Discrete random variables

Ignacio Cascos

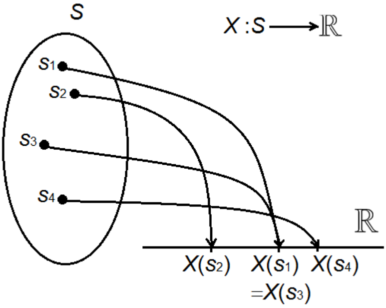
2018

Outline

Introduction

The outcome of a random experiment can often be presented in terms of a number. It is at least possible to summarize the relevant part of the outcome in a number.

A **random variable** is the numerical outcome of a random experiment, it is a number *at random*.



2.1 Definition of random variable

Associated with a random experiment we have a family of events with the σ-algebra structure, while in $\mathbb{R}$, the *borelian* sets (intervals, their complementaries, unions and intersections) do also form a σ-algebra.

Random variable

A **random variable** is a *measurable* mapping from the sample space asociated with a random experiment into the set of real numbers, $X : S \mapsto \mathbb{R}$.
It associates each outcome of a random experiment with a real number and is *measurable* because the inverse image of every borelian set does belong to the σ-algebra of events.

Events associated with a random variable

For any borelian $A \subset \mathbb{R}$, its inverse image through X, given by

$$X^{-1}(A) = \{s \in S : X(s) \in A\}$$

is an event, and as such, we can compute its probability.

$$\begin{aligned} P(X \in A) &= P(X^{-1}(A)) \\ &= P(\{s \in S : X(s) \in A\}) . \end{aligned}$$

If A is a singleton, then $P(X = x) = P(X^{-1}(x)) = P(\{s \in S : X(s) = x\})$.

We can ignore the original sample space S and consider a probability in $\mathbb{R}$ given by $P_X(A) = P(X \in A)$ for any borelian $A \subset \mathbb{R}$.

The support of a r.v. (discrete and continuous variables)

The **support** or **range** of a random variable $X(S)$ is the set of all values that it can assume.

- if $X(S)$ is a finite or denumerable set, then X is a **discrete** random variable.
- if $X(S)$ contains all the elements in an interval of real numbers, then X is a **continuous** random variable.

2.2 Discrete r.v.s, probability mass function, and cumulative distribution function

Probability mass function

A **probability mass function** associates each real number x with the probability that the random variable X exactly matches it,

$$p(x) = P(X = x).$$

Properties of the probability mass function

- $0 \le p(x) \le 1$ for every $x \in \mathbb{R}$
- if $X(S) = \{x_i\}_{i \in I}$, then $\sum_{i \in I} p(x_i) = 1$

The probability that X lies in any Borelian $A \subset \mathbb{R}$ is

$$P(X \in A) = \sum_{x_i \in X(S) \cap A} p(x_i).$$

2.2 Discrete r.v.s, probability mass function and cumulative distribution function

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,

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i), \quad \text{where } x_i \in \mathbb{R}.$$

Properties of the cdf of a discrete random variable

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

Example (6-face fair die)

$X \equiv$'outcome of a roll of a 6-face fair die'

$$p(x) = P(X = x) = \begin{cases} 1/6 & \text{if } x \in \{1,2,3,4,5,6\} \\ 0 & \text{otherwise} \end{cases},$$

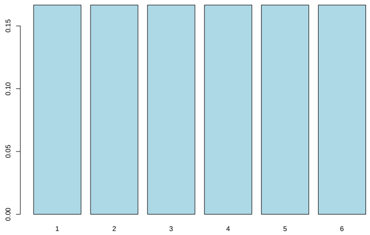
$$F(x) = P(X \le x) = \begin{cases} 0 & \text{if } x < 1 \\ 1/6 & \text{if } 1 \le x < 2 \\ 2/6 & \text{if } 2 \le x < 3 \\ 3/6 & \text{if } 3 \le x < 4 \\ 4/6 & \text{if } 4 \le x < 5 \\ 5/6 & \text{if } 5 \le x < 6 \\ 1 & \text{if } x \ge 6 \end{cases}.$$

