Bus 701: Advanced Statistics

Harald Schmidbauer

İSTANBUL BİLGİ ÜNİVERSİTESİ

About These Slides

- The present slides are not self-contained; they need to be explained and discussed. They contain only a small part of the course Bus 701.

- Even though being a "work in progress" and subject to revision, the slides constitute copyrighted material.
  If you want to reproduce or copy anything from the slides, please ask:

Harald Schmidbauer	**harald** at **hs-stat** dot **com**
Angi Rösch	**angi.r** at **t-online** dot **de**

- The slides were produced using LaTeX and R (the R project; www.R-project.org) on a Linux system.

- R files used for this course are available upon request.

Chapter 7:

Discrete Probability Distributions

7.1 Introduction

Discrete random variables and discrete distributions.

- A random variable is called discrete if it can take on only isolated values (for simplicity: 0, 1, 2, ...)
- The distribution of a discrete random variable is called a discrete distribution.

Example:

- X = result of throwing a die once
- The distribution of X is

$$P(X = i) = 1/6 \quad \text{for } i = 1, \ldots, 6.$$

7.1 Introduction

Example:

- Random variable:

$$X = \#\text{ successes in } n \text{ independent trials, where the probability of success is } p \text{ in each trial}$$

- Distribution of X:

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i} \quad \text{for } i = 0, \ldots, n.$$

This is the binomial distribution with parameters n and p.
In short, we write: $X \sim \mathrm{B}(n, p)$

7.1 Introduction

Example:

- Random variable:

$$X = \# \text{ number of customers calling a call center Friday evening between 6 and 7 p.m.}$$

- Distribution of X:

$$P(X = i) = \textbf{???}$$

7.1 Introduction

The probability function.

- A discrete probability distribution is given by the terms

$$p_i = P(X = i).$$

- As in the case of a discrete variable in Chapters 3 and 4, we can
  - display the distribution,
  - compute location and variation measures.

7.1 Introduction

The expectation: a location measure.

- For a discrete random variable X,

$$\mathrm{E}(X) = \sum_i i \cdot P(X = i).$$

- Same principle as arithmetic mean, with probabilities $p_i = P(X = i)$ substituted for relative frequencies f_i.

7.1 Introduction

The variance: a variation measure.

- For a discrete random variable X,

$$\begin{aligned}\mathrm{var}(X) &= \sum_i (i - \mathrm{E}(X))^2 \cdot P(X = i) \\ &= \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] = \mathrm{E}(X^2) - \mathrm{E}^2(X) \\ &= \sum_i i^2 \cdot P(X = i) - \left(\sum_i i \cdot P(X = i)\right)^2 .\end{aligned}$$

- Same principle as (empirical) variance, with probabilities $p_i = P(X = i)$ substituted for the relative frequencies f_i.

7.1 Introduction

Example 1: A Bernoulli experiment.

- Consider the random variable X with

$$P(X = 1) = p, \quad P(X = 0) = 1 - p$$

- Expectation and variance are:

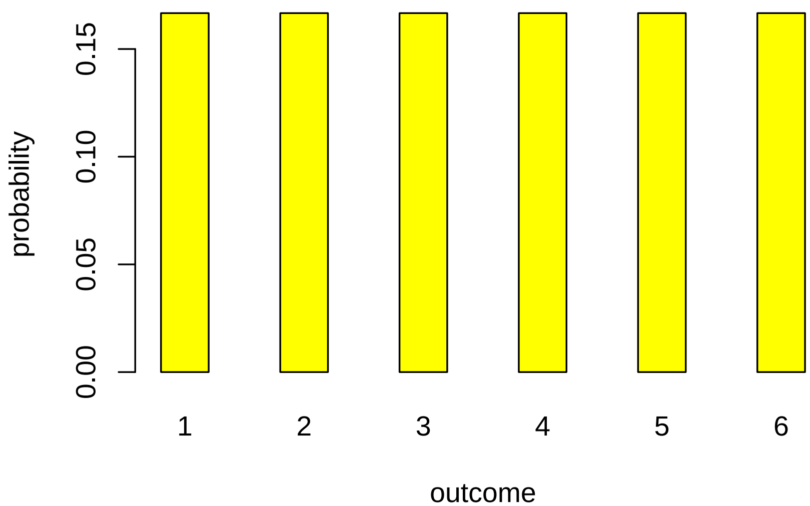
$$\begin{array}{lclcl} \mathrm{E}(X) & = 1 \cdot p + 0 \cdot (1-p) & = & p, \\ \mathrm{var}(X) & = 1^2 \cdot p + 0^2 \cdot (1-p) - p^2 & = & p(1-p). \end{array}$$

