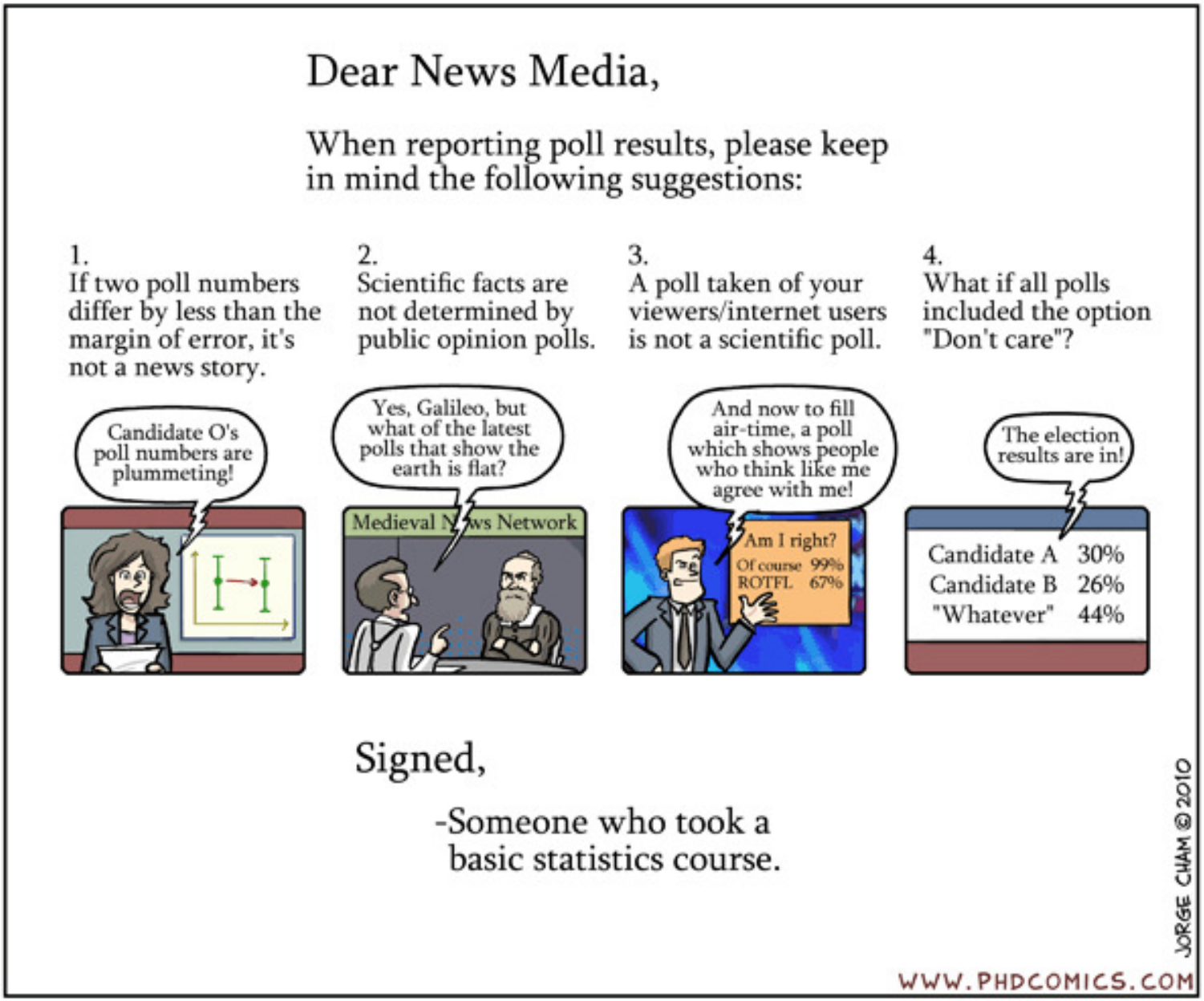
# Introduction to Statistics

## Notion of Statistics and Probability

Alexandra Posekany
alexandra.posekany@gmail.com


Dear News Media,
When reporting poll results, please keep in mind the following suggestions:
1.
If two poll numbers differ by less than the margin of error, it's not a news story.
2.
Scientific facts are not determined by public opinion polls.
3.
A poll taken of your viewers/internet users is not a scientific poll.
4.
What if all polls included the option "Don't care"?
Candidate O's poll numbers are plummeting!
Yes, Galileo, but what of the latest polls that show the earth is flat?
Medieval News Network
And now to fill air-time, a poll which shows people who think like me agree with me!
Am I right?
Of course 99%
ROTFL 67%
The election results are in!
Candidate A 30%
Candidate B 26%
"Whatever" 44%
Signed,
-Someone who took a basic statistics course.
JORGE CHAM © 2010
WWW.PHDCOMICS.COM

# Statistics - better than its reputation?

There are three kinds of lies:
lies, damned lies, and statistics.

[Mark Twain (referring to Benjamin Disraeli)]

Statistics can be used to support anything –
especially statisticians.

[Franklin P. Jones]

And thirdly, the code is more what you'd call
"guidelines" than actual rules.

[Barbossa (Pirates of the Carribbean)]

# What is 'statistics' actually?

# What do we need statistics for?

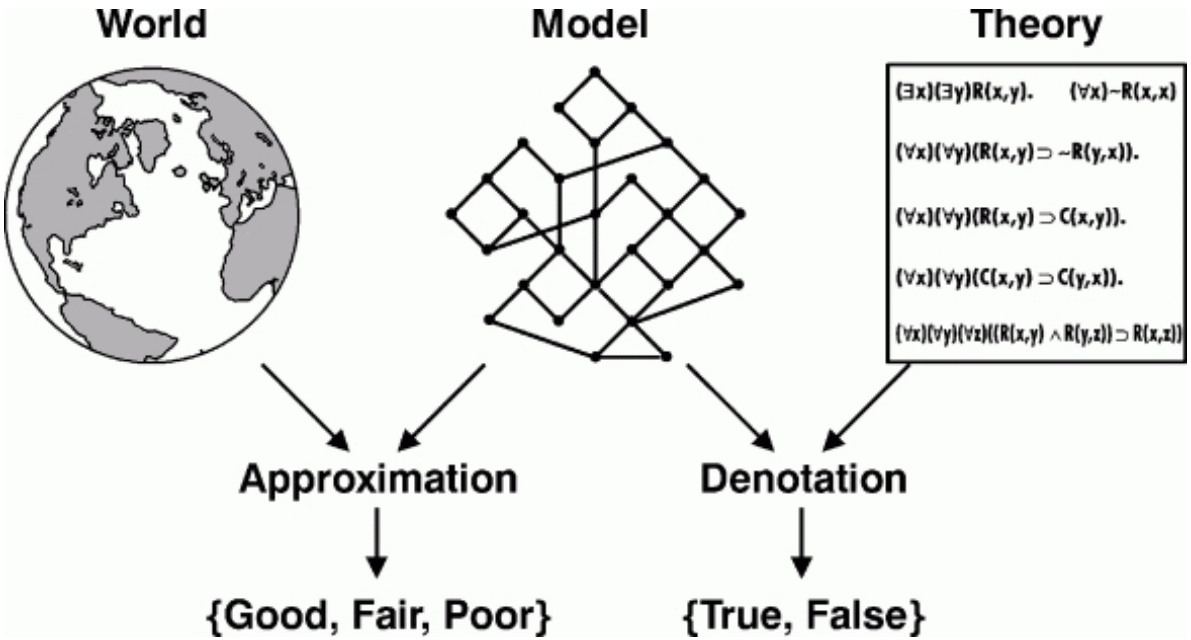
Statistics is the art/science of qualitative and quantitative assessment of data which can be described by "summary statistics"
How to answer the question for the differences between samples and goals correctly?
How to show that a scientific or technical hypothesis is valid?

The answers depend on the data, the exact formulation of the question etc., statistics provides a set of tools for dealing with such questions in a *reproducible* way.

# Moral of the story

Essentially, all models are wrong,
but some are useful.

[George Box]



# Goals of Statistics

- **Descriptive** Statistics: Organising and summarising data
  exploratory data analysis, using plots, summary statistics, sample estimators
- **Inferential** Statistics: Analysing the data
  answering specific questions, i.e. testing hypotheses, building a model for prediction (forecasting)

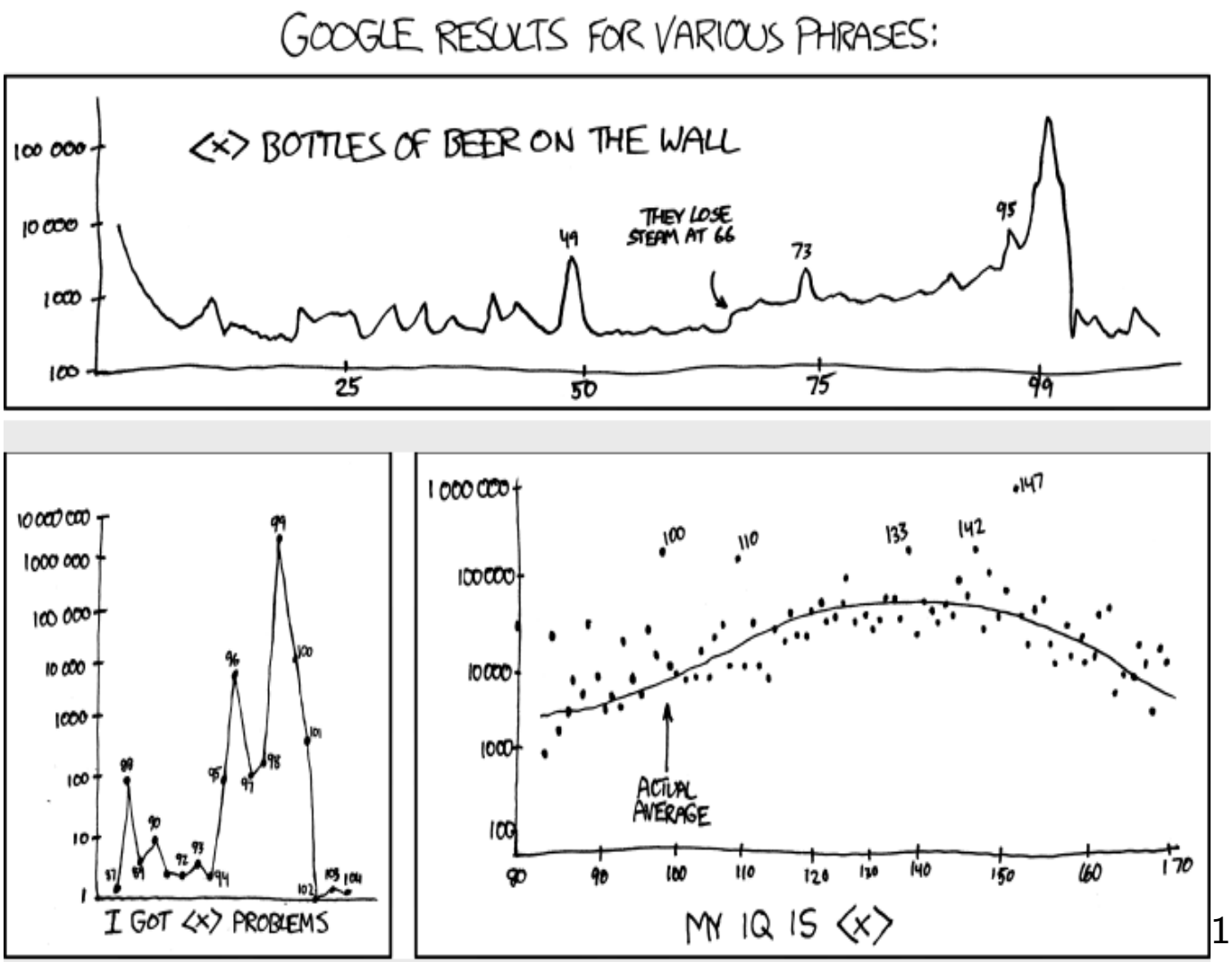
# Basic terms

- **Case**: individual, object or action observed
- **Population**: The complete collection of all possible cases
- **Sample**: A subset of the population
- **Random sample**: a sample with each element of the population having an equal chance of being selected
- **Census**: A sample of the entire population
- **Variable**: characteristic, property of the cases or population

# Watch out with 'random samples'!

- Voluntary response or self-selected samples not random
- Observational study $\neq$ experiment
- Beware of small samples!

