Course 02402 Introduction to Statistics

Lecture 2: Random variables and discrete distributions

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Agenda

1. Random variables and density functions
2. Distribution functions
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 of discrete distributions

Overview

Random variables

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

Random variables

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

- A throw of a dice.
- The number of six'es in ten dice throws.
- Fuel consumption of a car.
- Measurement of glucose level in blood sample.
- ...

Discrete and continuous random variables

- We distinguish between *discrete* and *continuous* random variables.
- Discrete:
  - Number of people in this room who wear glasses.
  - Number of planes departing from CPH within the next hour.
- Continuous:
  - Wind speed measurement.
  - Transport time to DTU.
- Today: Discrete. Next week: Continuous.

Random variable

Before the experiment is carried out, we have a random variable

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

indicated with capital letters.

Random variable

Before the experiment is carried out, we have a random variable

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

indicated with capital letters.

After the experiment is carried out, we have a *realization* or *observation*

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

indicated with lowercase letters.

Simulate rolling a dice in R

```r
# One random draw from (1,2,3,4,5,6)
# with equal probability for each outcome
sample(1:6, size = 1)
```

[1] 1

Discrete distributions

- Random variables are used to describe an experiment before it is carried out.
- How to do this without yet knowing the outcome?

Discrete distributions

- Random variables are used to describe an experiment before it is carried out.
- How to do this without yet knowing the outcome?
- Solution: Use a *density function*.

Density function, discrete random variable: Definition 2.6

The *density function* (probability density function, pdf) of a discrete random variable:

Definition

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

Describes the probability that X takes the value x when the experiment is carried out.

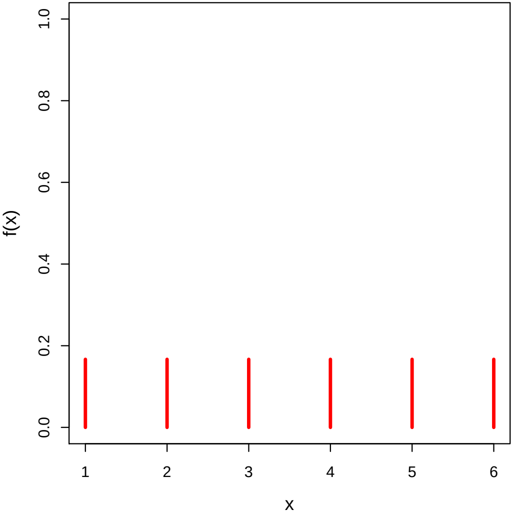
Density function, discrete random variable, Definition 2.6

The density function of a discrete random variable satisfies two properties:

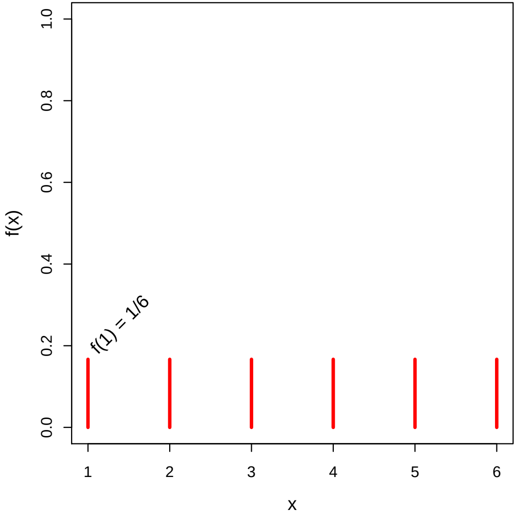
Definition

$$f(x) \geq 0 \text{ for all } x \quad \text{and} \quad \sum_{\text{all } x} f(x) = 1$$

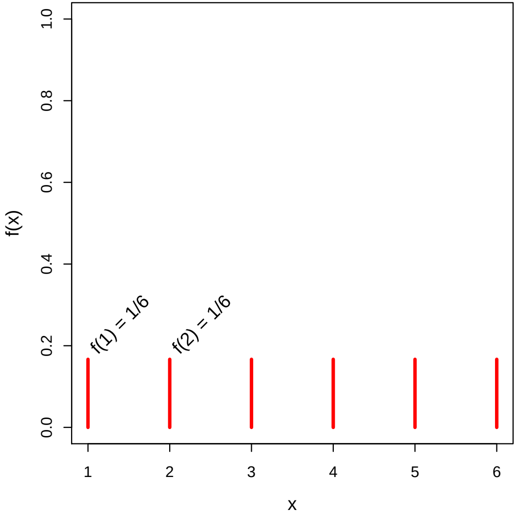
Density function for a fair dice



Density function for a fair dice



Density function for a fair dice



Sample

If we only have a single observation, can we see the distribution?

