Course 02402 Introduction to Statistics Lecture 2:

Random variables and discrete distributions

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda

1. Random Variables and the density function
2. Distribution function
3. Specific discrete distributions I: The binomial
   - Example 1
4. Specific distributions II: The hypergeometric
   - Example 2
5. Specific distributions III: The Poisson
   - Example 3
6. Distributions in R
7. Mean and Variance
   - Mean and variances for specific discrete distributions

Oversigt

Random Variables

A random variable represents a value of the outcome *before* the experiment is carried out

- A dice throw
- The number six'es in 10 dice throws
- km/l for for a car
- Measurement of glucose level in blood sample
- ...

Discrete or continuous

- We distinguish between discrete and continuous
- Discrete are countable:
  - How many use glassse in this room
  - The number of plains departing the next hour
- Kontinuert:
  - Wind speed measurement
  - Transport time to DTU

Today: discrete. Next week: Continuous

Random variable

Before the experiment is carried out, random variable:

$$X \text{ (or } X_1, \ldots, X_n)$$

indicated with capital letters
Then the experiment is carried out, and then we have a *realization* or *observation*

$$x \text{ (or } x_1, \ldots, x_n)$$

indicated with small letters

Simulate rolling a dice

Make a random draw from (1,2,3,4,5,6) with equal probability for each outcome

Discrete distributions

- The concept is to described the experiment before it is carried out
- What to do when we do not know the outcome?
- Solution: us the density function

Density function

A random variable has a *density function* (probability density function (pdf))

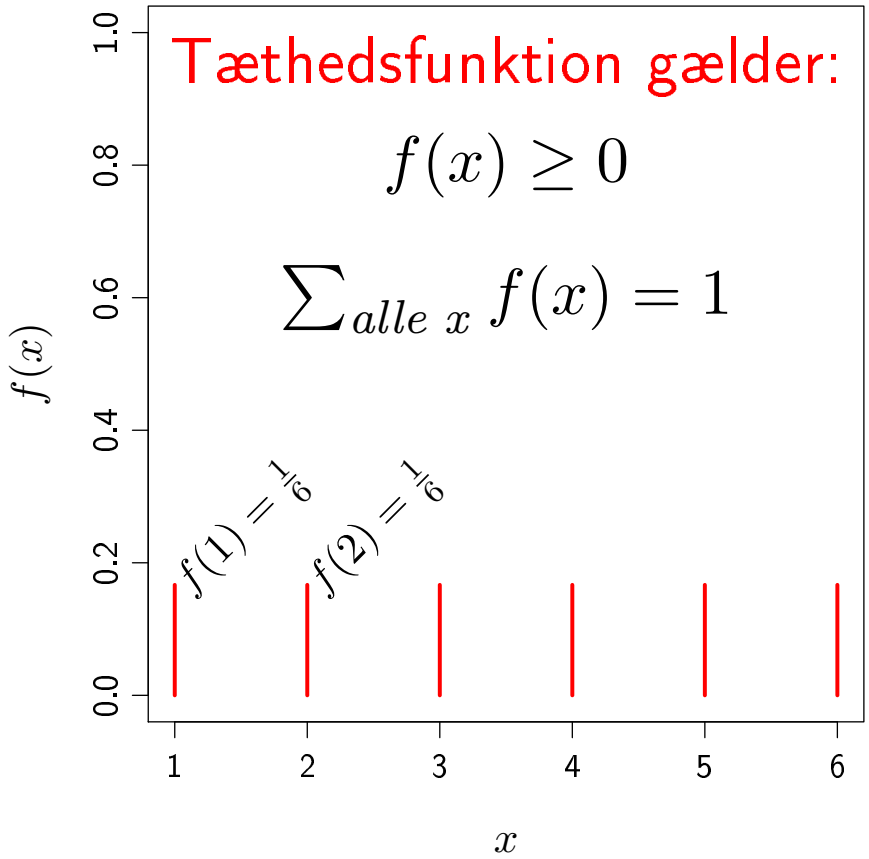
Definition

$$f(x) = P(X = x)$$

The probability that the X becomes x when the experiment is carried out

Density function

A fair dice density function



Sample

If we only have a single observation, can we see the distribution? No
but if we have n observations, then we have a *sample*