# Self-selected samples




[1]

[1] http://xkcd.com/

# Sample vs. experiment

- Voluntary response or self-selected samples not random
- Observational study $\neq$ experiment
- Beware of small samples!

# Sample vs. experiment

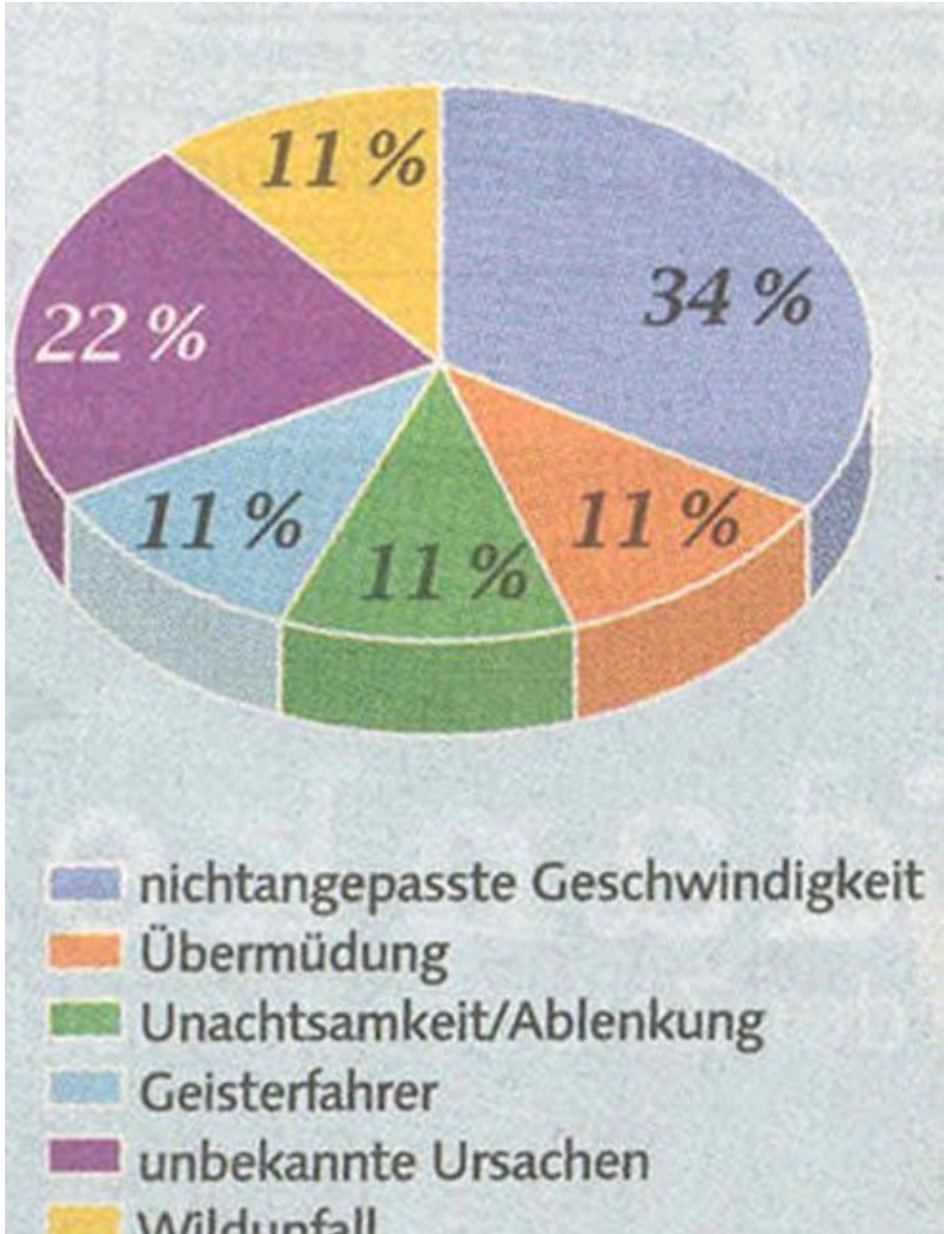
- **Census**: parliamentary votes; state census (household wise); firm inventory; exams (Matura)

  data: Statistik Austria
- **Random sample**:capturing animals; surveys for marketing asking about existing products; opinion polls; PISA; IQ tests
- **Experiment**: scientific experiment; calibration; surveys for marketing asking about future products with specific values to set; designing questions with 'test subjects' for studies (PISA, central Matura)

# Beware of small samples!

- Voluntary response or self-selected samples not random
- Observational study $\neq$ experiment
- Beware of small samples!

# Sample size?!


11 %
34 %
22 %
11 %
11 %
11 %
nichtangepasste Geschwindigkeit
Übermüdung
Unachtsamkeit/Ablenkung
Geisterfahrer
unbekannte Ursachen
Wildunfall

# Measurement Scales

**Measurement**: Mapping of observable phenomena to numbers.

- **categorial** variables; discrete categories (R: factor)
  - **nominal**; no ordering
    male/female; different products in marketing
  - **ordinal**; ordered
    Credit Ratings (AAA – D); Quality categories; grades
- **metric**; real valued measurements
  - **interval**; ordering matters, differences matter $(-\infty, \infty)$
    temperatures, companies' sales
  - **ratio**; ordering and ratios matter $(0, \infty)$, absolute zero
    incomes, heights, lengths, expenditures, stock prices

# Measurement Scales: Examples

## Classification into data types

| **Item** | **possible values** | **n/o/m** | **d/c** |
|---|---|---|---|
| Germ types | Bacteria (=1) Fungi (=2), Viruses (=3), etc. | nominal | discrete |
| Credit Ratings | AAA (=1), AA (=2), ..., D (=18) | ordinal | discrete |
| logarithmic concentrations (pH) | 1.40, 6.32 EUR, 9.6 EUR, ... | metric (ratio) | contin. |
| Years | 1999, 2012, ... | metric (interval) | discrete |
| cell counts | 20,500; 4,746; ... | metric (ratio) | discrete |
| Temperatures | 37.0°C, 38.1°C, ... | metric (interval) | cont. |

# The Notion of Probability

**Events**

events are subsets of the population set $\Omega$ of all possible events

e. g. rolling '6' with a die, drawing a specific card out of the deck

**Laplace Probability**

fraction of all interesting events (A) among all possible events (the population $\Omega$)

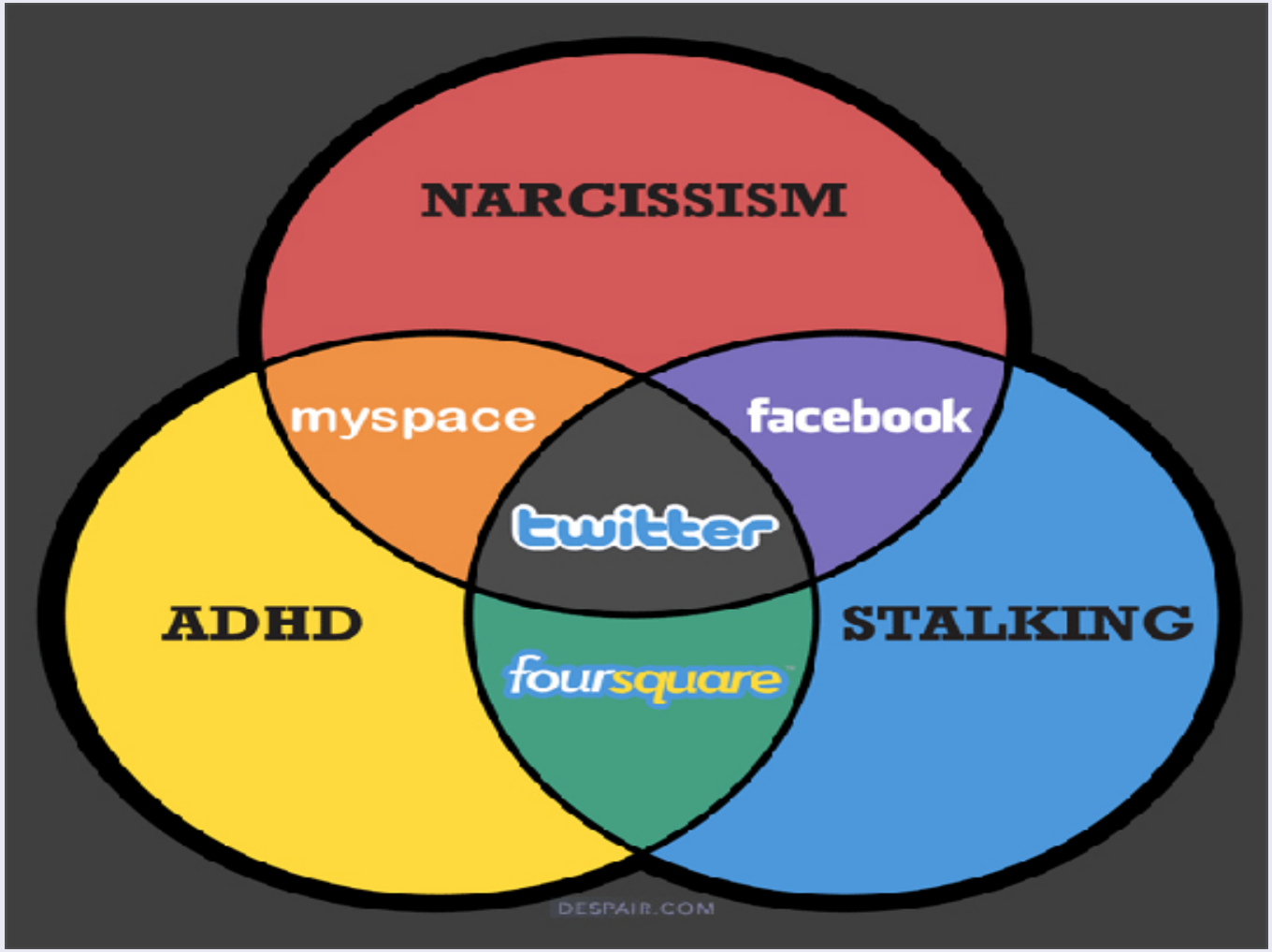
$$\mathbb{P}[A] = \frac{|A|}{|\Omega|}$$

e. g. $\mathbb{P}[\text{'rolling an even number with a die'}] = \frac{3}{6} = 0.5$

