsto \mathbb{R}$ is **continuous** if its support $X(S)$ contains an interval of real numbers, or more precisely if its probability law can be described in terms of a nonnegative real function f_X (**density mass function**) in such a way the the probability that X lies on any (borelian) $A \subset \mathbb{R}$ can be computed as

$$P(X \in A) = \int_A f_X(x)dx\,.$$

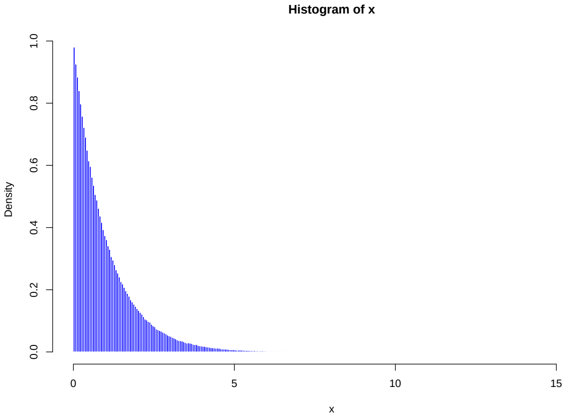
Introduction (histogram, 1000 observations)

```
set.seed(1)
x=rexp(1000)
hist(x,probability=T)
```



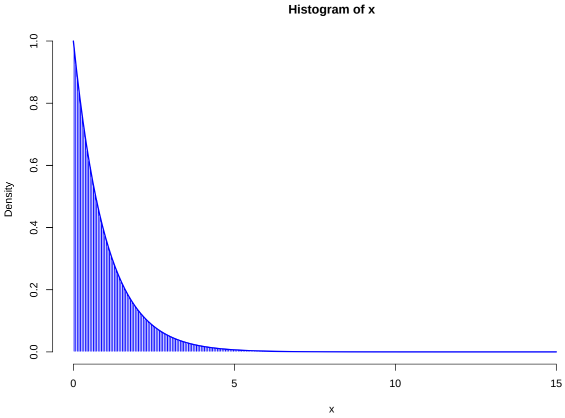
Introduction (histogram, 1000000 observations)

```
set.seed(1)
x=rexp(1000000)
hist(x,nclass=250,probability=T,border="white",col="blue")
```



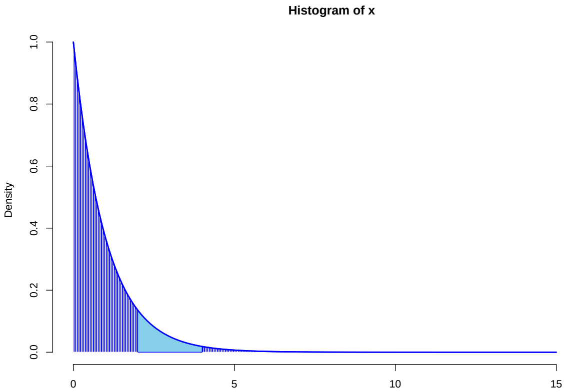
Introduction (histogram and density)

```
hist(x,nclass=250,probability=T,border="white",col="blue")
t=seq(0,15,by=.1)
points(t,dexp(t),type="l",col="blue",lwd=2)
```



Introduction (probability of an interval, $P(2 < X < 4)$)

```
cord.x <- c(2,seq(2,4,0.01),4)
cord.y <- c(0,dexp(seq(2,4,0.01)),0)
polygon(cord.x,cord.y,col='skyblue')
points(t,dexp(t),type="l",col="blue",lwd=2)
```



3.1 Density mass function and cdf

Definition of density mass function

Every **density mass function** $f : \mathbb{R} \mapsto \mathbb{R}$ satisfies 1. $f(x) \geq 0$,; 2. $\int_{-\infty}^{+\infty} f(x)dx = 1$,.

Density mass function of a continuous r.v.