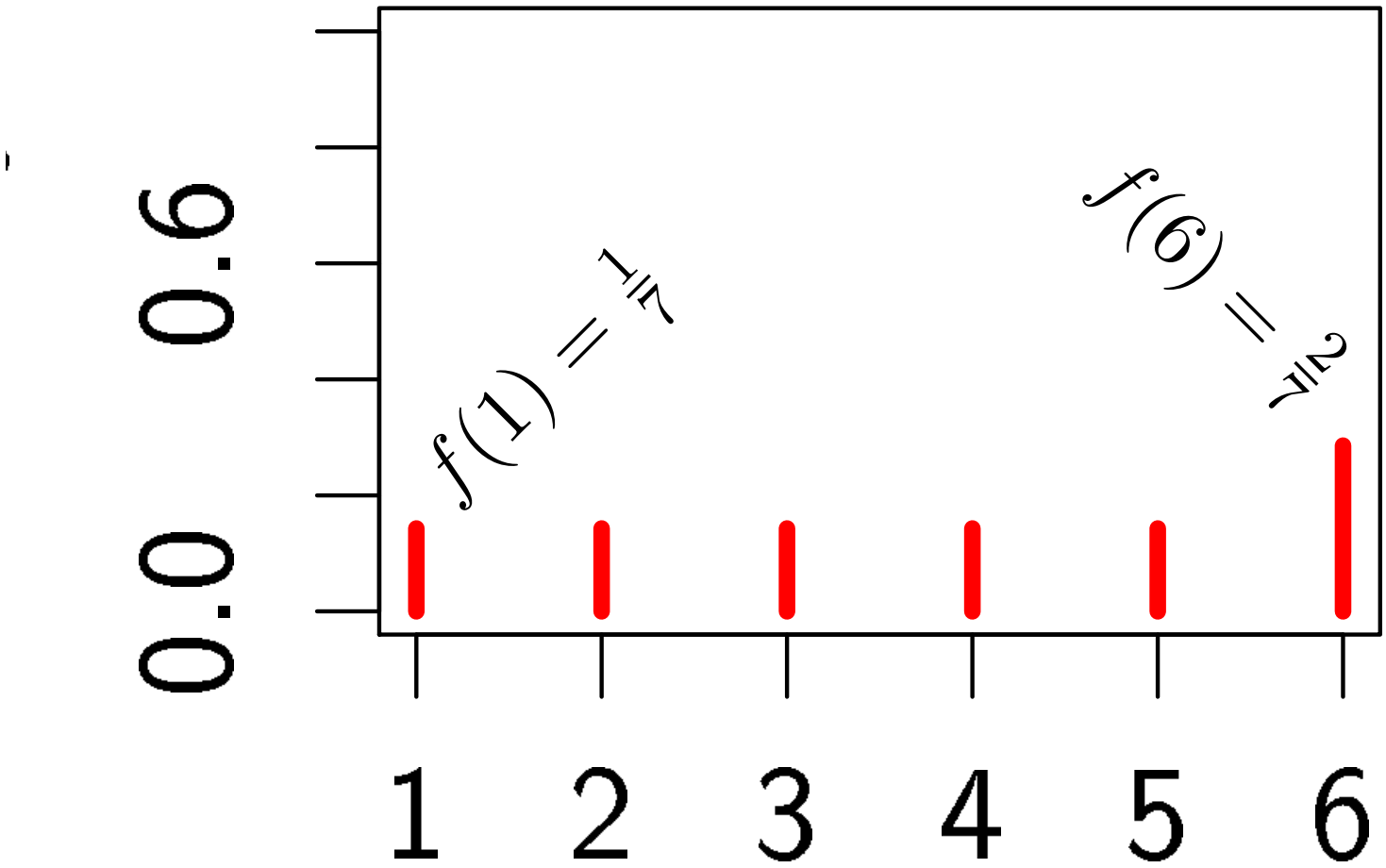
$$\{x_1, x_2, \ldots, x_n\}$$

and we can begin to "see" the distribution.