# Venn Diagrams

## Venn Diagram

Motivation: visualising coincidence of events



# Independence

## Independence

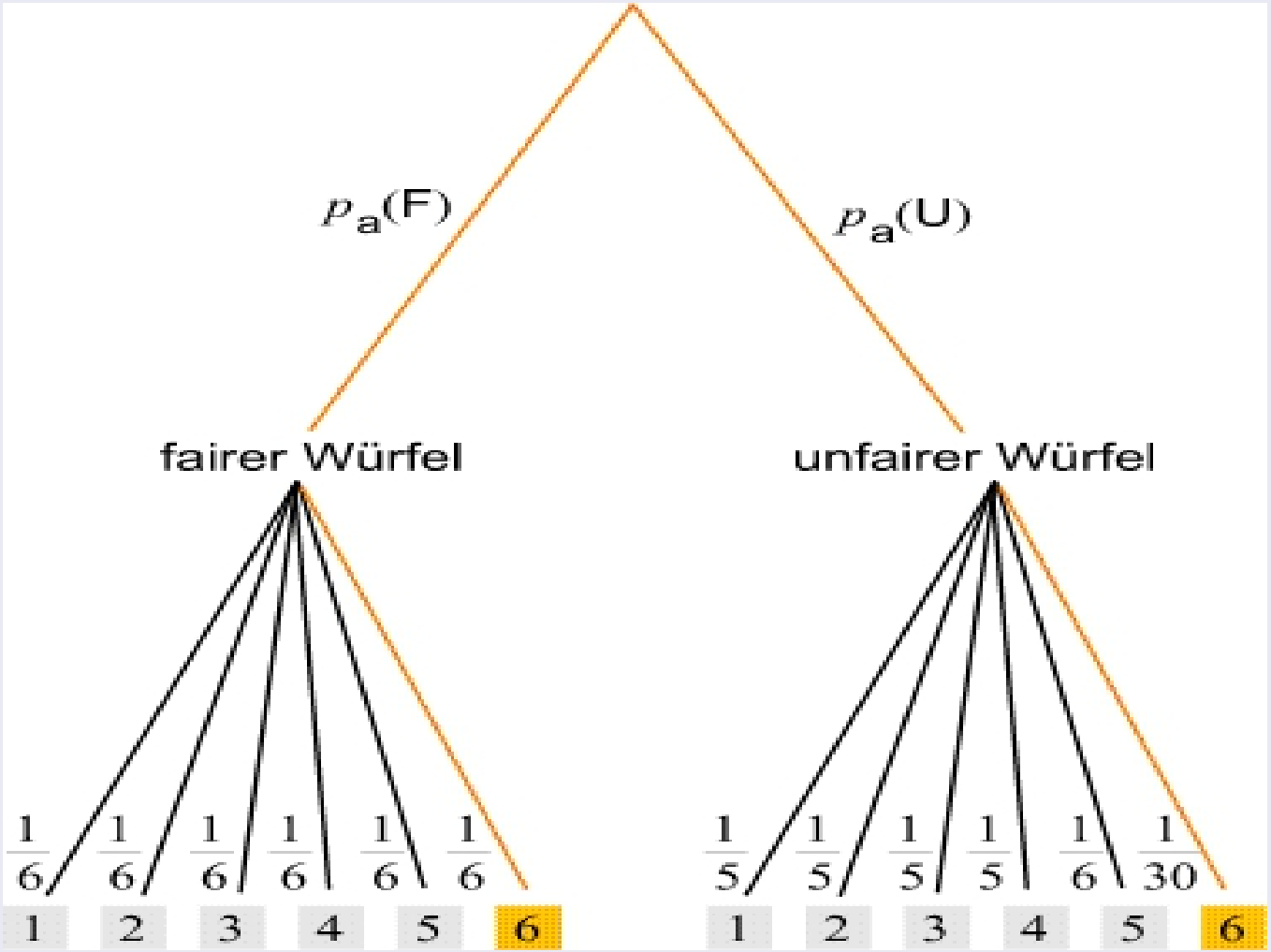
Events A and B are statistically independent, if

$$\mathbb{P}[A \text{ and } B] = \mathbb{P}[A \cup B] = \mathbb{P}[A]\mathbb{P}[B].$$

# Tree Diagrams

## Tree Diagram

Motivation: visualising events and their dependence



# The "gambling game" behind BLAST

**Example: DNA die**

bases {A,C,G,T} form the DNA

one of the bases is located in a single spot of the DNA
what is the probability it is A?
What is the probability of the sequence ACCTAAGGGC?

# Combinatorics

## samples

| | Variation | Combination |
|---|---|---|
| without replacement | $\frac{n!}{(n-k)!}$ | $\binom{n}{k}$ |
| with replacement | $n^k$ | $\frac{(n+k-1)!}{k!(n-1)!}$ |

All these options are implemented in R:
sample(x, size, replace = FALSE)

# DNA Examples

## Example: DNA sequences

1. How many sequences of 2 different bases can you make with the four bases {A,C,G,T}?
2. How many ways are there to pick 2 different bases out of the four bases {A,C,G,T}?
3. How many sequences of 2 bases can you make with the four bases {A,C,G,T}?
4. How many ways are there to pick 2 bases out of the four bases {A,C,G,T}?
5. What is the probability of the sequence ACCTAAGGGC?

# Conditional Probability

## Conditional Probability

The probability of an event A, knowing that event B has happened

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cup B]}{\mathbb{P}[B]}$$

$$\text{in words: conditional prob.} = \frac{\text{joint prob.}}{\text{marginal prob.}}$$

W. r. t. conditional probabilities, **independence** means

$$\mathbb{P}[A|B] = \mathbb{P}[A].$$

## marginal probability

The marginal probability of event A, when knowing all probabilties of A given several events $B_k$ and occurences of $B_k$ is

$$\mathbb{P}[A] = \sum \mathbb{P}[A|B_k]\mathbb{P}[B_k] \quad (1)$$

# Bayes' theorem

## Bayes' theorem

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$$

In other words:

$$\text{posterior probability} = \frac{\text{likelihood * prior probability}}{\text{marginal probability}}$$

## First guess, then calculate

The probability that a woman of a certain age has breast cancer is 0.8%, and the probability that if she does have breast cancer her mammogram will be positive is 90%, the probability that if she does not have breast cancer her mammogram will be positive is 7%. If her test comes back positive, what is the probability that she actually has breast cancer?

# Discrete data distributions

- uniform
- Bernoulli
- binomial
- hypergeometric
- Poisson
- negative binomial; geometric

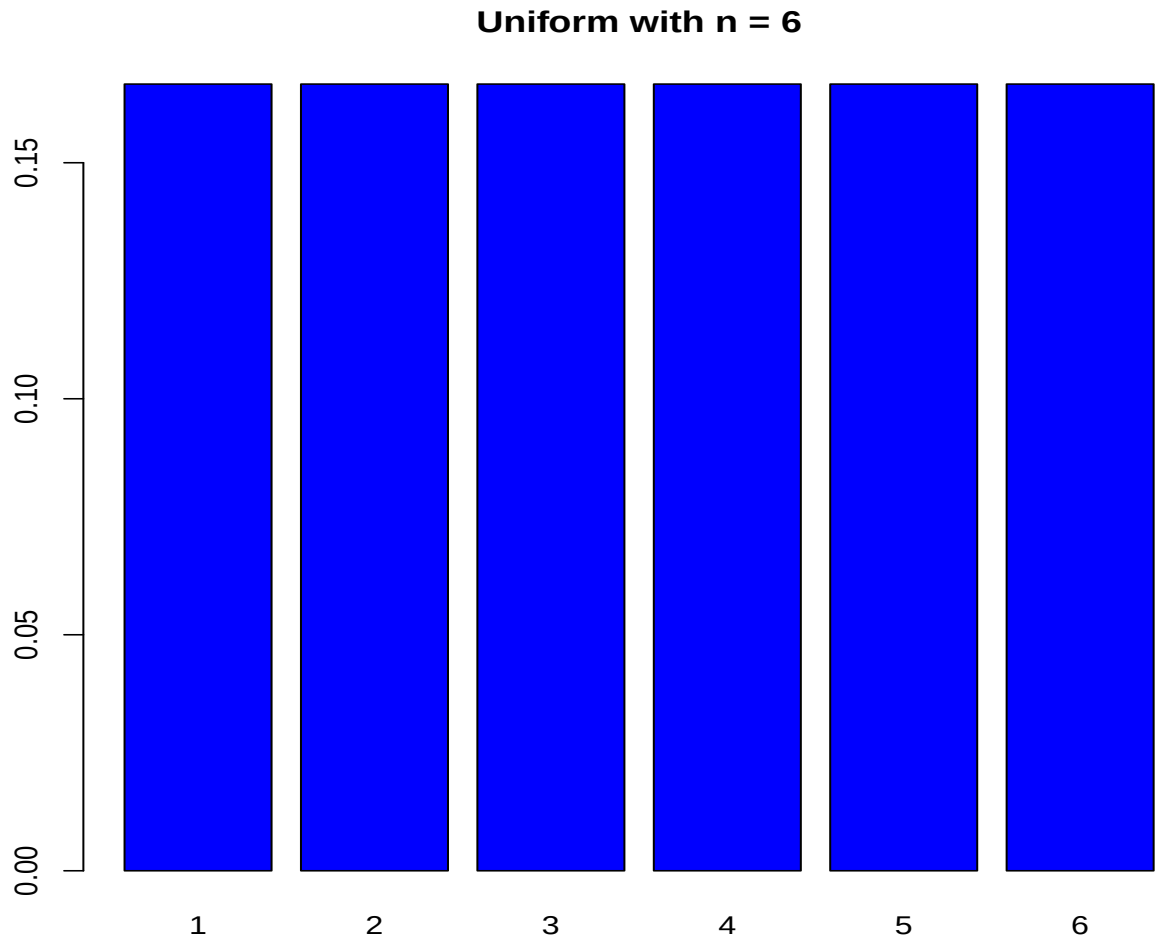
# Uniform distribution (discrete) I

$X$ may take values from 1 to $n$, each with probability $1/n$.

$$\mathbb{P}[X = k] = \begin{cases} 1/n & \text{if } k \in \{1, \ldots, n\}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} \mathbb{E}[X] &= \frac{n+1}{2} \\ \mathbb{V}[X] &= \frac{n^2-1}{12} \end{aligned}$$