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*

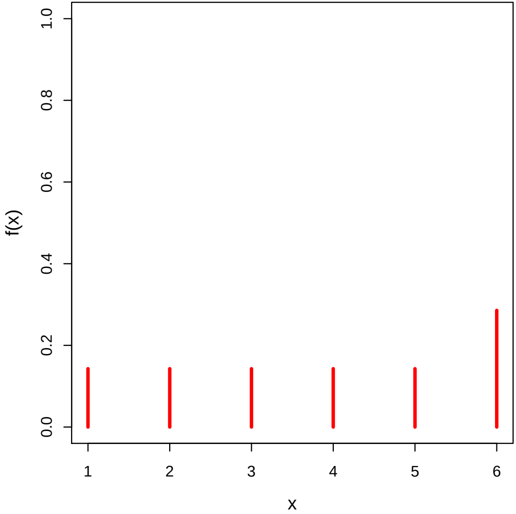
$$\{x_1, x_2, \ldots, x_n\}$$

and we can begin to get an idea of the distribution.

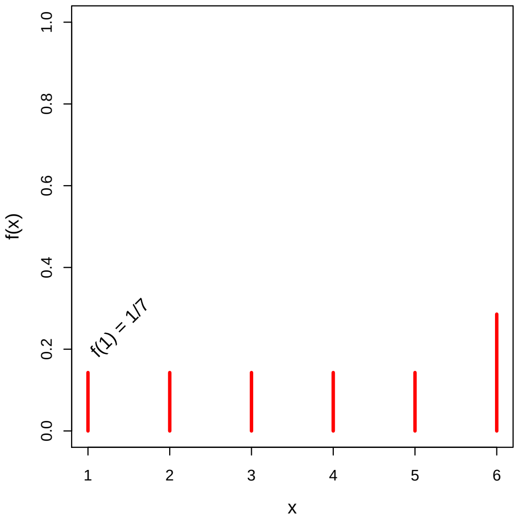
Simulation of n rolls with a fair dice

```r
# Number of simulated realizations (sample size)
n <- 30

# n independent random draws from the set (1,2,3,4,5,6)
# with equal probability of each outcome
xFair <- sample(1:6, size = n, replace = TRUE)
xFair

# Count 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 f(x)")
# Add the true 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","True pdf"), lty = 1, col = c(1,2),
       lwd = c(5, 2), cex = 0.8)
```

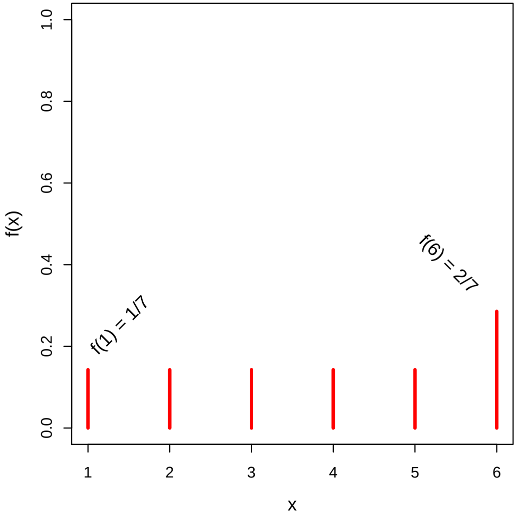
Density function for an unfair dice



Density function for an unfair dice



Density function for an unfair dice



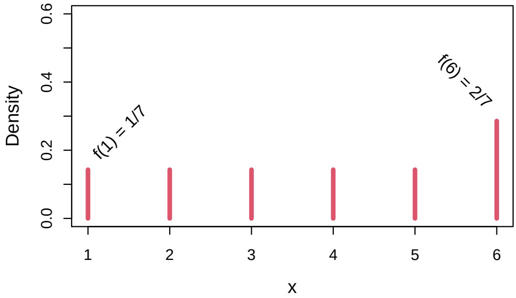
Simulation of n rolls with an unfair dice

```r
# Number of simulated realizations (sample size)
n <- 30

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

# Plot the empirical pdf
plot(table(xUnfair)/n, lwd = 10, ylim = c(0,1), xlab = "x",
     ylab = "Density f(x)")
# Add the true pdf to the plot
lines(c(rep(1/7,5),2/7), lwd = 4, type = "h", col = 2)
# Add a legend to the plot
legend("topright", c("Empirical pdf","True pdf"), lty = 1, col = c(1,2),
       lwd = c(5, 2), cex = 0.8)
```

Some questions

Let X describe one throw with the *unfair* dice. What is:

- The probability of getting a 4?
- The probability of getting a 5 or a 6?
- The probability of getting less than 3?