- Here, we used: $\mathrm{var}(X) = \mathrm{E}(X^2) \; - \; \mathrm{E}^2(X)$

7.1 Introduction

Example 2: Throwing a die once.

- Let X = result of throwing a die once
- A bar chart of the distribution:



$$\begin{aligned} \mathrm{E}(X) &= 3.5 \\ \mathrm{var}(X) &= 2.92 \end{aligned}$$

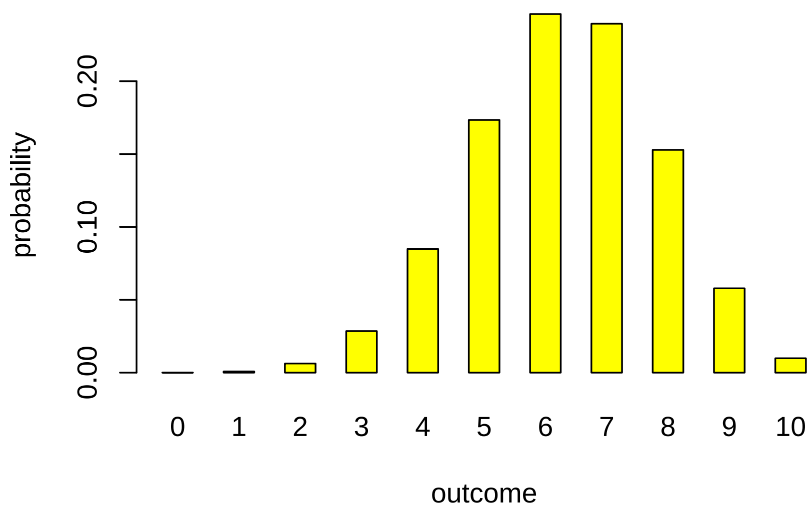
7.1 Introduction

Example 3: Bernoulli trials.

- Let $X \sim \mathsf{B}(10, 0.63)$, that is:

$$X = \# \text{ successes in 10 independent trials, success probability in each trial: 0.63}$$

- A bar chart of the distribution:



$$E(X) = 6.3$$
$$\mathrm{var}(X) = 2.33$$

7.2 The Binomial Distribution

Some properties of the binomial distribution.

- Let $X \sim \mathsf{B}(n, p)$, that is:.

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i} \quad \text{for } i = 0, \ldots, n.$$

- Then:

$$\begin{aligned} \mathrm{E}(X) &= n \cdot p \\ \mathrm{var}(X) &= n \cdot p \cdot (1-p) \end{aligned}$$

7.2 The Binomial Distribution

Example 1: A public opinion poll.

- Q: "Do you think New Orleans should be rebuilt?"
- Define: $p =$ share of American adults who say "YES"
- **Assume** $p = 0.63$. (We will never know if this is true.)
- Suppose we select 10 people randomly.
- Let $X = \#$ of those who answer "YES" among the 10.
  Then, $X \sim \mathrm{B}(10, 0.63)$.

7.2 The Binomial Distribution

Example 1: A public opinion poll.