# Uniform distribution (discrete) II



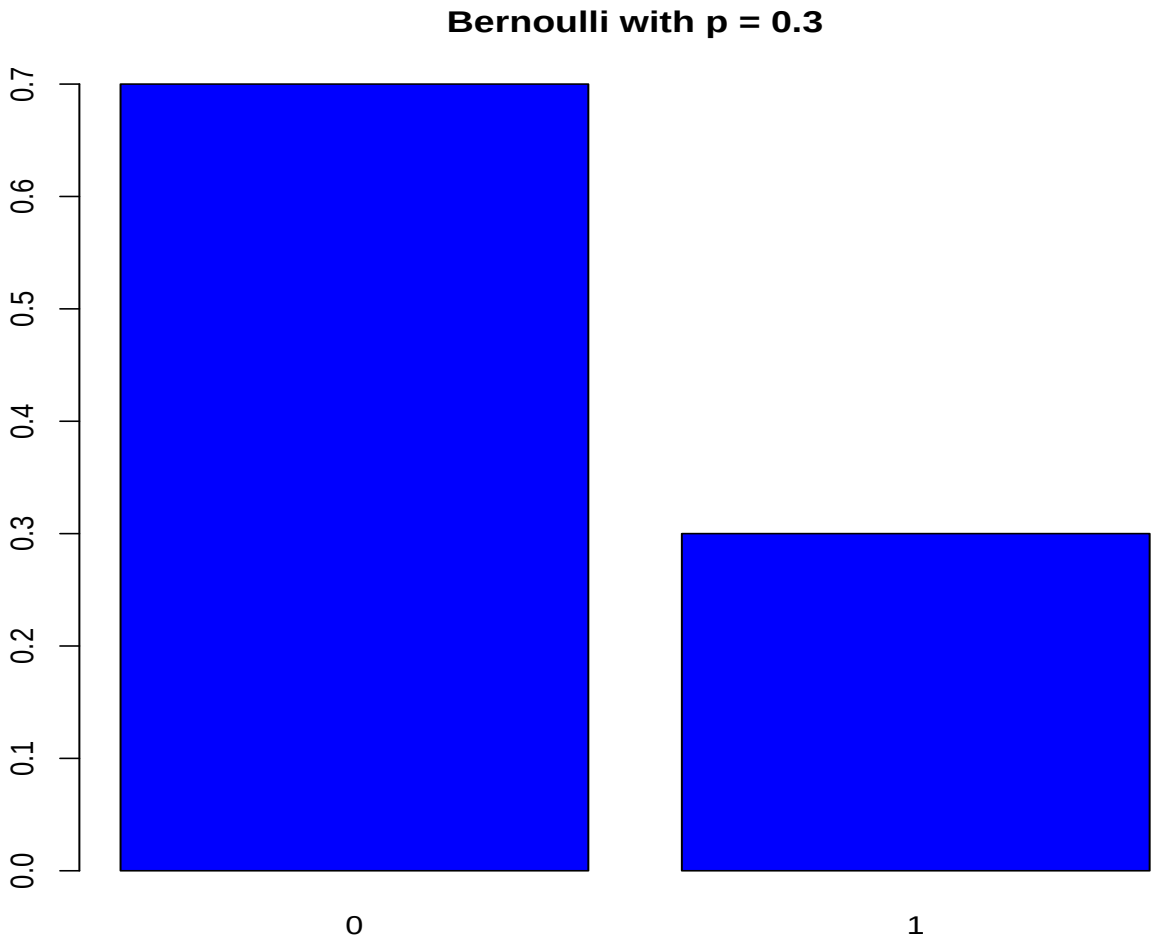
# Bernoulli distribution I

$X$ takes value 1 with success probability $p$ and value 0 with failure probability $1 - p$.

$$\mathbb{P}[X = k] = \begin{cases} p & \text{if } k = 1, \\ 1 - p & \text{if } k = 0, \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} \mathbb{E}[X] &= p \\ \mathbb{V}[X] &= p(1 - p) \end{aligned}$$

# Bernoulli distribution II



# Binomial distribution I

Bernoulli chain $\rightarrow$ binomial distribution

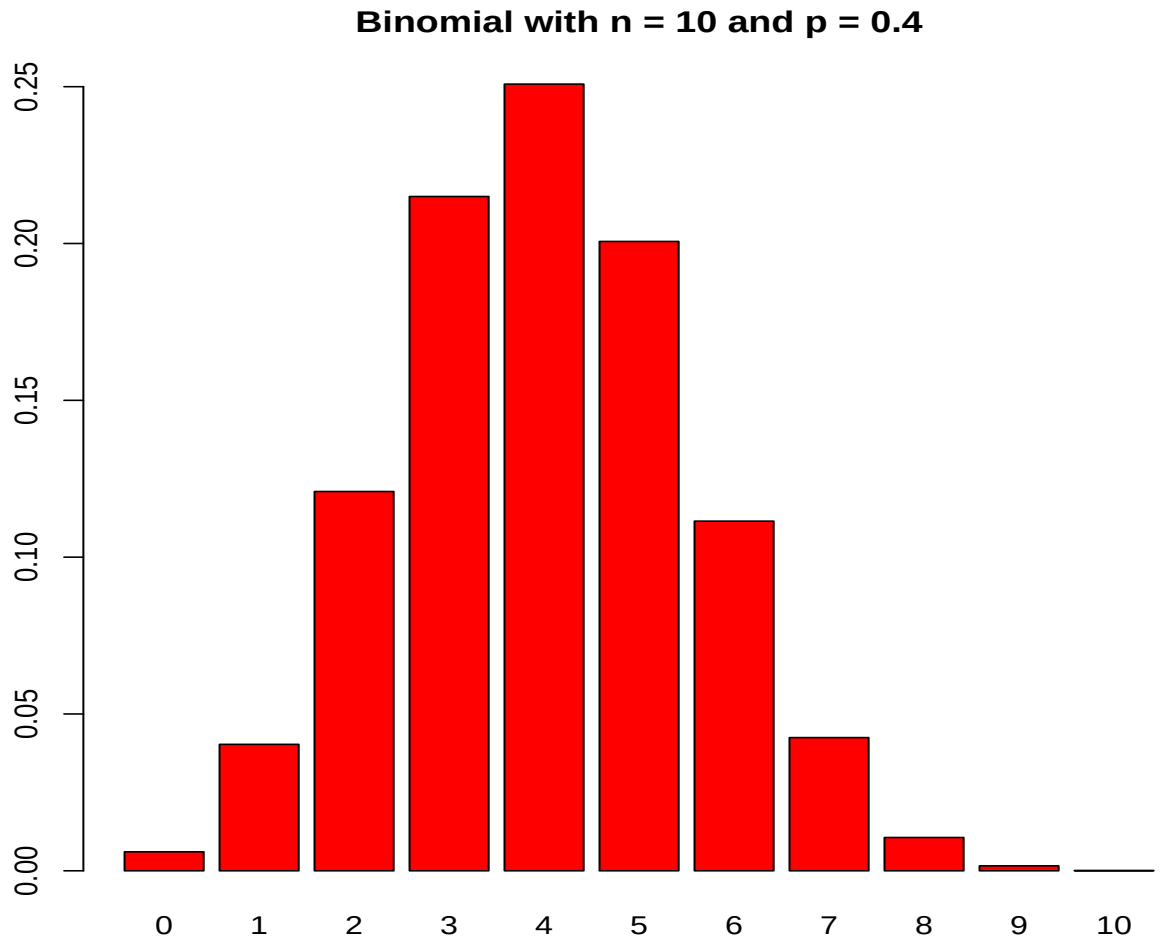
$X$ takes values from 0 to $n$, counting the number of successes of $n$ independent Bernoulli experiments with success probability $p$.

$$\mathbb{P}[X = k] = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k \in \{0, \ldots, n\}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} \mathbb{E}[X] &= np \\ \mathbb{V}[X] &= np(1-p) \end{aligned}$$

"Drawing with replacement while counting the number of successes"

# Binomial distribution II



## Bernoulli & binomial distribution

Microarray gene expression analysis:
each gene can be differentially expressed or non-differentially expressed
for the whole array with N genes: number of differentially expressed genes, if each gene has a random chance $p$ of being differentially expressed
several arrays within one experiment $\rightarrow X \sim Bin(N, p)$

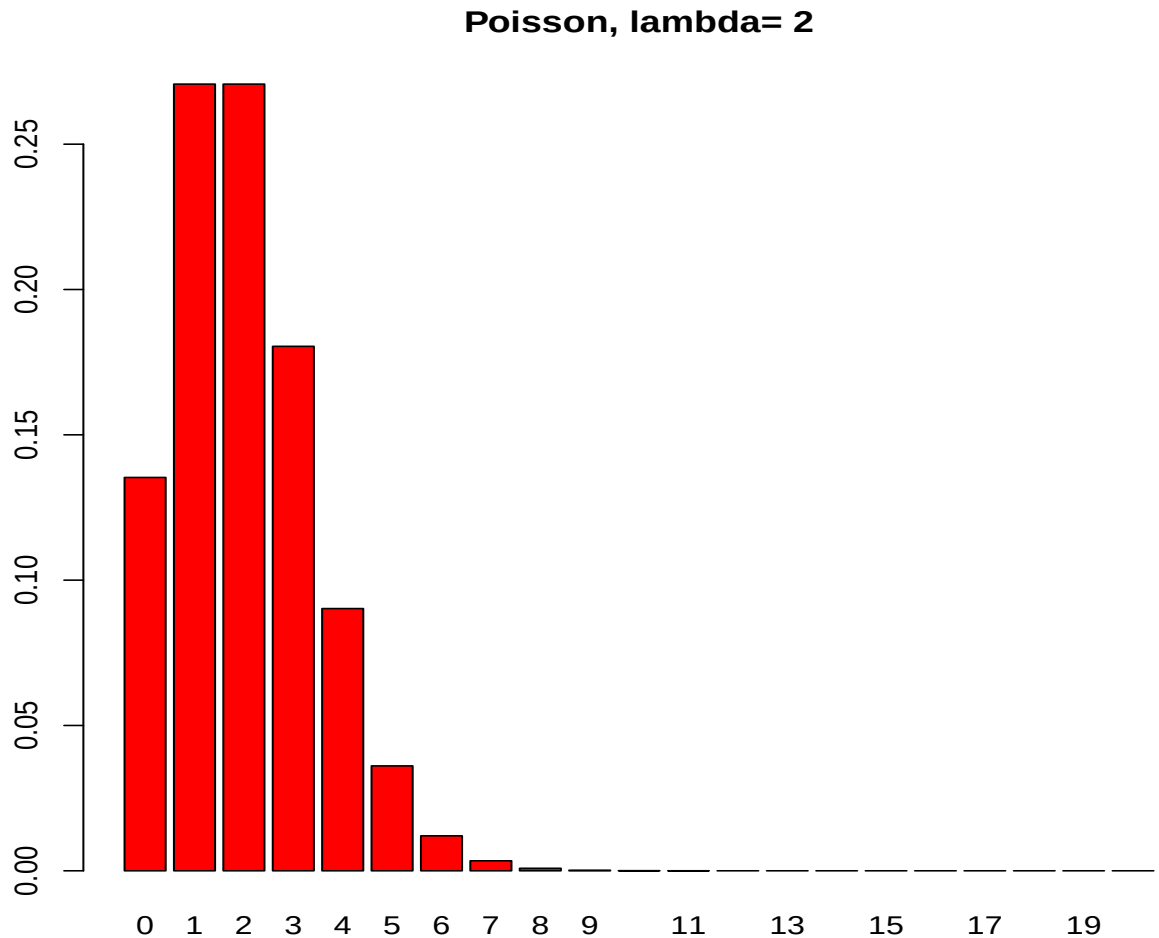
# Poisson distribution I

$X$ takes values from 0 to $\infty$, counting the number of events occurring during an interval of time or space.

$$\mathbb{P}[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}$$

For $n \geq 100$ and $\lambda = np \geq 10$ the binomial distribution and Poisson distribution become almost the same.