If X is a *continuous* r.v. and f_X its associated density mass function, the probability that X lies in any (borelian) $A \subset X$ is

$$P(X \in A) = \int_A f_X(x)dx\,.$$

When $A = [a, b]$, we have $P(a \leq X \leq b) = \int_a^b f_X(x)dx\,.$

Properties of continuous r.v.s

- $P(X = a) = \int_a^a f_X(x)dx = 0$ for any $a \in \mathbb{R}$;
- $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$.

Example

$$f_X(x) = \begin{cases} e^{-x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

$$P(2 < X < 4) = \int_2^4 e^{-x} dx = -(e^{-4} - e^{-2}) = 0.117\,.$$

3.1 Density mass function and cdf

Cumulative distribution function, cdf

The **cumulative distribution function (cdf)** of r.v. X evaluated at $x \in \mathbb{R}$ is the probability that X is not greater than x,

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt.$$

Properties of the cdf of a continuous random variable

- $\lim_{x \to -\infty} F(x) = 0$;
- $\lim_{x \to +\infty} F(x) = 1$;
- F is nondecreasing;
- F is continuous.

3.1 Density mass function and cdf

The probability that X lies in the interval $[a, b]$ is computed in terms of its cdf as

$$P(a \leq X \leq b) = F_X(b) - F_X(a) \,.$$

Relationship between density mass function and cdf

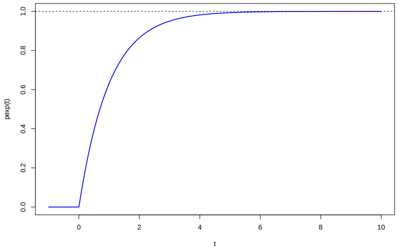
- The cdf is a primitive of the density mass function, $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$.
- The density mass function is the derivative of the cdf, $f_X(x) = F'_X(x)$.

Example (cdf)

$$F_X(x) = \begin{cases} 1 - e^{-x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

$$P(2 < X < 4) = F_X(4) - F_x(2) = (1 - e^{-4}) - (1 - e^{-2}) = 0.117$$

```
t=seq(-1,10,by=.1)
plot(t,pexp(t),type="l",col="blue",lwd=2)
abline(h=1,lty=2)
```



Mean or expectation

The **mean** or **expectation** of X is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x) dx, .$$

Properties of the mean

For any real numbers $a, b \in \mathbb{R}$, any function $g : \mathbb{R} \mapsto \mathbb{R}$, and r.v. X,

- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$;
- $\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$;
- $\mathbb{E}[(X - \mathbb{E}[X])^2] = \min_{x \in \mathbb{R}} \mathbb{E}[(X - x)^2]$.

3.2 Mean, variance, and quantiles

Variance

The **variance** is a measure of the *scatter* of the distribution of r.v. X.
It is the expected squared distance of X to its mean,

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])^2 f_X(x) d(x) \,.$$

Properties of the variance

- $\mathrm{Var}[X] \geq 0$;
- $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$;
- $\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]$, for any $a, b \in \mathbb{R}$.

Standard deviation

The **standard deviation** of X is the (positive) square root of its variance,

$$\sigma_X = \sqrt{\mathrm{Var}[X]} \,.$$

Example (mean, variance)

X with the previous density.

$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x) dx = \int_0^{+\infty} x e^{-x} dx = [-xe^{-x}]_0^{+\infty} + \int_0^{+\infty} e^{-x} dx = 1.$

$\mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2 f_X(x) dx = \int_0^{+\infty} x^2 e^{-x} dx = 2 \int_0^{+\infty} x e^{-x} dx = 2.$

$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1.$

```
set.seed(1)
x=rexp(10000)
mean(x)
```

```
## [1] 0.9983612
```

```
var(x)
```

```
## [1] 1.031541
```

3.2 Mean, variance, and quantiles

Median

The **median** is the most central value with respect to the distribution of a random variable X in the sense that

$$F_X(\mathrm{Me}_X) = P(X \leq \mathrm{Me}_X) = 1/2\,.$$

Example

Solve $F_X(\mathrm{Me}_X) = 1/2$, then $1 - e^{-\mathrm{Me}_X} = 1/2$, and $\mathrm{Me}_X = -\log(1/2) = \log(2) = 0.693$.