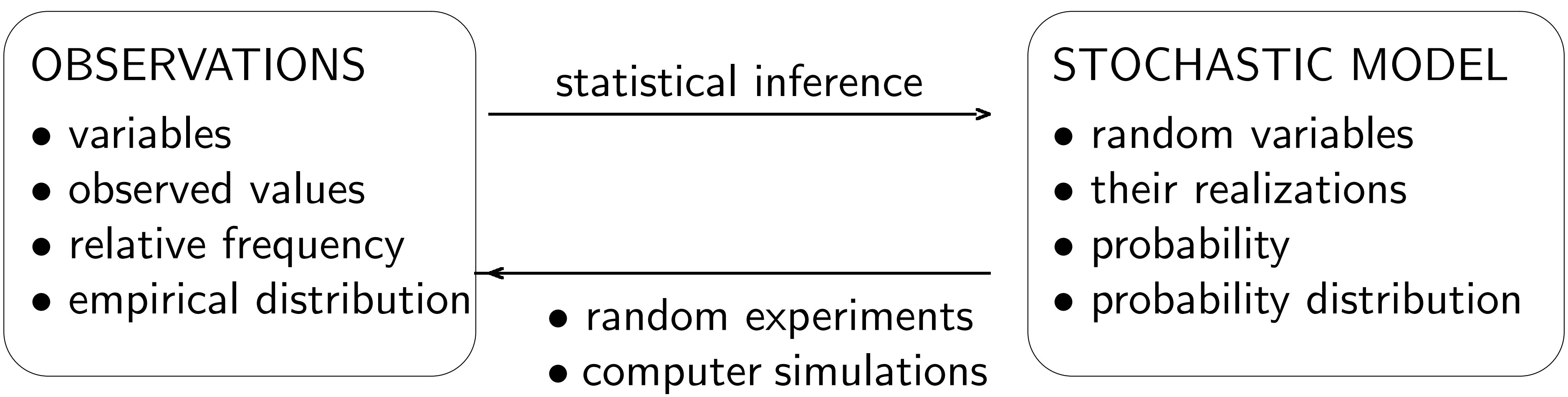
- What is the probability that there are at least 8 in the sample of 10 who say "YES"?
- Since $X \sim \mathrm{B}(10, 0.63)$:

$$
\begin{array}{lcll}
P(X=8) & = & \binom{10}{8}(0.63)^{8}(0.37)^{2} & = 0.153 \\
P(X=9) & = & \binom{10}{9}(0.63)^{9}(0.37)^{1} & = 0.058 \\
P(X=10) & = & \binom{10}{10}(0.63)^{10}(0.37)^{0} & = 0.010 \\
\hline
P(X \geq 8) & = & \sum_{i=8}^{10} \binom{10}{i}(0.63)^{i}(0.37)^{10-i} & = 0.22
\end{array}
$$

7.2 The Binomial Distribution

Example 1: A public opinion poll.

- But in many real-world applications, we won't know the parameter p?!
- This is true, but let's not forget where we are!



7.2 The Binomial Distribution

Example 2: Success days in stock indices.

- Consider a daily series of stock index returns.
- We look at "windows" of length n.
- Define

$$X = \# \text{ days in an } n\text{-day window with positive returns}$$

- If a certain random walk hypothesis holds: $X \sim \mathrm{B}(n, p)$
- Is this true for real stock markets?

7.2 The Binomial Distribution

Example 2: Success days in stock indices.