# Poisson distribution II



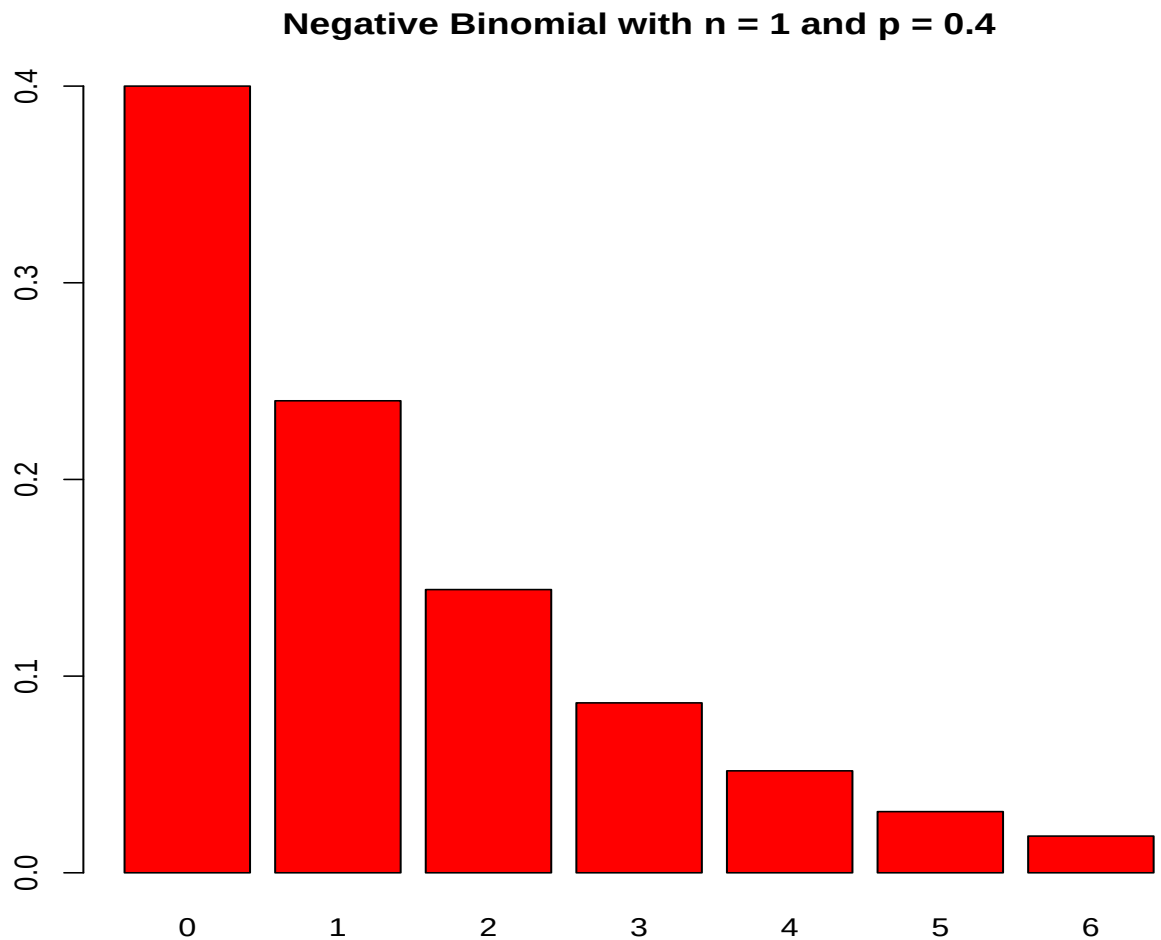
# Negative binomial distribution I

$X$ takes values from 0 to $\infty$, counting the number of 'successes' in a sequence of Bernoulli trials until a specified number of 'failures' $r$ has occurred.

$$\mathbb{P}[X = k] = \binom{k + r - 1}{k} (1 - p)^r p^k$$

Special case: geometric distribution (r=1)

# Negative Binomial distribution II



# Continuous data distributions

- uniform
- normal/Gaussian
- Student's t
- gamma; $\chi^2$
- beta

# Objective

When modeling continuous data we move from a finite (or countably infinite) number of "states of the world" to uncountably infinitely many outcomes, usually some interval $[a, b] \in \mathbb{R} \cup \{\pm\infty\}$. Examples:

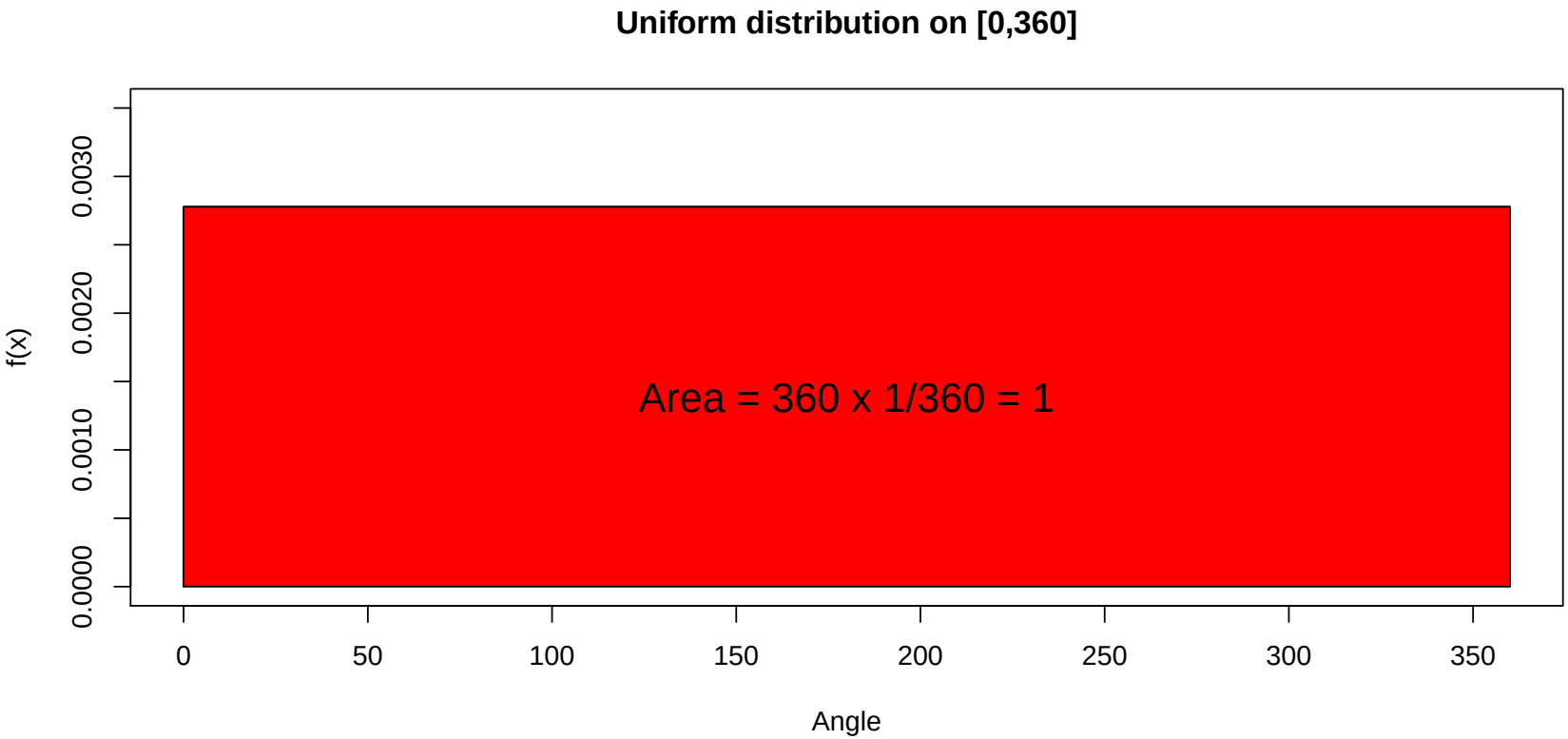
- Body heights: $\Omega = \mathbb{R}^+$
- Temperature (on Celsius scale): $\Omega = (-273.15, \infty)$
- Angle when playing Wheel of Fortune: $\Omega = [0, 360)$

Thus, we move from summation to integration:

$$\mathbb{P}[a \leq X \leq b] = \int_a^b f(x)\, dx,$$

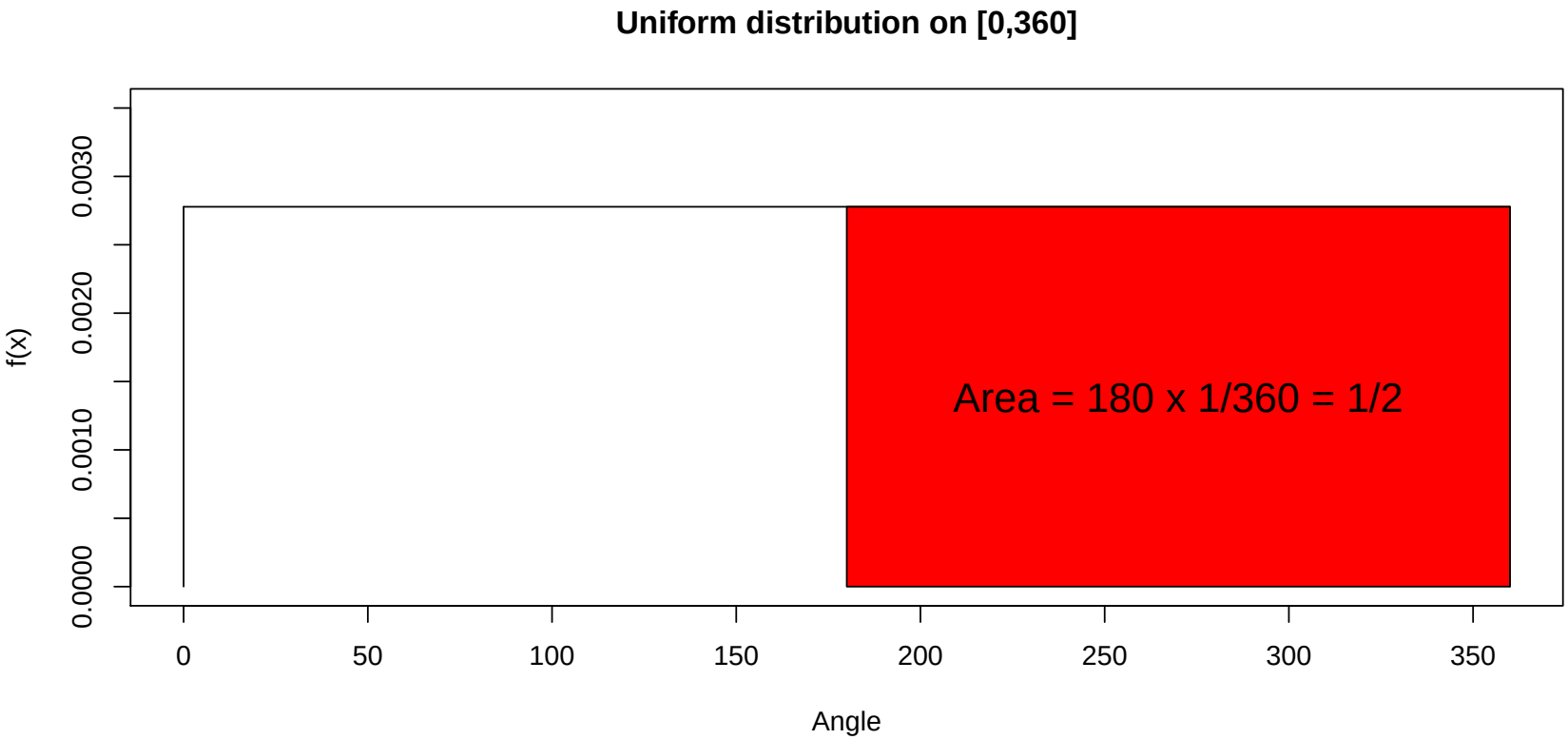
with $f(x) \geq 0$ and $\int_\Omega f(x)\, dx = 1$. This obviously implies that $\mathbb{P}[X = c] = \int_c^c f(x)\, dx = 0$ (no "point mass").