Properties of the median

- $\mathrm{Me}_{aX+b} = a\mathrm{Me}_X + b$, for any $a, b \in \mathbb{R}$;
- $\mathrm{Me}_{g(X)} = g(\mathrm{Me}_X)$ if g is monotone;
- $\mathbb{E}|X - \mathrm{Me}_X| = \min_{x\in\mathbb{R}} \mathbb{E}|X - x|$.

3.2 Mean, variance, and quantiles

Quantiles

For $0 < \alpha < 1$ the α-**quantile** of random variable X a number q_α such that

$$F_X(q_\alpha) = P(X \leq q_\alpha) = \alpha\,.$$

The **quantile function** of random variable X is defined as

$$F_X^{-1}(\alpha) = \inf\{x :\ F_X(x) \geq \alpha\}.$$

A quantile function defined like this is:

- $\lim\limits_{\alpha\downarrow 0} F_X^{-1}(\alpha) = \inf X(S)$;
- $\lim\limits_{\alpha\uparrow 1} F_X^{-1}(\alpha) = \sup X(S)$;
- *nondecreasing*;
- *left-continuous*.

Example (quantiles)

X with the previous density. If $F_X(x) = 1 - e^{-x} = y$, then $y = -\log(1-x)$, so

$$F_X^{-1}(x) = -\log(1-x)\,.$$

Half of the random variables with the distribution of X assume a value greater (or less) than $\mathrm{Me}_X = 0.693$, while 75% assume a value greater than $F^{-1}(0.25) = -\log(0.75) = 0.288$.

```
median(x)
```

```
## [1] 0.6946537
```

```
quantile(x,0.25)
```

```
##       25%
## 0.2810167
```

3.3 Uniform distribution `unif(min=0,max=1)`

A **Uniform** random variable in the interval $[a, b]$ represents a number at random in that interval selected in such a way that the probability that it lies in any subinterval of $[a, b]$ is proportional to the width of the subinterval.

$$X \sim \mathrm{U}(a, b)$$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases} .$$

```r
dunif(x,min=a,max=b)
```

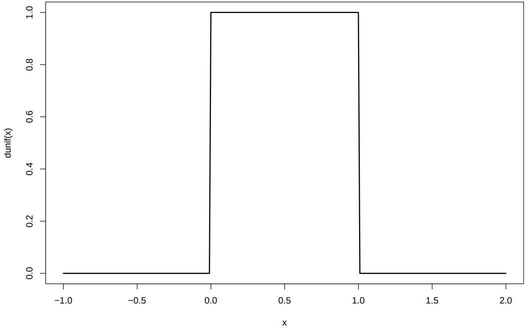
$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \le x \le b \\ 1 & \text{if } x > b \end{cases} .$$

```r
punif(x,min=a,max=b)
```

$$\mathbb{E}[X] = \frac{a+b}{2} \quad ; \quad \mathrm{Var}[X] = \frac{(b-a)^2}{12}$$

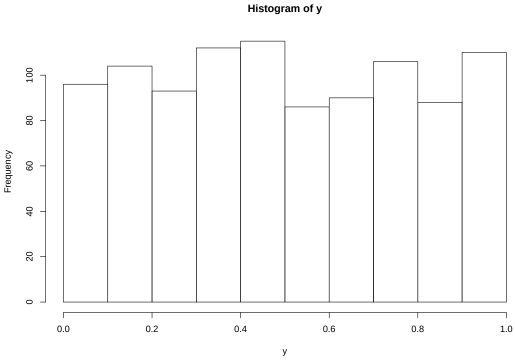
Uniform density mass function `dunif(min=0,max=1)`

```
x=seq(-1,2,by=.01)
plot(x,dunif(x),type="l",lwd=2)
```