Simulate n rolls with a fair dice

```
n <- 30

## Draw independently from the set (1,2,3,4,5,6) with
## equal probability
xFair <- sample(1:6, size=n, replace=TRUE)

## Print the values
xFair

## Count the number of each outcome using the table function
table(xFair)

## Plot the empirical pdf
plot(table(xFair)/n, lwd=10, ylim=c(0,1), xlab="x", ylab="Density")

## Add the pdf to the plot
lines(rep(1/6,6), lwd=4, type="h", col=2)
## Add a legend to the plot
legend("topright", c("Empirical pdf","pdf"), lty=1, col=c(1,2),
       lwd=c(5,2), cex=0.8)
```

An unfair dice density function



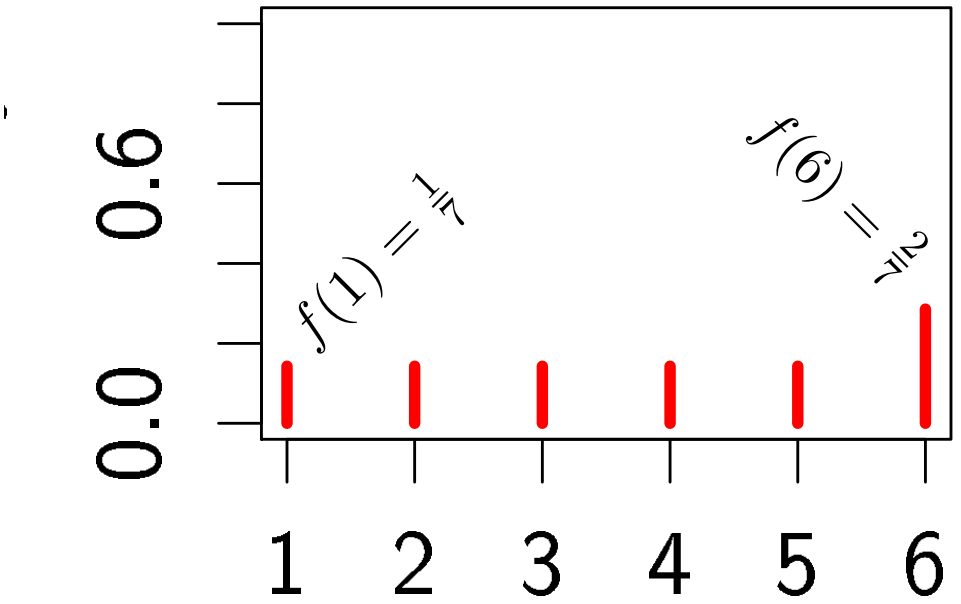
Simulate n rolls with an unfair dice

```
## Number of simulated realizations
n <- 30

## Draw independently from the set (1,2,3,4,5,6) with
## higher probability for a six
xUnfair <- sample(1:6, size=n, replace=TRUE, prob=c(rep(1/7,5),2/7))

## Plot the empirical density function
plot(table(xUnfair)/n, lwd=10, ylim=c(0,1), xlab="x", ylab="Density")

## Add the pdf to the plot
lines(c(rep(1/7,5),2/7), lwd=4, type="h", col=2)

## Add a legend
legend("topright", c("Empirical pdf","pdf"), lty=1, col=c(1,2), lwd=c(5,2))
```

Some questions

Find some probabilities for X^{unFair}:

- The probability to get a 4?
- The probability to get a 5 or a 6?
- The probability to get less than 3?



Oversigt

Distribution function or cumulative density function (cdf))

Definition

The distribution function(cdf) is the cumulated density function:

$$F(x) = P(X \leq x) = \sum_{j \text{ where } x_j \leq x} f(x_j)$$

Fair dice example

Let X represent one throw with a fair dice
Find the probability to get below 3:

$$
\begin{aligned}
P(X<3) &= P(X\leq 2) \\
&= F(2) \quad \textit{the distribution function} \\
&= P(X=1)+P(X=2) \\
&= f(1)+f(2) \quad \textit{the density function} \\
&= \frac{1}{6}+\frac{1}{6}=\frac{1}{3}
\end{aligned}
$$

Fair dice example

Find the probability to above or equal to 3:

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) \\ &= 1 - F(2) \text{ } \textit{the distribution function} \\ &= 1 - \frac{1}{3} = \frac{2}{3} \end{aligned}$$

Oversigt

Specific discrete distributions

- A number of statistical distributions exists that can be used to describe and analyse different kind of problems
- Today we consider discrete distributions:
  - The binomial distribution
  - The hypergeometric distribution
  - The Poisson distribution