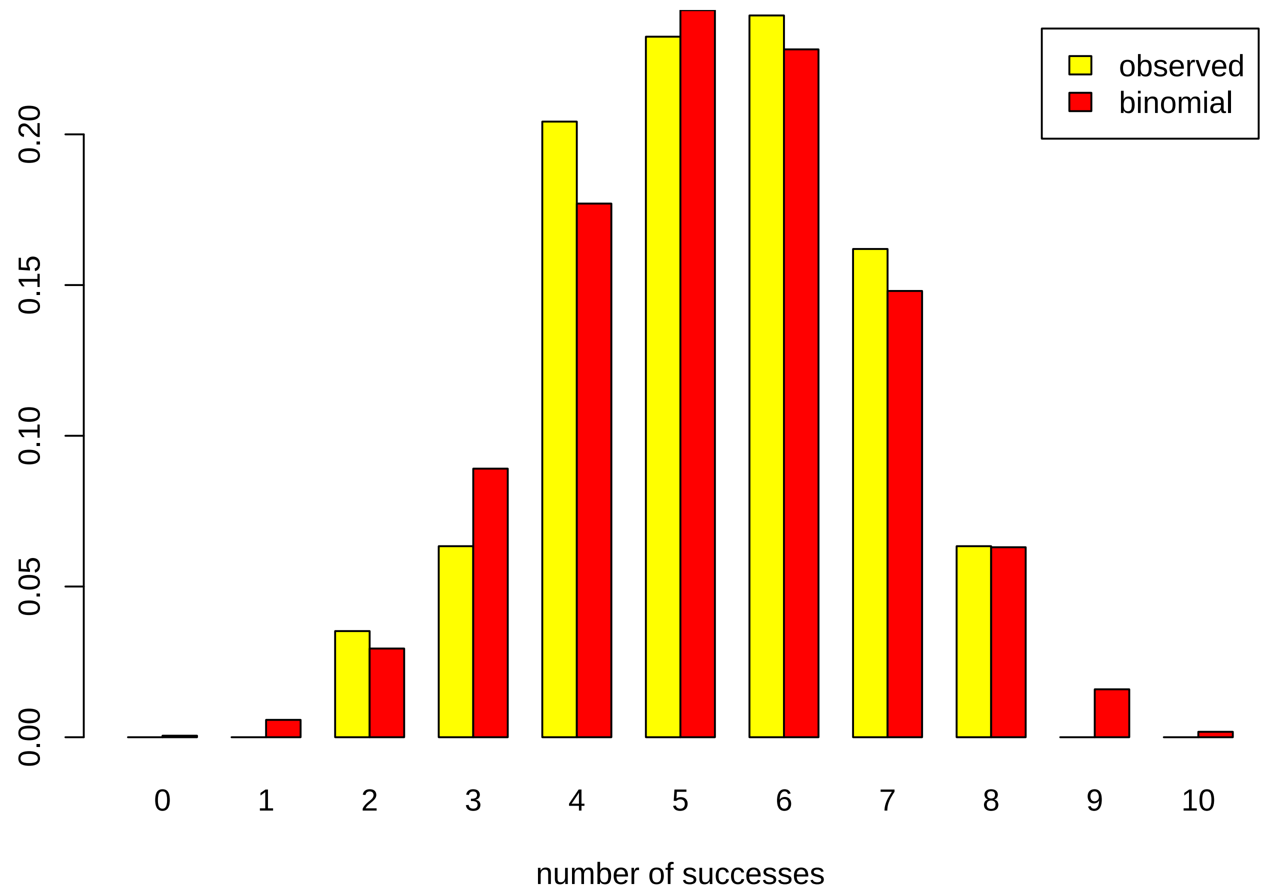
- For İMKB 100, from 2001-01-02 to 2006-09-18 (1427 days) with $n = 10$:

window	indicators	X
1	1, 0, 1, 0, 1, 1, 1, 0, 0, 0	5
2	1, 1, 0, 1, 0, 0, 1, 1, 0, 0	5
3	0, 1, 0, 0, 0, 1, 0, 0, 0, 0	2
4	1, 1, 0, 1, 0, 1, 0, 1, 1, 1	7
⋮	⋮	⋮

- There are 142 windows of length $n = 10$ days in this period of time.

7.2 The Binomial Distribution

Example 2: Success days in stock indices.



7.3 The Hypergeometric Distribution

Derivation of the hypergeometric distribution.

- Urn: N balls, M red, $N - M$ blue.
- We draw n balls randomly without replacement. Let

  X = number of red balls among the n drawn balls.

- Then, for $i = \max(0, n + M - N), \ldots, \min(M, n)$:

$$P(X = i) = \frac{\binom{M}{i} \cdot \binom{N-M}{n-i}}{\binom{N}{n}}$$

7.3 The Hypergeometric Distribution

Properties of the hypergeometric distribution.

- It holds that:

$$\mathrm{E}(X) = n \cdot \frac{M}{N},$$

$$\mathrm{var}(X) = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \left(\frac{N-n}{N-1}\right).$$

- Compare this with the binomial distribution!

7.3 The Hypergeometric Distribution

Example: Sayısal loto 6/49.

- $N = 49$ balls
- 6 balls are red (those you ticked)
- 43 balls are blue (those you didn't tick)
- Let $X =$ number of hits in a single game. Then,

$$P(X = i) = \frac{\binom{6}{i} \cdot \binom{43}{6-i}}{\binom{49}{6}}, \quad i = 0, \ldots, 6.$$

7.3 The Hypergeometric Distribution

Example: Sayısal loto 6/49.