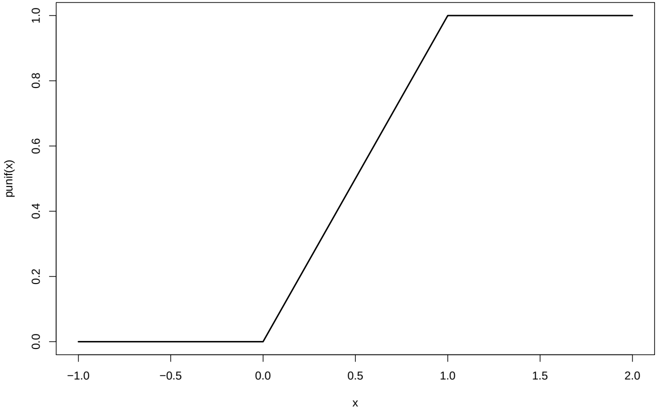
Uniform random observations runif(min=0,max=1)

```
set.seed(1)
y=runif(1000)
hist(y)
```



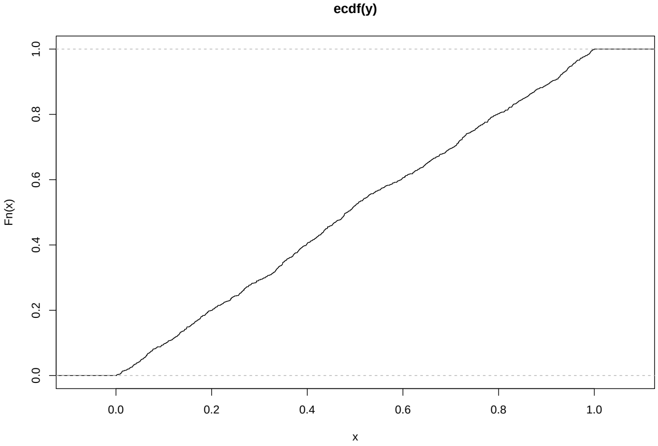
Uniform cdf punif(min=0,max=1)

```
x=seq(-1,2,by=.01)
plot(x,punif(x),type="l",lwd=2)
```



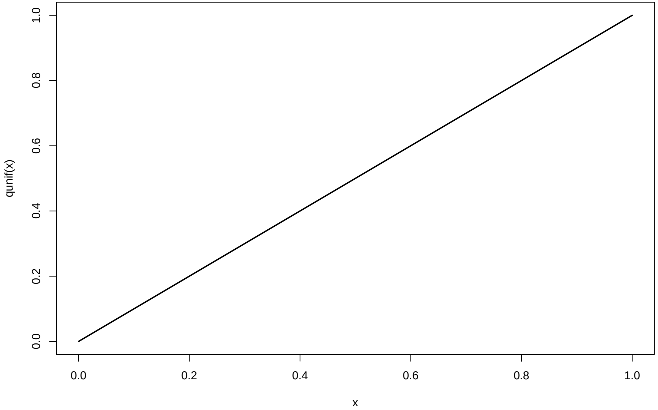
Uniform empirical cumulative distribution function

```r
plot(ecdf(y))
```



Uniform quantile function `qunif(min=0,max=1)`

```
x=seq(0,1,by=.01)
plot(x,qunif(x),type="l",lwd=2)
```



3.4 Transformations of a random variable

If X is a random variable and $g : \mathbb{R} \mapsto \mathbb{R}$ a function, then $Y = g(X)$ is a random variable.

If X is continuous and g continuous and increasing

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)),$$

where g^{-1} is the inverse function of g, that is, $g^{-1}(y) = x$ if $g(x) = y$.