# Uniform distribution (continuous)




$$\mathbb{P}[0 \leq X \leq 360] = \int_0^{360} f(x)dx = \int_0^{360} \frac{1}{360} dx = 1$$

# Uniform distribution (continuous)




$$\mathbb{P}[180 \leq X \leq 360] = \int_{180}^{360} f(x)dx = \int_{180}^{360} \frac{1}{360} dx = \frac{1}{2}$$

# Normal (Gaussian) distribution

## Normal distribution

The probability density of the normal distribution is defined as

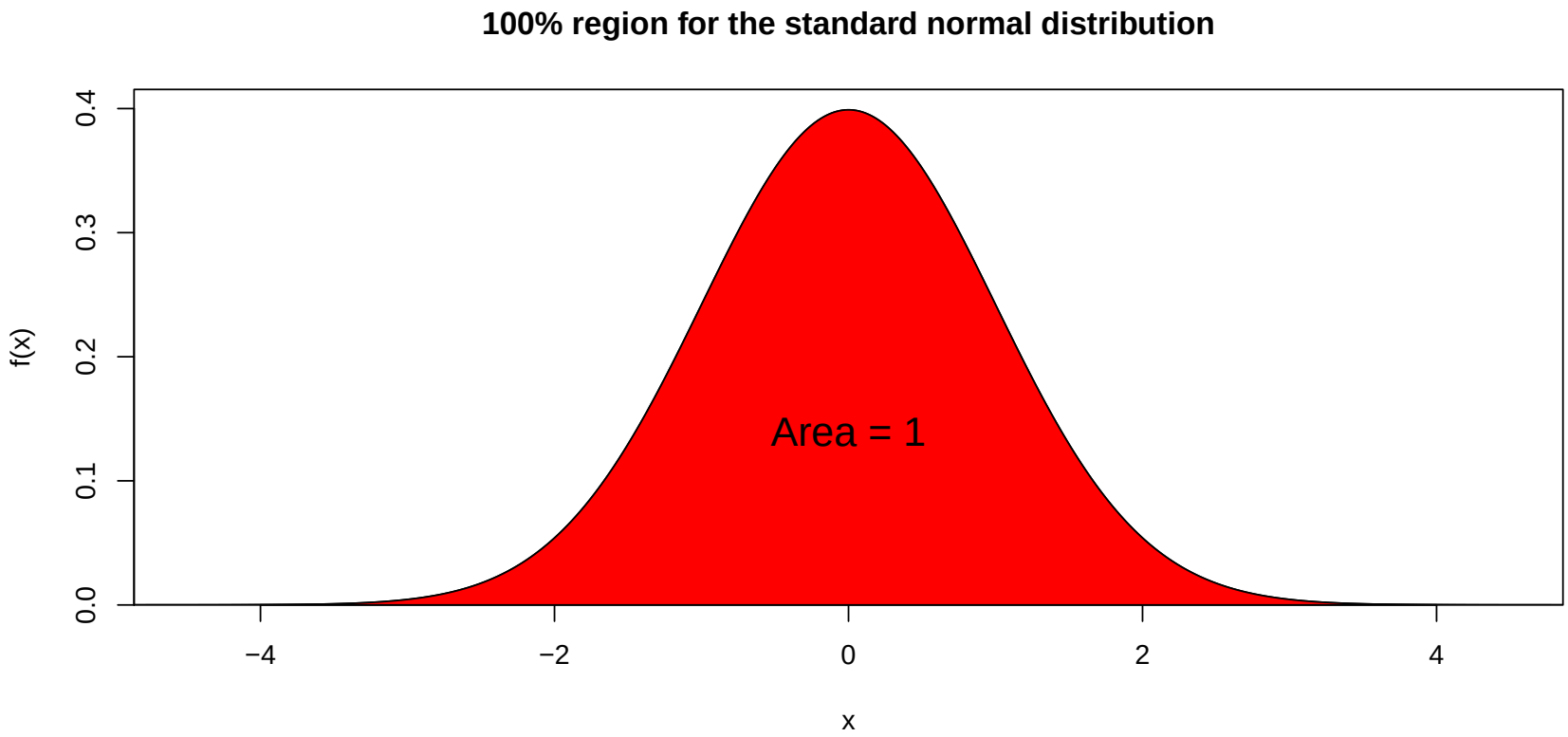
$$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

The distribution has expected value and Variance

$$\begin{aligned} \mathbb{E}[X] &= \mu \\ \mathbb{V}[X] &= \sigma^2 \end{aligned}$$

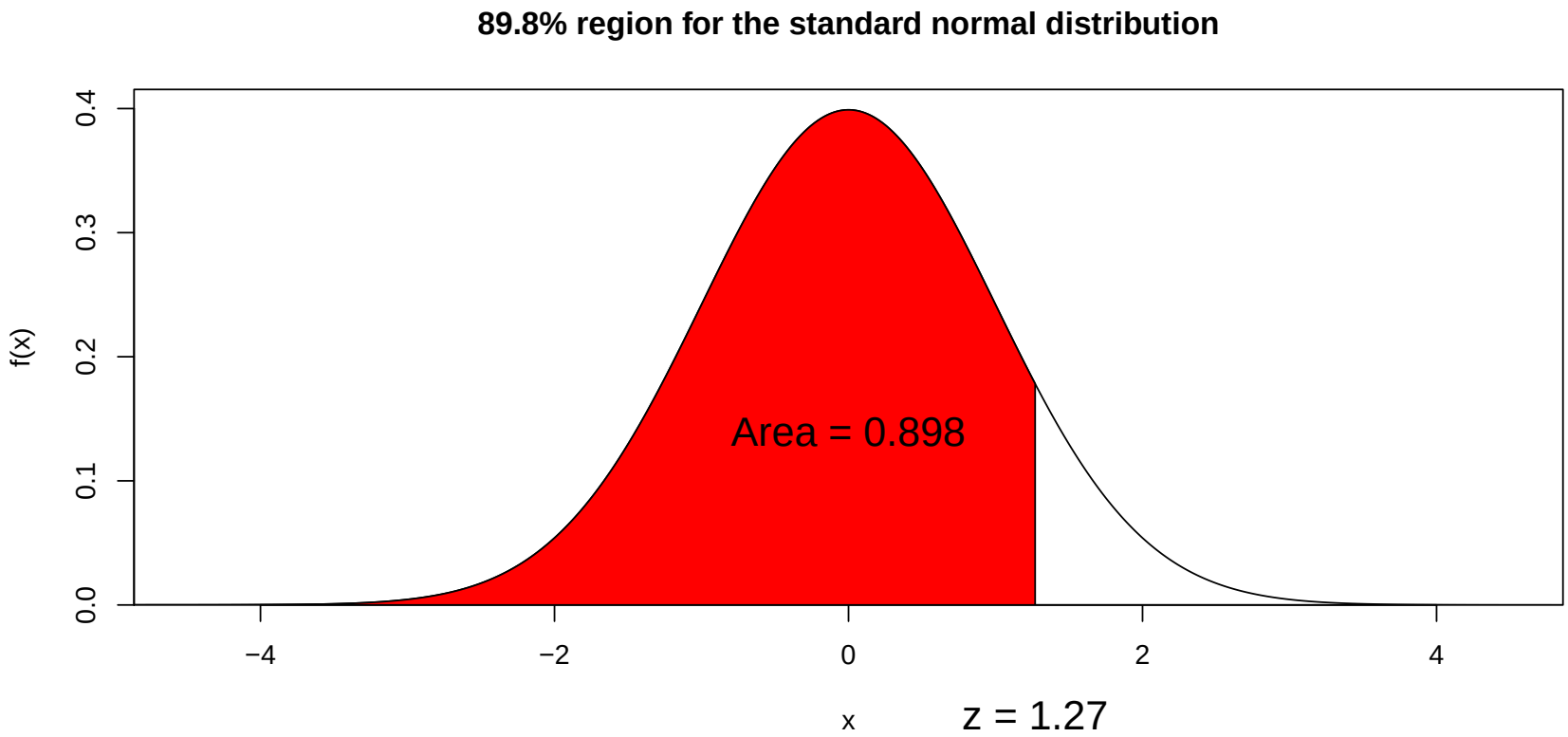
The standard normal distribution has mean $\mu = 0$ and variance $\sigma^2 = 1$.

# Standard normal (Gaussian) distribution




$$\mathbb{P}[-\infty \leq X \leq \infty] = \int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx =$$
$$= F(\infty) - F(-\infty) = 1 - 0 = 1$$

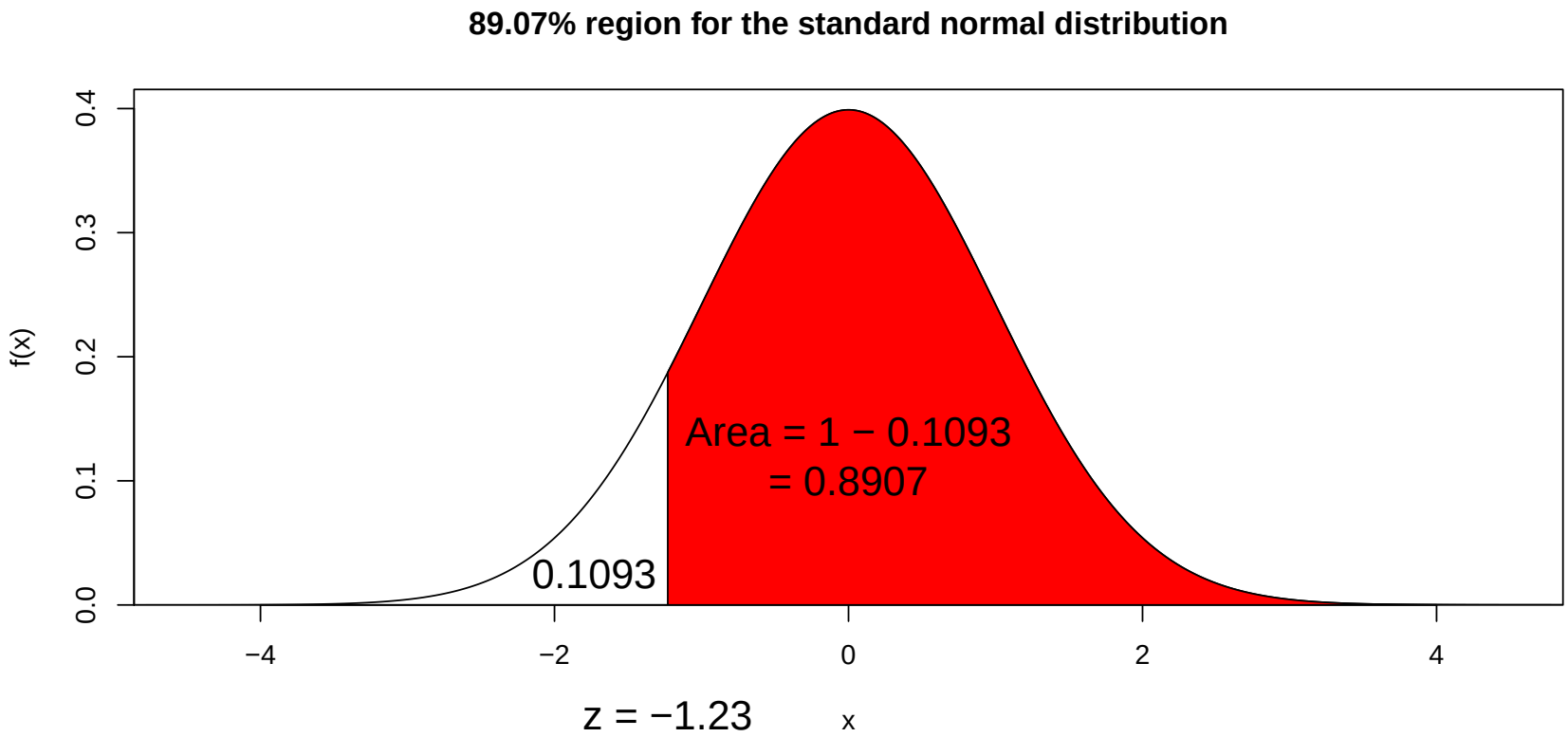
# Standard normal (Gaussian) distribution




$$\mathbb{P}[-\infty \leq X \leq 1.27] = \int_{-\infty}^{1.27} f(x)dx = \int_{-\infty}^{1.27} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx =$$

