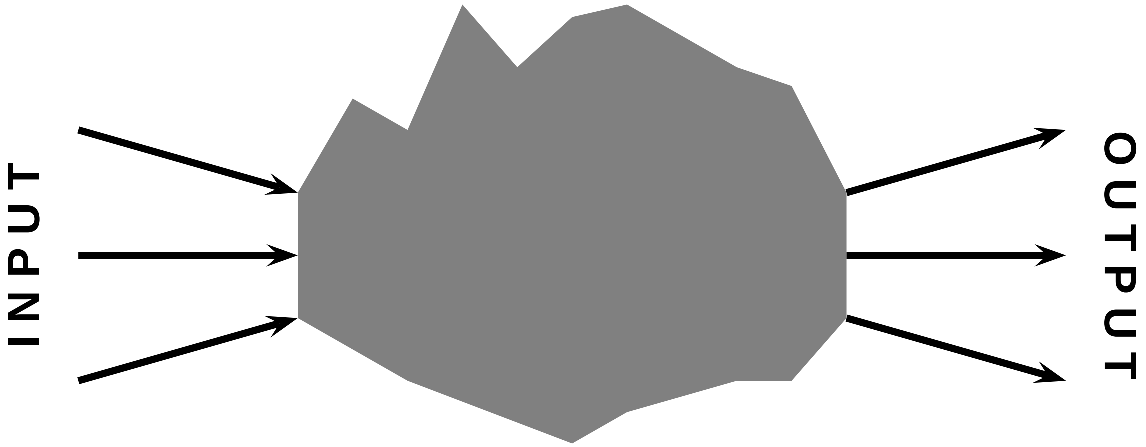
Modelling input

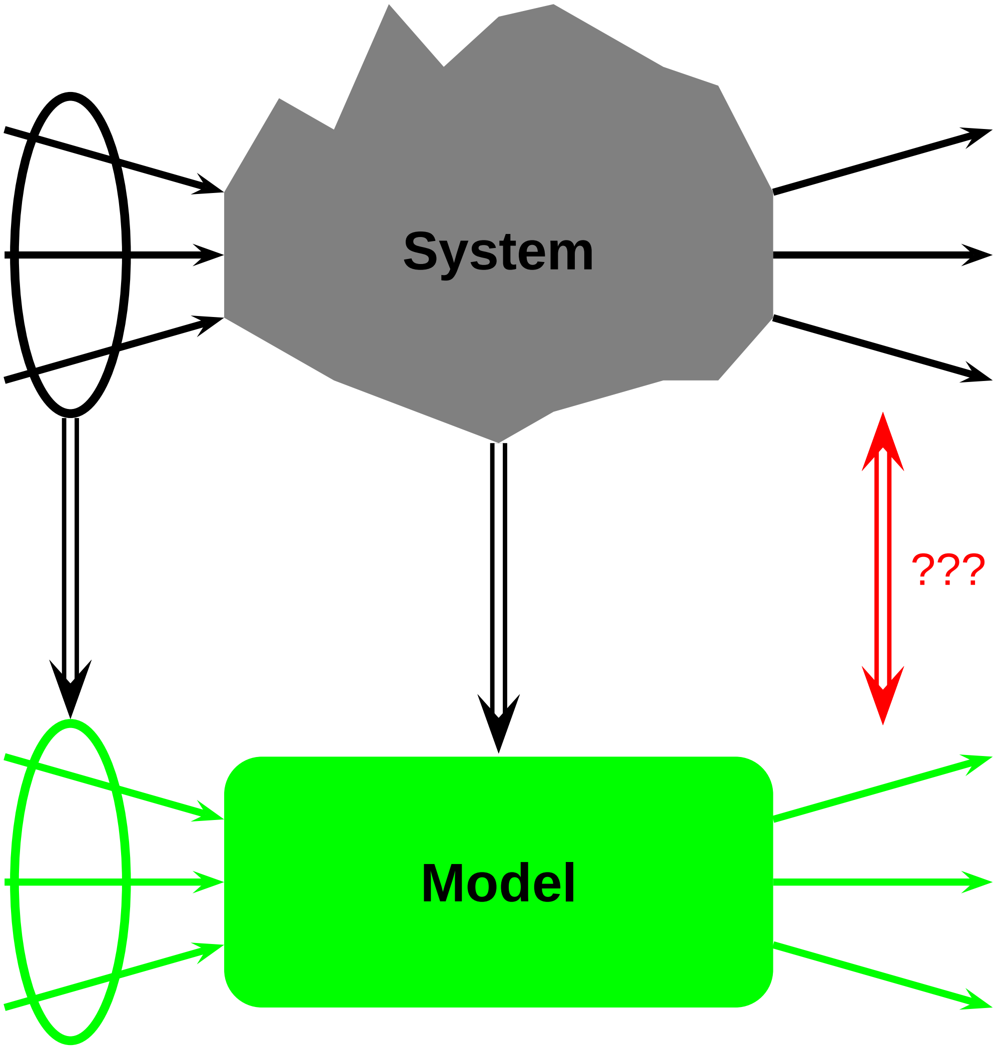
This is the first step in any simulation project. The input to the actual system is only partially known, therefore it has to be modelled with care.