In general, if g is injective (one-to-one) and derivable

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| .$$

Example

Consider $X \sim \mathrm{U}(0,1)$, determine the density mass function of $Y = -\log(1-X)$.
Clearly the support of Y is $(0,+\infty)$, consider $y > 0$

$$F_Y(y) = P(Y \le y) = P(-\log(1-X) \le y) = P(\log(1-X) \ge -y)$$
$$= P(1-X \ge e^{-y}) = P(-X \ge e^{-y}-1) = P(X \le 1-e^{-y}) = 1-e^{-y}\,.$$

$$F_Y(y) = \begin{cases} 1-e^{-y} & \text{if } y \ge 0 \\ 0 & \text{if } y < 0 \end{cases}\,.$$

Inverse transform method for simulation

If $X \sim \mathrm{U}(0,1)$, then $F^{-1}(X)$ is a random variable with cdf F.

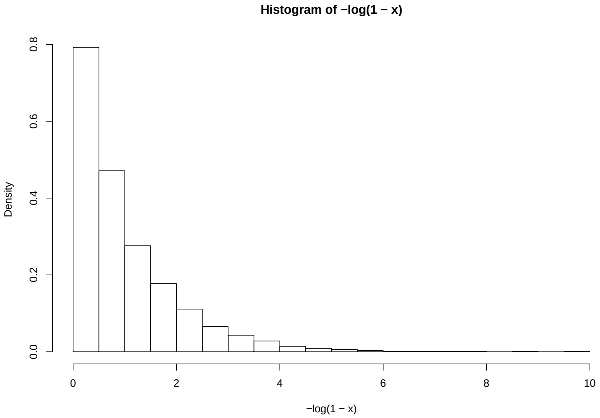
$$P(F^{-1}(X) \leq x) = P(X \leq F(x)) = F_X(F(x)) = F(x)$$

Example

Observe that if $F(x) = 1 - e^{-x}$ for $x \geq 0$, then $F^{-1}(x) = -\log(1-x)$. The cdf of $Y = -\log(1-X)$ is F and we can use this to simulate from such a distribution.

Inverse transform method for simulation (Example)

```
set.seed(1)
x=runif(10000)
hist(-log(1-x),probability=T)
```



3.5 Exponential distribution `exp(rate=1)`

If $X \sim \mathcal{P}(\lambda)$ represents the number of events that occur in a given time period (independently and with constant rate λ events per time units in the period), then the time between two consecutive events follows an **Exponential** distribution with parameter λ.

$$X_t \equiv \text{'number of events in [0,t]'}$$

$$T \equiv \text{'time until first event occurs'}$$

$$X_t \sim \mathcal{P}(\lambda t)$$

Take $t > 0$,

$$F_T(t) = P(T \leq t) = 1 - P(T > t) = 1 - P(X_t = 0) = 1 - e^{-\lambda t}\,.$$

Exponential distribution exp(rate=1)

$T \sim \mathrm{Exp}(\lambda)$

- cdf

$$F_T(t) = \begin{cases} 1 - e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

- density

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

$$\mathbb{E}[T] = \lambda^{-1} \quad ; \quad \mathrm{Var}[T] = \lambda^{-2}$$

Exponential distribution exp(rate=1)

Lack of memory property

If $T \sim \mathrm{Exp}(\lambda)$ and $t_1, t_2 > 0$, then

$$P(T > t_1 + t_2 | T > t_1) = P(T > t_2).$$

Proof:

$$P(T > t_1 + t_2 | T > t_1) = \frac{P((T > t_1 + t_2) \cap (T > t_1))}{P(T > t_1)} = \frac{P(T > t_1 + t_2)}{P(T > t_1)}$$

$$= \frac{1 - F_T(t_1 + t_2)}{1 - F_T(t_1)} = \frac{e^{-\lambda(t_1 + t_2)}}{e^{-\lambda t_1}} = e^{-\lambda t_2} = P(T > t_2).$$