Example (6-face fair die: probability mass function)

```
library(prob)
table(rolldie(1)$X1)
```

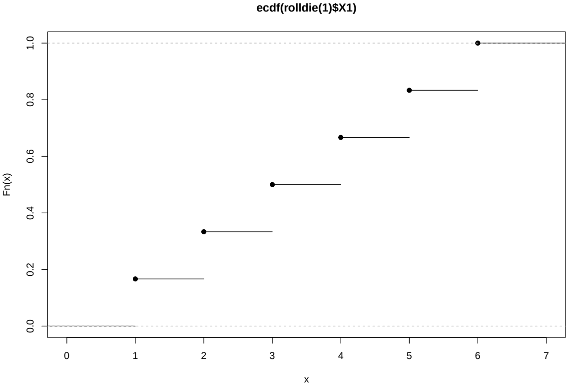
```
##
## 1 2 3 4 5 6
## 1 1 1 1 1 1
```

```
barplot(table(rolldie(1)$X1)/6,col="light blue")
```



Example (6-face fair die: cdf)

```
plot(ecdf(rolldie(1)$X1))
```




```
ecdf(rolldie(1)$X1)(3)
```

2.3 Mean, variance, and quantiles

Mean or expectation

The **mean** or **expecation** of X is its probability-weighted average

$$\mathbb{E}[X] = \sum_{i \in I} x_i p_X(x_i)\,.$$

Transformation of a random variable

If X discrete r.v. and $g : \mathbb{R} \mapsto \mathbb{R}$ function, then $g(X)$ is a discrete r.v. with probability mass function $p_{g(X)}(y) = P(g(X) = y) = \sum_{g(x_i)=y} P(X = x_i)$.

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,

- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$;
- $\mathbb{E}[g(X)] = \sum_{i \in I} g(x_i) p_X(x_i)$;
- $\mathbb{E}[(X - \mathbb{E}[X])^2] = \min_{x \in \mathbb{R}} \mathbb{E}[(X - x)^2]$.

2.3 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,

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_i (x_i - \mathbb{E}[X])^2 p_X(x_i).$$

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 (6-face fair die: mean)

$$X \equiv \text{'outcome of a roll of a 6-face fair die'}$$

$$\mathbb{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5\,.$$

```
set.seed(1)
x=sim(probspace(rolldie(1)),ntrials=1000)
mean(x$X1)
```

```
## [1] 3.488
```

```
sample((1:6),size=1000,replace=T)
```

Example (6-face fair die: variance and std. dev.)

$$X \equiv \text{'outcome of a roll of a 6-face fair die'}$$

$$\mathbb{E}[X^2] = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} = 15.1667\,,$$

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 2.9167\,,$$

$$\sigma_X = \sqrt{\mathrm{Var}[X]} = 1.7078\,.$$

```r
var(x$X1)
```

```
## [1] 2.880737
```

```r
sd(x$X1)
```

```
## [1] 1.697273
```

Median

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

$$P(X \leq \mathrm{Me}_X) \geq 1/2 \quad \text{and} \quad P(X \geq \mathrm{Me}_X) \geq 1/2\,.$$

Example (roll of a 6-face fair die)}

Any value in the interval $[3,4]$ is a median of the outcome of the die.

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;
- $\mathbb{E}|X - \mathrm{Me}_X| = \min_{x\in\mathbb{R}} \mathbb{E}|X - x|$.

Quantiles

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

$$P(X \leq q_\alpha) \geq \alpha \quad \text{and} \quad P(X \geq q_\alpha) \geq 1 - \alpha\,.$$

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

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

A quantile function defined like this is:

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

Example (6-face fair die quantiles)

$X \equiv$'outcome of a roll of a 6-face fair die'

$$F^{-1}(x) = \begin{cases} 1 & \text{if } \quad 0 < x \leq 1/6 \\ 2 & \text{if } 1/6 < x \leq 2/6 \\ 3 & \text{if } 2/6 < x \leq 3/6 \\ 4 & \text{if } 3/6 < x \leq 4/6 \\ 5 & \text{if } 4/6 < x \leq 5/6 \\ 6 & \text{if } 5/6 < x \leq 1 \end{cases} .$$

2.4 The Bernoulli process

Bernoulli trial

A **Bernoulli trial** is a random experiment that can only result in two possible outcomes. Commonly we refer to these outcomes as *success* (S) and *failure* (F). The probability of *success* is denoted by $0 < p < 1$ and this experiment can be repeated ***independently*** as many times as needed.