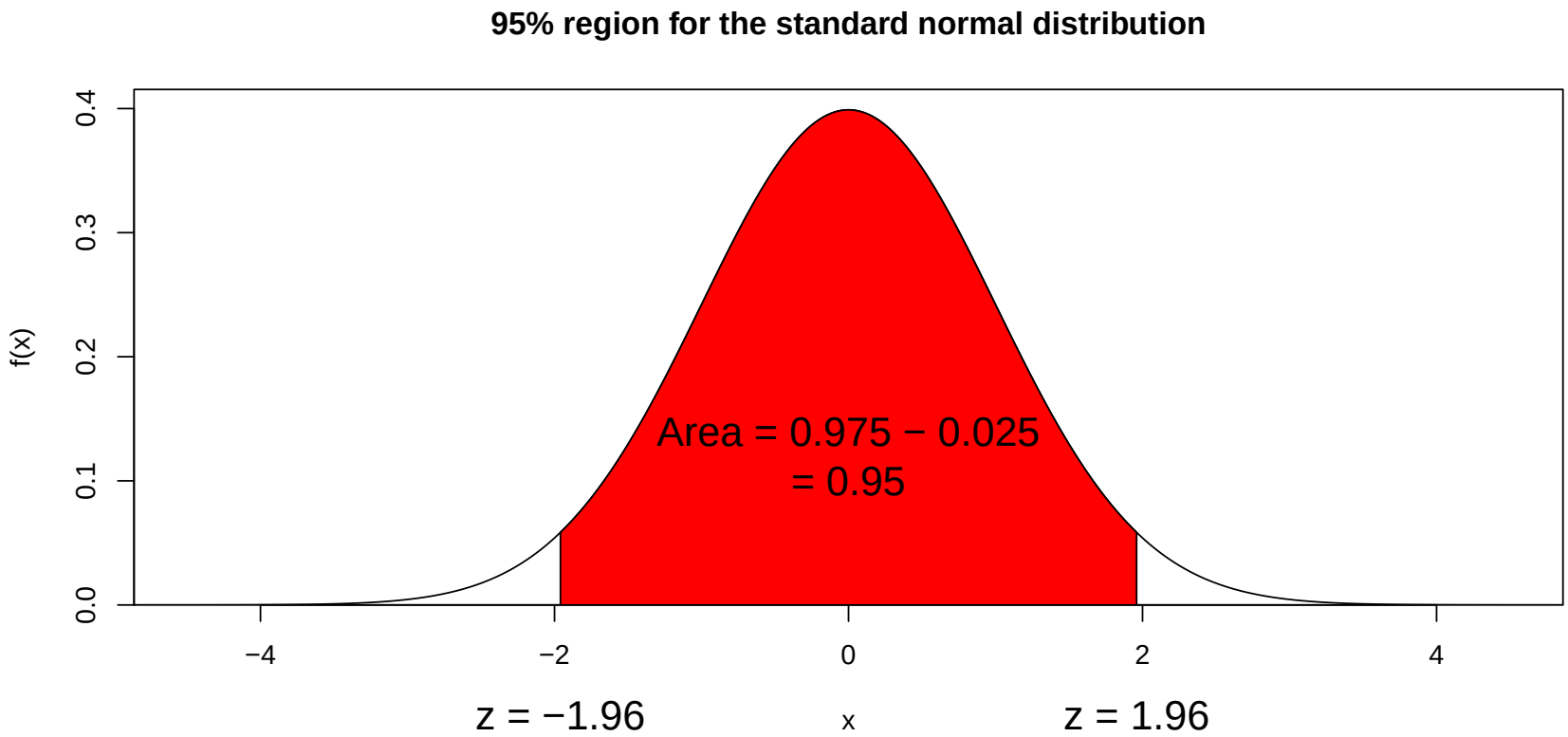
$$= F(1.27) - F(-\infty) = 0.8980 - 0 = 0.8980$$

# Standard normal (Gaussian) distribution




$$\mathbb{P}[-1.23 \leq X \leq \infty] = \int_{-1.23}^{\infty} f(x)dx = \int_{-1.23}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx =$$
$$= F(\infty) - F(-1.23) = 1 - 0.1093 = 0.8907$$

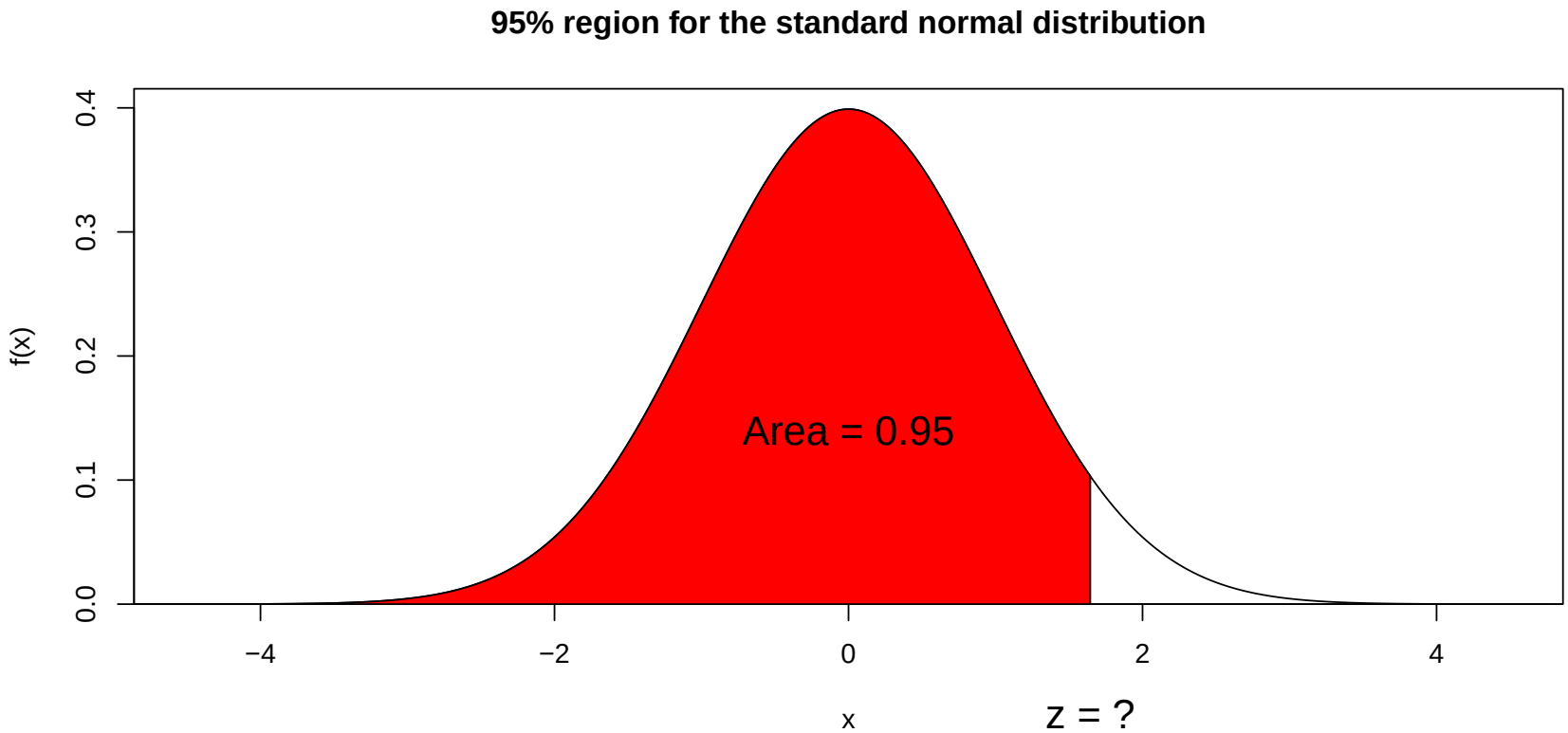
# Standard normal (Gaussian) distribution




$$\mathbb{P}[-1.96 \leq X \leq 1.96] = \int_{-1.96}^{1.96} f(x)dx = \int_{-1.96}^{1.96} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx =$$