These probabilities are:

i	$P(X = i)$
0	0.4360
1	0.4130
2	0.1324
3	0.01765
4	0.0009686
5	0.00001845
6	0.00000007151

7.4 The Poisson Distribution

Derivation of the Poisson distribution.

- Poisson's limit theorem (Siméon-Denis Poisson, 1781–1840):

$$\lim_{\substack{n\to\infty \\ p\to 0 \\ n\cdot p\to\lambda}} \binom{n}{i} p^i (1-p)^{n-i} = \frac{\lambda^i}{i!}\mathrm{e}^{-\lambda}$$

- A random variable X with

$$P(X=i) = \frac{\lambda^i}{i!}\mathrm{e}^{-\lambda} \quad \text{for } i = 0, 1, 2, \ldots, \quad \lambda > 0$$

  is said to have a Poisson distribution with parameter λ.

- In short, we write: $X \sim \mathrm{Po}(\lambda)$.

7.4 The Poisson Distribution

Properties of the Poisson distribution.

- If $X \sim \text{Po}(\lambda)$:

$$\begin{aligned} \mathrm{E}(X) &= \lambda, \\ \mathrm{var}(X) &= \lambda. \end{aligned}$$

- This can be seen as a consequence of Poisson's limit theorem!

7.4 The Poisson Distribution

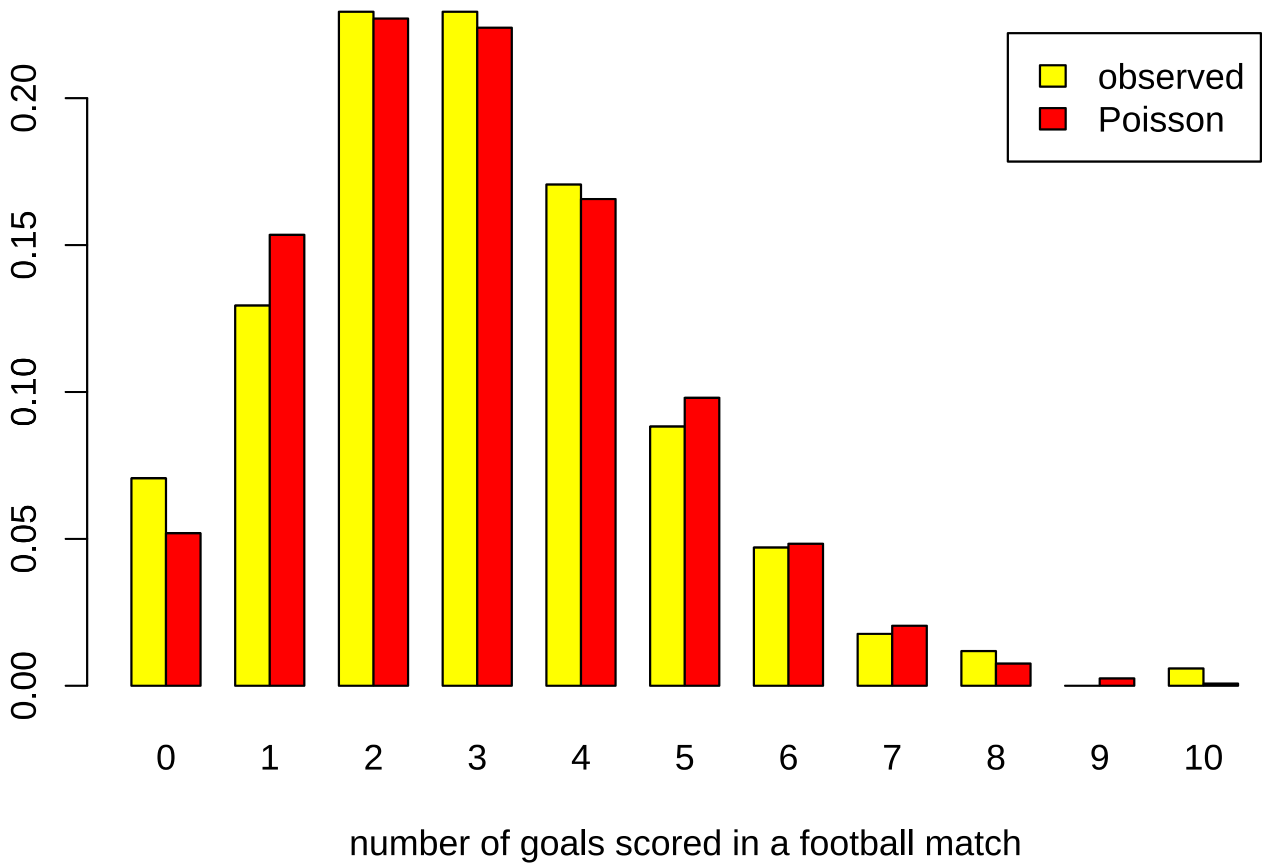
The Poisson distribution.