Binomial distribution `binom(size,prob)`

Consider a Bernoulli trial with probability of success p and that is carried out *independently* n times, a **Binomial** random variable X with parameters n and p represents the number of trials that result in success.

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

$$P(X = r) = \binom{n}{r} p^r (1-p)^{n-r}, \quad r \in \{0, 1, \ldots, n\}$$

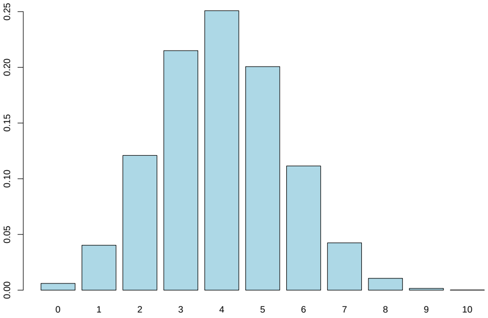
```
dbinom(r,size=n,prob=p)
```

$$\mathbb{E}[X] = np \quad ; \quad \mathrm{Var}[X] = np(1-p)$$

If $X \sim \mathrm{B}(n_1, p)$ and $Y \sim \mathrm{B}(n_2, p)$ are *indep.*, then $X + Y \sim \mathrm{B}(n_1 + n_2, p)$.

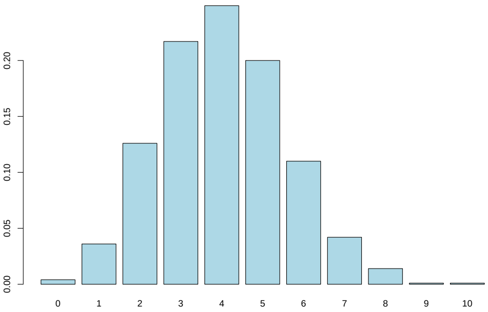
Binomial probability mass function dbinom

```
r=c(0:10)
barplot(dbinom(r,size=10,prob=0.4),
        names.arg=r,col="light blue")
```



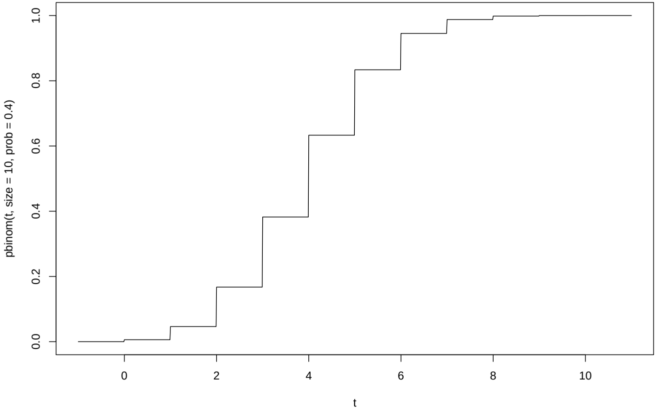
Binomial random observations generation rbinom

```
set.seed(1)
x=rbinom(1000,size=10,prob=0.4)
barplot(table(x)/1000,col="light blue")
```