3.6 Normal distribution `norm(mean=0,sd=1)`

Random variable X follows a normal distribution with mean μ and standard deviation σ, $X \sim \mathrm{N}(\mu, \sigma)$ if its density mass function is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

```
dnorm(x,mean=mu,sd=sigma)
```

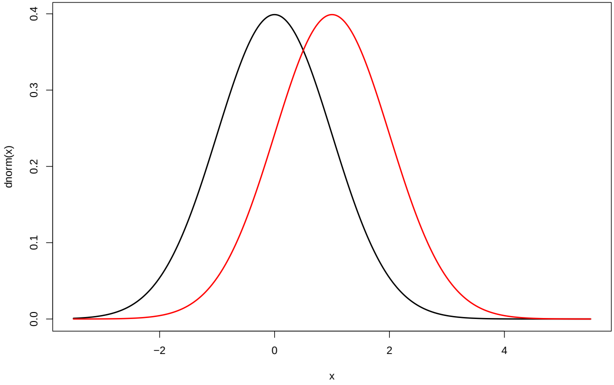
We refer to $Z \sim \mathrm{N}(0,1)$ as **standard normal** random variable,

$$f_Z(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}.$$

```
dnorm(x)
```

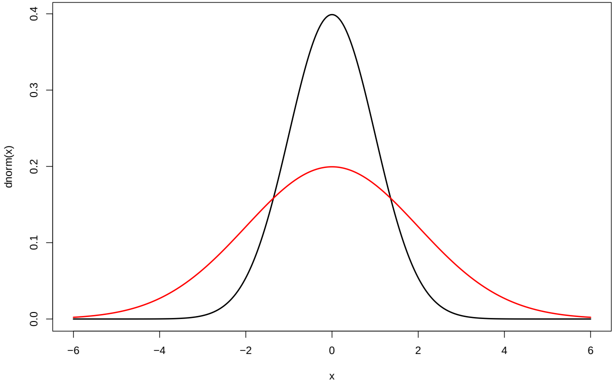
Normal density (location shift) `dnorm(mean=0,sd=1)`

```
x=seq(-3.5,5.5,by=.01)
plot(x,dnorm(x),type="l",lwd=2)
points(x,dnorm(x,mean=1,sd=1),type="l",lwd=2,col="red")
```



Normal density (scale shift) dnorm(mean=0,sd=1)

```
x=seq(-6,6,by=.01)
plot(x,dnorm(x),type="l",lwd=2)
points(x,dnorm(x,mean=0,sd=2),type="l",lwd=2,col="red")
```

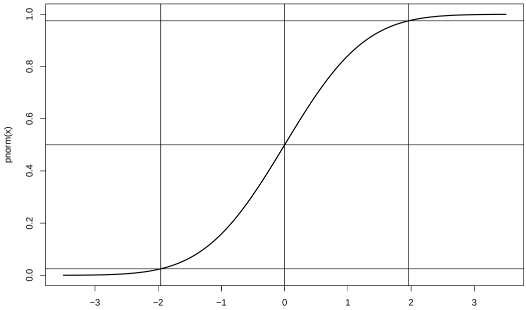


Normal cdf pnorm(mean=0,sd=1)

There is no analitic expression for the cdf of a normal r.v.

If $Z \sim \mathrm{N}(0,1)$, $F_Z(x) = P(Z \leq z) = \int_{-\infty}^{x} \phi(t)dt = \Phi(x)$.

```
x=seq(-3.5,3.5,by=.01)
plot(x,pnorm(x),type="l",lwd=2)
abline(h=c(0.025,0.5,0.975),v=c(-1.96,0,1.96))
```