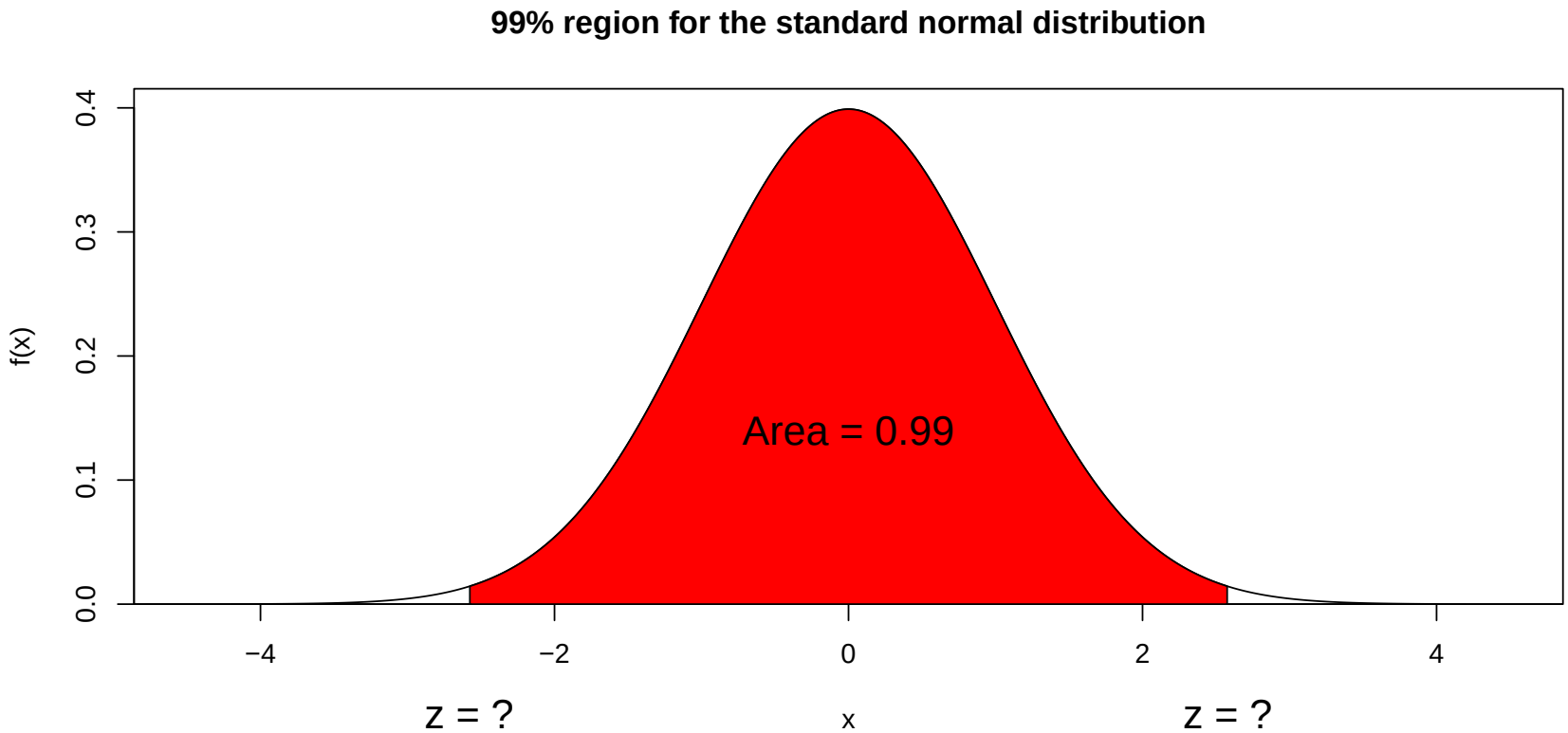
$$= F(1.96) - F(-1.96) = 0.975 - 0.025 = 0.95$$

# Standard normal (Gaussian) distribution



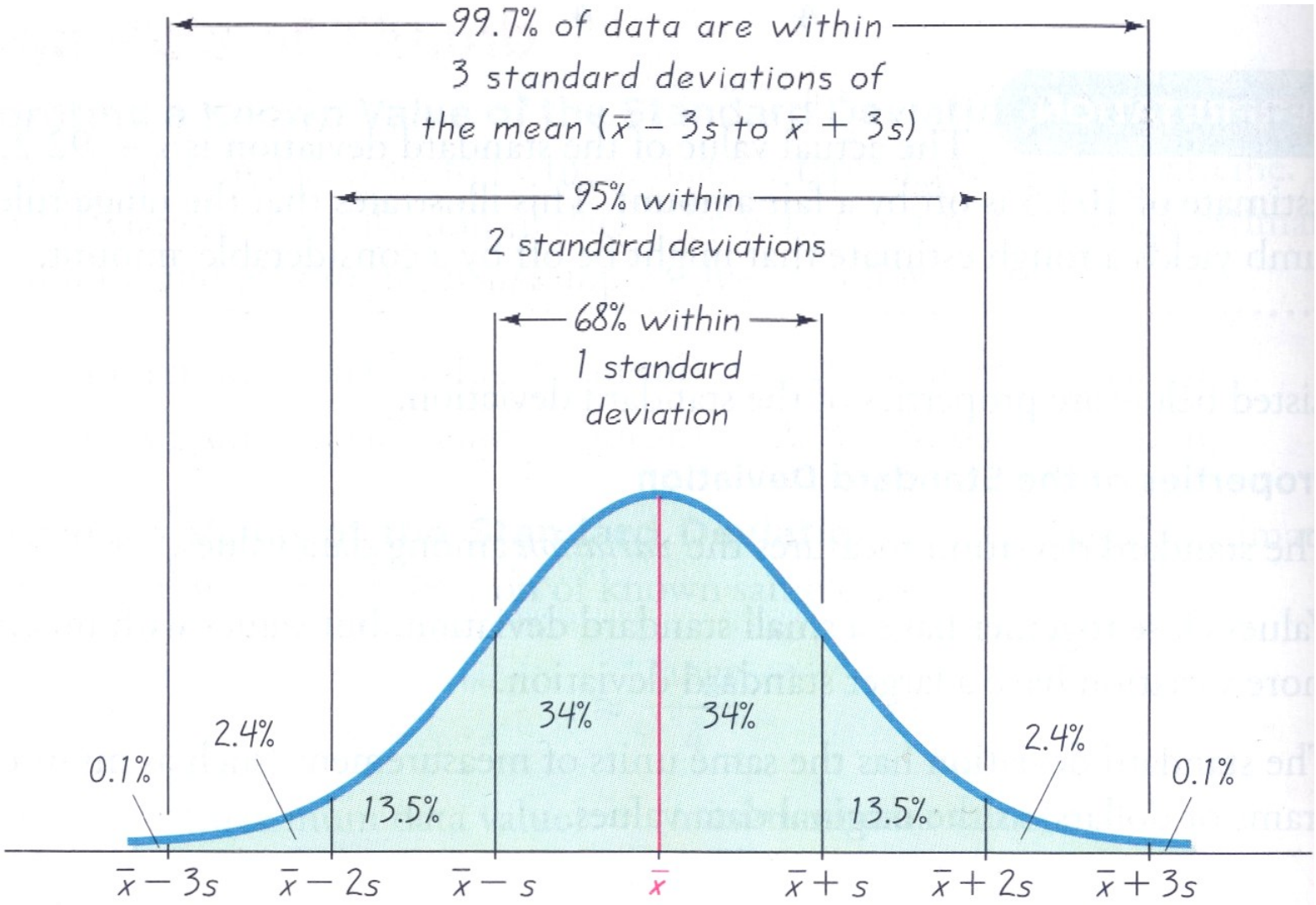

$$\mathbb{P}[-\infty \leq X \leq z] = 0.95 \Longrightarrow z = 1.645$$

# Standard normal (Gaussian) distribution




$$\mathbb{P}[p_{0.5} \leq X \leq p_{99.5}] = 0.99 \Longrightarrow z = \pm 2.575$$

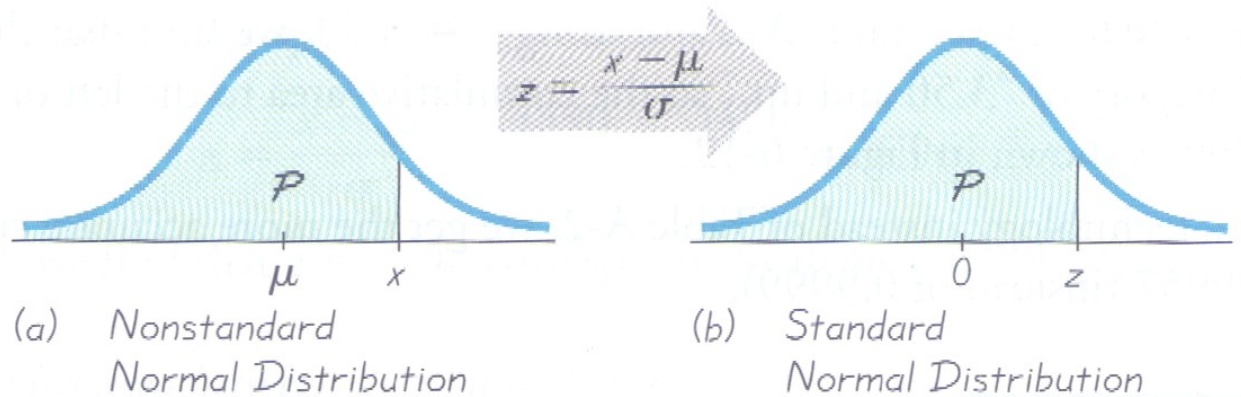
# Empirical (or 68-95-99.7) rule for Gaussian data



# Converting from a nonstandard to a standard normal distribution

## $z$-Transformation

$$z = \frac{x - \mu}{\sigma}$$




(a) Nonstandard Normal Distribution

(b) Standard Normal Distribution

# Example: Finding values from known areas

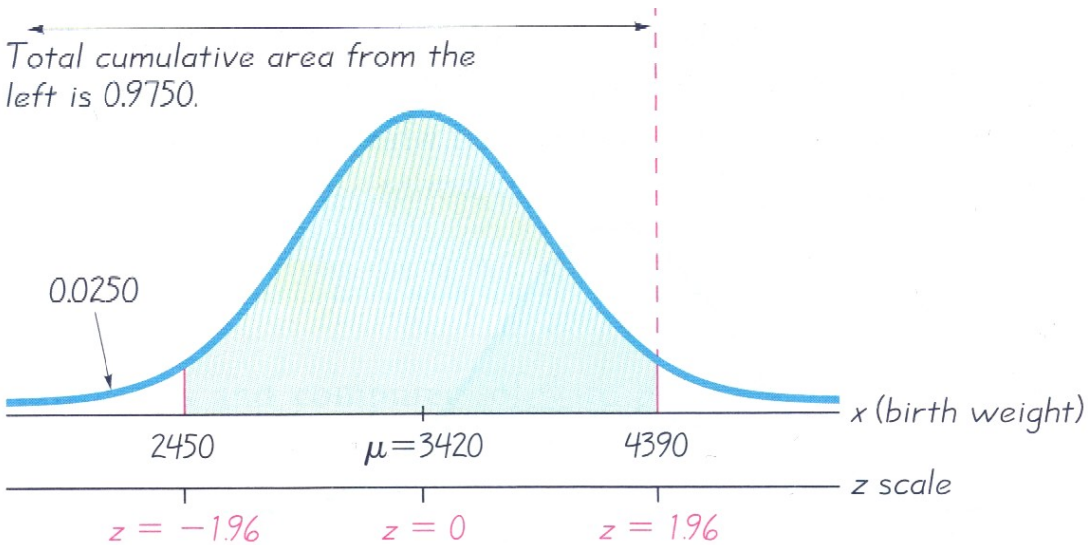
**Birth weights**

Birth weights in the US are normally distributed with mean 3420g and a standard deviation of 495g. The Newport General Hospital requires special treatment for babies that are less than 2450g (unusually light) or more than 4390g (unusually heavy).

# Example: Finding values from known areas

$$z_{\text{low}} = \frac{2450 - 3420}{495} \approx -1.96$$

$$z_{\text{high}} = \frac{4390 - 3420}{495} \approx 1.96$$




Thus,

- 95% of the babies don't require special treatment
- 2.5% of the babies are considered to be too light
- 2.5% of the babies are considered to be too heavy

# Example: Finding areas from known values

**Birth weights**

The Newport General Hospital wants to redefine the minimum and maximum birth weights that require special treatment because they are unusually low or unusually high. After considering relevant factors, a committee recommends special treatment for birth weights in the lowest 3% and highest 1%.

$$\begin{aligned} x_{\text{low}} &= \mu + z\sigma = 3420 + (-1.88) \times 495 = 2489.4 \\ x_{\text{high}} &= \mu + z\sigma = 3420 + 2.33 \times 495 = 4573.35 \end{aligned}$$

Thus,

- 2489g separates the lowest 3% of birth weights,
- 4573g separates the highest 1% of birth weights.