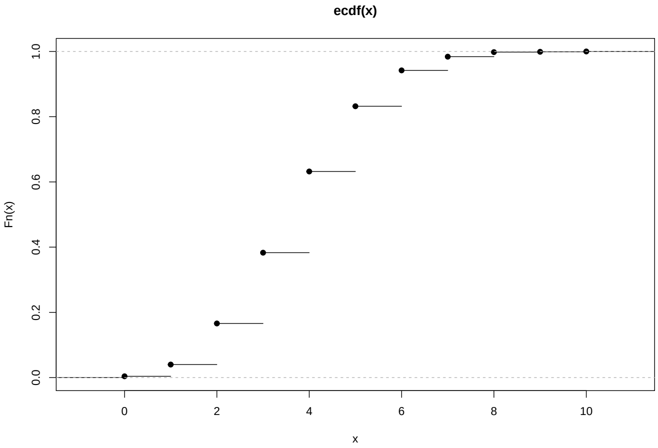
Binomial cumulative distribution function pbinom

```
t=seq(-1,11,by=.01)
plot(t,pbinom(t,size=10,prob=0.4),type="l")
```



Binomial empirical cumulative distribution function

```
plot(ecdf(x))
```



Binomial ecdf and cdf

```
sum(x<=4)/1000
```

```
## [1] 0.632
```

```
ecdf(x)(4)
```

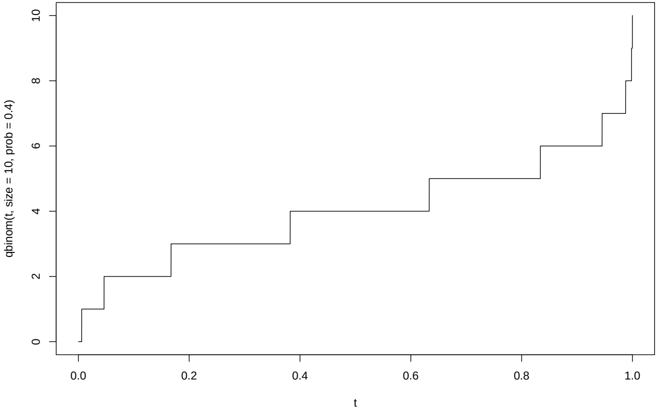
```
## [1] 0.632
```

```
pbinom(4,size=10,prob=0.4)
```

```
## [1] 0.6331033
```

Binomial quantile function

```
t=seq(0,1,by=.0001)
plot(t,qbinom(t,size=10,prob=0.4),type="l")
```



Geometric (Pascal's) distribution `geom(prob)`

Consider a Bernoulli trial with probability of success p, the number of *independent* trials that result in failure obtained before the first success follows a **Geometric** distribution with parameter p.

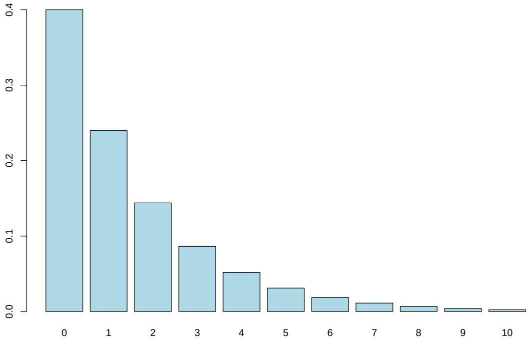
$$X \sim \mathcal{G}(p)$$

$$P(X = r) = p(1-p)^r, \quad r \in \{0, 1, 2, \ldots\}$$

$$\mathbb{E}[X] = \frac{1-p}{p} \quad ; \quad \mathrm{Var}[X] = \frac{1-p}{p^2}$$

Geometric probability mass function dgeom

```
r=c(0:10)
barplot(dgeom(r,prob=0.4),
        names.arg=r,col="light blue")
```



Negative Binomial distribution `nbinom(size,prob)`

Consider a Bernoulli trial with probability of success p, the number of failures (*independent* trials that result in failure) before the k-th success (trials that result in sucess) follows a **Negative Binomial** distribution with parameters k and p.

$$X \sim \mathrm{NB}(k, p)$$

$$P(X = r) = \binom{r+k-1}{r} p^k (1-p)^r, \quad r \in \{0, 1, 2, \ldots\}$$

$$\mathbb{E}[X] = \frac{k(1-p)}{p} \quad ; \quad \mathrm{Var}[X] = \frac{k(1-p)}{p^2}$$