Overview

Distribution function, discrete random variable: Definition 2.9

The *distribution function* (cumulative distribution function, cdf) of a discrete random variable:

Definition

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

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 3:

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 3:

$$P(X < 3)$$

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 3:

$$P(X < 3) = P(X \leq 2)$$

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 3:

$$\begin{aligned} P(X < 3) &= P(X \leq 2) \\ &= F(2) \quad \textit{the distribution function} \end{aligned}$$

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 3:

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

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 3:

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

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 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 of throwing greater than or equal to 3:

Fair dice example

Find the probability of throwing greater than or equal to 3:

$$P(X \geq 3)$$

Fair dice example

Find the probability of throwing greater than or equal to 3:

$$P(X \geq 3) = 1 - P(X \leq 2)$$

Fair dice example

Find the probability of throwing greater than or equal to 3:

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

Fair dice example

Find the probability of throwing greater than or equal to 3:

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

Overview

Specific discrete distributions

- A number of different statistical distributions exist, which may be used to describe and analyse different types of problems.
- Today, we consider only discrete distributions:
  - The binomial distribution
  - The hypergeometric distribution
  - The Poisson distribution

The Binomial distribution

- An experiment with two outcomes, "success" or "failure", is repeated (independent repetitions).
- X is the number of successes after n repetitions.

The Binomial distribution

- An experiment with two outcomes, "success" or "failure", is repeated (independent repetitions).
- X is the number of successes after n repetitions.
- Then X follows a binomial distribution:

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

  - n: number of repetitions
  - p: probability of success in each repetition

The density function of 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

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

Example: Binomial distribution

Suppose that $X \sim B(4, p)$, i.e. $n = 4$. Find the probability of 3 successes.

Example: Binomial distribution

Suppose that $X \sim B(4, p)$, i.e. $n = 4$. Find the probability of 3 successes.

- Probability of 3 successes: $P(X = 3)$.

Example: Binomial distribution

Suppose that $X \sim B(4, p)$, i.e. $n = 4$. Find the probability of 3 successes.

- Probability of 3 successes: $P(X = 3)$.
- Three successes can be obtained in four "ways":
  SSSF, SSFS, SFSS, FSSS.

Example: Binomial distribution

Suppose that $X \sim B(4,p)$, i.e. $n = 4$. Find the probability of 3 successes.

- Probability of 3 successes: $P(X = 3)$.
- Three successes can be obtained in four "ways":
  SSSF, SSFS, SFSS, FSSS.
- Thus,

$$\binom{n}{x} = \binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 1} = 4,$$

and

$$P(X = 3) = 4p^3(1-p).$$

Simulation from a binomial distribution

```r
## Probability of success
p <- 0.1

## Number of repetitions
nRepeat <- 30

## Simulate Bernoulli experiment 'nRepeat' times
tmp <- sample(c(0,1), size = nRepeat, prob = c(1-p,p), replace = TRUE)

# Compute 'x'
sum(tmp)

## Or: Use the binomial distribution simulation function
rbinom(1, size = 30, prob = p)
```

Example: Fair dice

```r
# Number of simulated realizations (sample size)
n <- 30

# n independent random draws from the set (1,2,3,4,5,6)
# with equal probability for each outcome
xFair <- sample(1:6, size = n, replace = TRUE)

# Count the number of six'es
sum(xFair == 6)

## Do the same using 'rbinom()' instead
rbinom(n = 1, size = 30, prob = 1/6)
```

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

- Step 1) What should be described by a random variable X?

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

- Step 1) What should be described by a random variable X?

  The number of corrected errors.

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

- Step 1) What should be described by a random variable X?

  The number of corrected errors.

- Step 2) What is the distribution of X?

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

- Step 1) What should be described by a random variable X?

  The number of corrected errors.

- Step 2) What is the distribution of X?

  A binomial distribution with $n = 6$ and $p = 0.7$.

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

- Step 3) Which probability should be computed?

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

- Step 3) Which probability should be computed?

  $\underline{P(X=6)=f(6;6,0.7)}$

Overview

The hypergeometric distribution

- Again, X is the the number of successes, but now *without* replacement when repeating.