The Binomial distribution

- An experiment with two outcomes (succes or failure) is repeated
- X is the number of successes after n repeats
- So X follows a binomial distribution

$$X \sim B(n,p)$$

- n number of repeats
- p the probability of success in each repeat

The density function for the binomial distribution:

The probability of x successes:

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

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Binomial distribution simulation

```
## Probability  of success
p <- 0.1
## Number of repeats
nRepeat <- 30
## Simulate Bernoulli experiment nRepeat times
tmp <- sample(c(0,1), size=nRepeat, prob=c(1-p,p), replace=TRUE)
## x is now
sum(tmp)

## Make similar with binomial distribution simulation function
rbinom(1, size=30, prob=p)

################
## Fair dice example

## Number of simulated realizations
n <- 30
## Sample independent from the set (1,2,3,4,5,6) with same probabilities
xFair <- sample(1:6, size=n, replace=TRUE)
## Count the number of 6'es
sum(xFair == 6)

## Make similar with rbinom()
rbinom(n=1, size=30, prob=1/6)
```

Example 1

In a call center in a phone company the costumer satisfaction is an issue. It is especially important that when errors/faults occur, then they are corrected within the same day.
Assume that the probability of an error being corrected within the same is $p = 0.7$.
Assume that the probability of an error being corrected within the same is $p = 0.7$.

- Step 1) What is the random variable: X is number of corrected errors
- Step 2) What dstribution: X follows The binomial distribution
- Step 3) What probability: $P(X = x) = f(x;n,p)$ $P(X = 6) = f(6;n,p)$
- Step 4)
  - What is the number of repeats? $n = 6$
  - What is the probability of success? $p = 0.7$

Example 1

What is the probability that 2 or less of the errors is corrected within the same day?

- Step 1) What is the random variable: X is number of corrected errors
- Step 2) What dstribution: X follows The binomial distribution
- Step 3) What probability: $P(X\ ?\ ?)$ $P(X \leq 2) = F(2; n, p)$
- Step 4)
  - What is the number of repeats? $n = 6$
  - What is the probability of success? $p = 0.7$

Oversigt

The hypergeometric distribution

- X is again the the number of successes, but now *WITHOUT replacement when repeating*
- X follows the hypergeometric distribution

$$X \sim H(n, a, N)$$

- n is the number of draws (repeats)
- a the number of successes in the population
- N is the number of elements in the (entire) population

The hypergeometric distribution

- The probability to get x successes is

$$f(x;n,a,N) = P(X=x) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}$$

- n is the number of draws (repeats)
- a the number of successes in the population
- N is the number of elements in the (entire) population

Example 2

In a shipment of 10 hard disks 2 of them have small scrathes.

A random sample of 3 hard disks is taken. What is the probability that at least 1 of them has scratches?

- Step 1) What is the random variable: X is number with scratches
- Step 2) What distribution: X follows the hypergeometric distribution
- Step 3) What probability:

  $P(X\ ?\ ?)\ P(X \geq 1) = 1 - P(X = 0) = 1 - f(0; n, a, N)$
- Step 4)
  - What is number of draws? $n = 3$
  - How many successes is there? $a = 2$
  - How many disks all together? $N = 10$

Binomial vs. hypergeometric

- The binomial distribution is also used to analyse samples with replacement
- The hypergeometric distribution is used to analyse samples without replacement

Oversigt

The Poisson distribution

- The Poisson distribution is often use as distribution (model) for counts which do not have a natural upper bound
- The Poisson distribution is often characterized as intensity, that is on the form number/unit
- The parameter λ gives the gives the intensity in the Poisson distribution

The Poisson distribution

$$X \sim P(\lambda)$$

The density function:

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

The distribution function:

$$F(x) = P(X \leq x)$$

Example 3.1

Assume that on average 0.3 patients per day are put in hospital in Copenhagen due to air pollution.

What is the probability that at most two patients are put in hospital in Copenhagen due to air pollution on a given day?

- Step 1) What is the random variable: X is the number of patients on a day
- Step 2) What distribution: X follows the Poisson distribution
- Step 3) What probability: $P(X\ ?\ ?)P(X \leq 2)$
- Step 4) What is the intensity: $\lambda = 0.3$ patients per day

Example 3.2

Assume that on average 0.3 patients per day are put in hospital in Copenhagen due to air pollution.

What is the probability that exactly two patients are put in hospital in Copenhagen due to air pollution on a given day?