2.5 Hypergeometric distribution `hyper(m,n,k)`

Consider a finite population with $N_1 + N_2$ objects, such that N_1 are of type 1 and N_2 are of type 2. A total number of k objects are selected from the population *without replacement*. The number of objects of type N_1 in the selection follows a **Hypergeometric** distribution with parameters N_1, N_2, and k.

$$X \sim \mathrm{H}(N_1, N_2, k)$$

$$P(X = r) = \frac{\binom{N_1}{r}\binom{N_2}{k-r}}{\binom{N_1+N_2}{k}}, \quad r \in \{\max\{0, k - N_2\}, \dots, \min\{k, N_1\}\}$$

$$\mathbb{E}[X] = \frac{kN_1}{N_1 + N_2} \quad ; \quad \mathrm{Var}[X] = k \cdot \frac{N_1 N_2}{(N_1 + N_2)^2} \cdot \frac{N_1 + N_2 - k}{N_1 + N_2 - 1}$$

2.6 Poisson distribution `pois(lambda)`

The number of events that occur in a region of space (or time) *independently* one from the others and at a constant rate $\lambda > 0$ follows a **Poisson** distribution with parameter λ.

$$X \sim \mathcal{P}(\lambda)$$

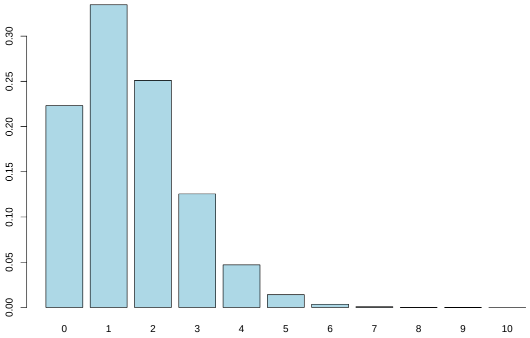
$$P(X = r) = e^{-\lambda}\frac{\lambda^r}{r!}, \quad r \in \{0, 1, 2, \ldots\}$$

$$\mathbb{E}[X] = \lambda \quad ; \quad \mathrm{Var}[X] = \lambda$$

If $X \sim \mathcal{P}(\lambda_1)$ and $Y \sim \mathcal{P}(\lambda_2)$ are *independent*, then $X + Y \sim \mathcal{P}(\lambda_1 + \lambda_2)$.

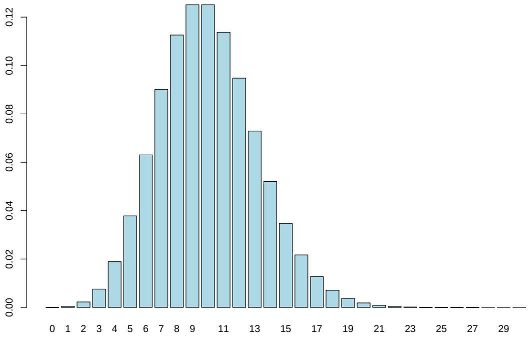
Poisson probability mass function $\lambda = 1.5$

```
t=c(0:10)
barplot(dpois(t,lambda=1.5),
        names.arg=t,col="light blue")
```



Poisson probability mass function $\lambda = 10$

```
t=c(0:30)
barplot(dpois(t,lambda=10),
        names.arg=t,col="light blue")
```



Discrete distributions in R

Distributions	R command
Binomial, $\mathrm{B}(n, p)$	`binom(size,prob)`
Geometric, $\mathcal{G}(p)$	`geom(prob)`
Negative Binomial, $\mathrm{NB}(k, p)$	`nbinom(size,prob)`
Hypergeometric, $\mathrm{H}(N_1, N_2, k)$	`hyper(m,n,k)`
Poisson, $\mathcal{P}(\lambda)$	`pois(lambda)`

Functions	R prefix
probability function	`d`
cumulative probability	`p`
quantile function	`q`
random numbers	`r`