The following random variables may have a Poisson distribution:

- # of typos on the page of a newspaper
- # of customer arrivals at a supermarket between 10:00 and 10:15 on a typical day
- # of traffic accidents in a town on a typical day
- # of goals in a football match
- many others which count the number of events, where an event may happen at any time but is unlikely to happen in a given short time interval

7.4 The Poisson Distribution

Example: Number of goals scored in matches of Beşiktaş.



7.5 Benford's Law

Distribution of the first digit.

- Consider a random variable X with distribution

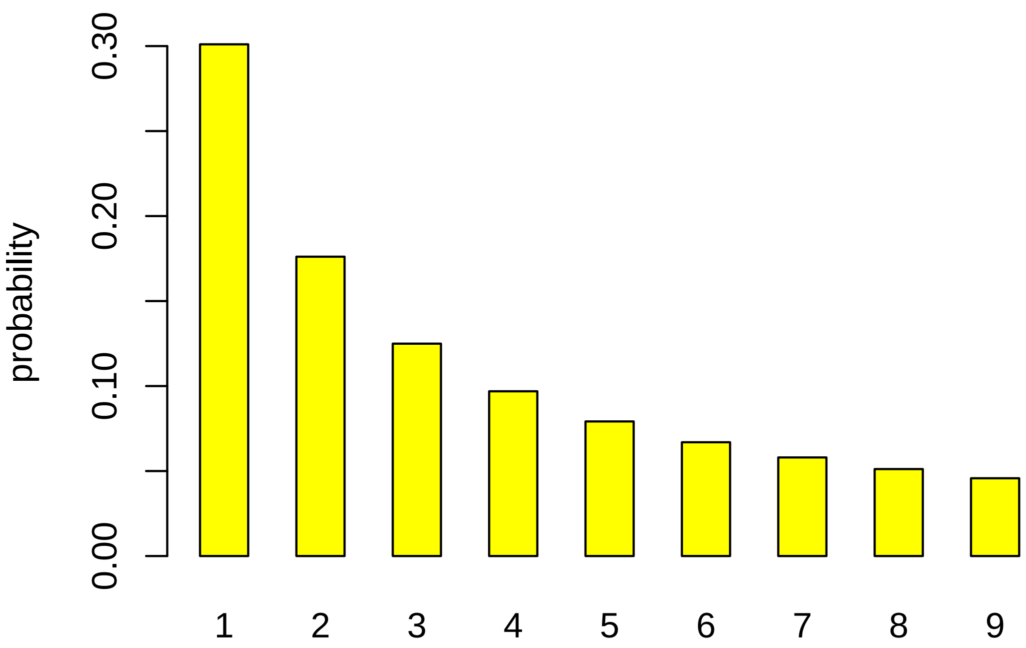
$$P(X = i) = \log_{10}(i + 1) - \log_{10}(i), \quad i = 1, 2, \ldots, 9.$$

- Benford's law (the first-digit law):

  In lists of numbers from many real-life sources of data, the leading digit is distributed according to this distribution.

7.5 Benford's Law

The probability function.



i	$P(X = i)$
1	0.301
2	0.176
3	0.125
4	0.097
5	0.079
6	0.067
7	0.058
8	0.051
9	0.046