- Step 3) What probability: $P(X\ ?\ ?)P(X = 2)$

Example 3.3

Assume that on average 0.3 patients per day are put in hospital in Copenhagen due to air pollution.

What is the probability that at least 2 patients are put in hospital in Copenhagen due to air pollution on a given day?

- Step 3) What probability:
  $P(X \; ? \; ?)P(X \geq 2) = 1 - P(X \leq 1)$

Example 3.4

What is the probability that exactly 1 patient is put in hospital in Copenhagen due to air pollution within 3 days?

- Step 1) What is the random variable:
  - From X number per day
  - To $X^{3\text{days}}$ which is *patients per 3 days*
- Step 2) What distribution has $X^{3\text{days}}$: The Poisson distribution
- Step 3) What probability: $P(X^{3\text{days}} = 1)$
- Step 4) Scale the intensity
  - From $\lambda = 0.3$ patientes/day to $\lambda_{3days} = 0.9$ patients/3days

Oversigt

R	Name
`binom`	Binomial
`hyper`	Hypergeometric
`pois`	Poisson

- `d` $f(x)$ (probability density function).
- `p` $F(x)$ (cumulative distribution function).
- `r` Random numbers from the distribution
- `q` quantiles (the inverse of $F(x)$)

Remember that function help etc. is achieved by putting '?' in front of the name.

Example binomial distribution: $P(X \leq 5) = F(5; 10, 0.6)$

```
pbinom(q=5, size=10, prob=0.6)
## Get the hep with
?pbinom
```

Oversigt

Mean (Expected value))

Mean of discrete random variable:

$$\mu = E(X) = \sum_{\text{all x}} xf(x)$$

- The "correct mean"
- Expresses the "center" of X

Example: Mean of a dice throw

$$\begin{aligned}\mu = E(X) &= \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= 3.5\end{aligned}$$

Link to sample mean - simulation learning

```
## NUmber of simulated realizations
n <- 30
## Sample independently from the set (1,2,3,4,5,6) with
## equal probability
xFair <- sample(1:6, size=n, replace=TRUE)

## Find the sample mean
mean(xFair)
```

The more observations, the close you get to the right mean (expected value)

$$\lim_{n \to \infty} \hat{\mu} = \mu$$

- Try it in R

Variance

Definition

$$\sigma^2 = \text{Var}(X) = \sum_{\text{all x}} (x-\mu)^2 f(x)$$

- Measures average spread
- The "correct standard deviation" of X (as opposed to sample variance))

Variance, example

Variance of dice throw

$$\begin{aligned}\sigma^2 = E[(X-\mu)^2] = & \\ &= (1-3.5)^2 \cdot \frac{1}{6} + (2-3.5)^2 \cdot \frac{1}{6} + (3-3.5)^2 \cdot \frac{1}{6} \\ &\quad + (4-3.5)^2 \cdot \frac{1}{6} + (5-3.5)^2 \cdot \frac{1}{6} + (6-3.5)^2 \cdot \frac{1}{6} \\ &\approx 2.92\end{aligned}$$

Link to sample variance - simulation learning

```
  ## NUmber of simulated realizations
  n <- 30
  ## Sample independently from the set (1,2,3,4,5,6) with
  ## equal probability
  xFair <- sample(1:6, size=n, replace=TRUE)

  ## Find the sample variance
var(xFair)
```

Mean and variances for specific discrete distributions

The binomial distribution:

- Mean:
  $\mu = n \cdot p$
- Variance:
  $\sigma^2 = n \cdot p \cdot (1 - p)$

Mean and variances for specific discrete distributions

The hypergeometric distribution:

- Mean:
  $\mu = n \cdot \frac{a}{N}$
- Variance:
  $\sigma^2 = \frac{na \cdot (N-a) \cdot (N-n)}{N^2 \cdot (N-1)}$

Mean and variances for specific discrete distributions

The poisson distribution:

- Mean:
  $\mu = \lambda$
- Variance:
  $\sigma^2 = \lambda$

Agenda

1. Random Variables and the density function
2. Distribution function
3. Specific discrete distributions I: The binomial
   - Example 1
4. Specific distributions II: The hypergeometric
   - Example 2
5. Specific distributions III: The Poisson
   - Example 3
6. Distributions in R
7. Mean and Variance
   - Mean and variances for specific discrete distributions