The hypergeometric distribution

- Again, X is 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 (repetitions)
  - a is the number of successes in the population
  - N is the number of elements in the (entire) population

The hypergeometric distribution

- The probability of getting 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 (repetitions)
  - a is the number of successes in the population
  - N is the number of elements in the (entire) population

Example 2

In a shipment of 10 harddisks, 2 of them have small scratches.

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

Example 2

In a shipment of 10 harddisks, 2 of them have small scratches.

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

- Step 1) What should be described by a random variable X?

Example 2

In a shipment of 10 harddisks, 2 of them have small scratches.

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

- Step 1) What should be described by a random variable X?

  Number of harddisks with scratches in the random sample.

Example 2

In a shipment of 10 harddisks, 2 of them have small scratches.

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

- Step 1) What should be described by a random variable X?
  Number of harddisks with scratches in the random sample.
- Step 2) What is the distribution of X?

Example 2

In a shipment of 10 harddisks, 2 of them have small scratches.

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

- Step 1) What should be described by a random variable X?

  Number of harddisks with scratches in the random sample.

- Step 2) What is the distribution of X?

  A hypergeometric distribution with $n = 3$, $a = 2$, $N = 10$.

- Step 3) Which probability should be computed?

Example 2

In a shipment of 10 harddisks, 2 of them have small scratches.

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

- Step 1) What should be described by a random variable X?

  Number of harddisks with scratches in the random sample.

- Step 2) What is the distribution of X?

  A hypergeometric distribution with $n = 3$, $a = 2$, $N = 10$.

- Step 3) Which probability should be computed?

  $P(X \geq 1) = 1 - P(X = 0) = 1 - f(0; 3, 2, 10)$

Binomial vs. hypergeometric

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

Overview

The Poisson distribution

- The Poisson distribution is often used as a distribution (model) for counts, which do not have a natural upper bound.
- The Poisson distribution is often characterized by its *intensity*, which is on the form "number/unit", and often denoted λ.

The Poisson distribution

$$X \sim Po(\lambda)$$

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

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

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

Example 3

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

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

- Step 1) What should be described by a random variable X?

Example 3

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

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

- Step 1) What should be described by a random variable X?

  The number of patients on a given day.

Example 3

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

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

- Step 1) What should be described by a random variable X?

  The number of patients on a given day.

- Step 2) What is the distribution of X?

Example 3

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

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

- Step 1) What should be described by a random variable X?

  The number of patients on a given day.

- Step 2) What is the distribution of X?

  A Poisson distribution with $\lambda = 0.3$.

Example 3

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

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

- Step 1) What should be described by a random variable X?

  The number of patients on a given day.

- Step 2) What is the distribution of X?

  A Poisson distribution with $\lambda = 0.3$.

- Step 3) Which probability should be computed?

Example 3

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

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

- Step 1) What should be described by a random variable X?

  The number of patients on a given day.

- Step 2) What is the distribution of X?

  A Poisson distribution with $\lambda = 0.3$.

- Step 3) Which probability should be computed?

  $P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$

Overview

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 of the distribution ("inverse" of $F(x)$)

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

```
pbinom(q = 5, size = 10, prob = 0.6)
```

[1] 0.37

```
# Get help with:
?pbinom
```

Overview

Mean (expectation, expected value)

Mean of a discrete random variable, Definition 2.13:

Definition

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

- The *"true mean"* of X (as opposed to the sample mean).
- Expresses the "center" of the distribution of X.

Example: Mean of a throw with a fair dice

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

Link to sample mean - learning from simulations

```r
# Number of simulated realizations (sample size)
n <- 30

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

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

```
[1] 3.3
```

Asymptotics, increasing the sample size

The more observations (the larger the sample size), the closer you get to the true mean:

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

- Try increasing n in the simulations in R.

Variance

Variance of a discrete random variable, Definition 2.16:

Definition

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

- Measures average dispersion/spread.
- The "true variance" of X (as opposed to the sample variance).

Example: Variance of a throw with a fair dice

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

Link to sample variance - learning from simulations

```r
# Number of simulated realizations (sample size)
n <- 30

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

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

[1] 2.4

Mean and variance of specific discrete distributions

The binomial distribution

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

Mean and variance of specific discrete distributions

The hypergeometric distribution

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

Mean and variance of specific discrete distributions

The poisson distribution

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

Agenda

1. Random variables and density functions
2. Distribution functions
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 of discrete distributions