Normal distribution `norm(mean=0,sd=1)`

A linear transformation of a normal random variable is normal

If $X \sim \mathrm{N}(\mu, \sigma)$ and $a, b \in \mathbb{R}$,

$$aX + b \sim \mathrm{N}(a\mu + b, |a|\sigma)\,.$$

Standardization

Among all linear tranformations of a normal r.v., the most relevant is the **standardization**, if $X \sim \mathrm{N}(\mu, \sigma)$,

$$\frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1)\,.$$

Normal distribution `norm(mean=0,sd=1)`

Examples

If $X \sim \mathrm{N}(\mu = 2, \sigma = 3)$, compute:

- $P(X \leq 4)$

```
pnorm(4,mean=2,sd=3)
```

```
## [1] 0.7475075
```

- $P(X \leq 4) = P((X-2)/3 \leq (4-2)/3) = \Phi(2/3)$

```
pnorm(2/3)
```

```
## [1] 0.7475075
```

Normal approximation to the Binomial distribution

DeMoivre-Laplace limit theorem

For $0 < p < 1$ and $r \in \{0, 1, 2, \ldots, n\}$

$$\frac{\sqrt{2\pi np(1-p)}\binom{n}{r}p^r(1-p)^{n-r}}{e^{-(r-np)^2/(2np(1-p))}} \stackrel{n\to+\infty}{\longrightarrow} 1$$

Consequence:
If $X \sim \mathrm{B}(n, p)$, then for any $a < b$, we have

$$P\left(a \leq \frac{X - np}{\sqrt{np(1-p)}} \leq b\right) \stackrel{n\to+\infty}{\longrightarrow} \Phi(b) - \Phi(a)$$

Good approximation for values of n satisfying $np(1-p) \geq 10$.

Normal approximation to the Binomial distribution

$X \sim \mathrm{B}(n = 40, p = 0.5)$

- $P(X = 20)$

```
dbinom(20,size=40,prob=0.5)
```

```
## [1] 0.1253707
```

- $P(X = 20) = P(19.5 \leq X \leq 20.5)$

```
pnorm(20.5,mean=20,sd=sqrt(10))-pnorm(19.5,mean=20,sd=sqrt(10)
```

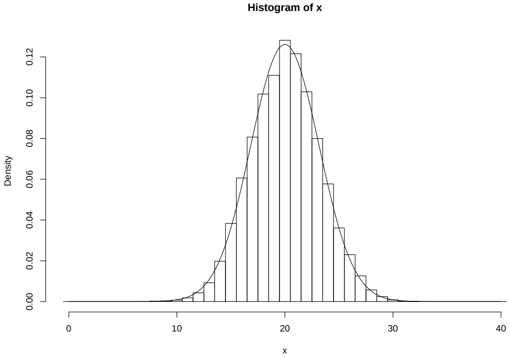
```
## [1] 0.1256329
```

```
dnorm(20,mean=20,sd=sqrt(10))
```

```
## [1] 0.1261566
```

Normal approximation to the Binomial distribution

```
set.seed(2)
x=rbinom(10000,size=40,prob=.5)
hist(x, breaks=seq(-0.5,40.5,1), probability=T)
t=seq(0,40,by=.01)
points(t,dnorm(t,mean=20,sd=sqrt(10)),type="l")
```



Continuous distributions in R

Distributions	R command
Uniform, $\mathrm{U}(a, b)$	`unif(min=0,max=1)`
Exponential, $\mathrm{Exp}(\lambda)$	`exp(rate=1)`
Normal, $\mathrm{N}(\mu, \sigma)$	`norm(mean=0,sd=1)`
Gamma, $\mathrm{Gamma}(k, \lambda)$	`gamma(shape,rate=1)`
Beta, $\mathrm{Beta}(\alpha, \beta)$	`beta(shape1,shape2)`
Chi-square, χ^2_n	`chisq(df)`
Student's t, t_n	`t(df)`
Fisher's F, F_{n_1,n_2}	`f(df1,df2)`

Functions	R prefix
density	`d`
cdf	`p`
quantile function	`q`
random numbers	`r`