The definition of “input” is simple: it is made of data **independent** of the behaviour of the system itself. Customer arrivals are a typical example of input data.

System



System vs. model



Data collection

This is real art with few algorithmic steps to follow.

The purpose is to sift through existing data in order to guess a pdf that will model adequately the behaviour of some input parameter.

In many cases it is easy–or at least looks easy. If we need to model the flow of customers into a new store being designed, we might be tempted to take actual data from an existing store. Not a good idea, though.

In the rare cases when existing data can be used to model an input parameter, one should be careful to do a thorough job.

An excellent example is provided by Banks et al.:

> *the modelling of the time needed to get through a metal detector.*

- An observer was placed near a metal detector and recorded 1000 durations of time needed by people to get through a metal detector.
- These 1000 samples had a sample mean of 30 seconds and a sample deviation of 30 seconds, strongly suggesting the use of exponential distribution.
- A shrewd analyst noted that 30 seconds are far more than the time needed to pass through and a more detailed analysis was done.
- It turned out that the sample consisted of two distinct subsets: those who passed without problems (764 people) and those who had to try a second time (236 people).

An additional twist was added when it was discovered that 9 people passed through the metal detector in negative time.

The existence of negative service times puts in doubt the value of all the other measurements; unfortunately this is a very common situation.

Random sampling

If we do not know the pdf of a random variable $\mathcal{X}$, we might try to guess it. By taking several measurements (observations) of the random variates (values of $\mathcal{X}$ observed in specific experiments), we get a **random sample** partially describing $\mathcal{X}$.

This sample is random only if the experiments are not correlated, i.e. are independent. If they are independent, we call these experiments **replications**.

Note that the set of values obtained by sampling $\mathcal{X}$ is not a random variable, even though each replication may yield a random variate.

Sample

A random sample is characterized by measures similar, but not identical, to those of random variables:

Central tendency is the equivalent of the mean.

$$\overline{\mathcal{X}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Dispersion is the equivalent of the standard deviation.

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{\mathcal{X}}^2}{n-1}$$

The size of the sample influences the correlation between these measures and the values of μ and σ of the actual random variable from which the sample was drawn.

Fitting experimental data

A random sample allows us to guess the pdf of the random variable that the sample was taken from. In some sense thie process is a reverse simulation; one should expect all the potential difficulties caused bt stochastic effects (which do or do not appear in the sample).

The simplest approach is to check if the sample looks similar to a sample generated from a known distribution (this is a crude method, but not totally useless).

Note that while guessing the pdf of a random variable, we can neither assume that it is continuous if the sample contains non–integer values nor assume that it is discrete if it contains integers only.

Consider a random sample made of 10 values, X_{10}, shown sorted for convenience (when sampling, the order in which variates are measured should not matter, unless it does, of course). Clearly, the sample is too small. X_{10} has a central tendency $\overline{X_{10}} = 12.8499$ and a dispersion $S_{10} = 28.1791$.

Our goal is to find the pdf of the random variable $\mathcal{X}$ from which X_{10} was taken. One reasonable guess is that the pdf is exponential and that the value of the dispersion (which should be close to the central tendency) is a fluke due to the occurrence of 92.6205 (the stochastic effect of a black swan in a small sample).

To check the hypothesis, we generate a 10–variate sample from the exponential distribution with a mean $\mu = \overline{X_{10}}$ and compare the samples.

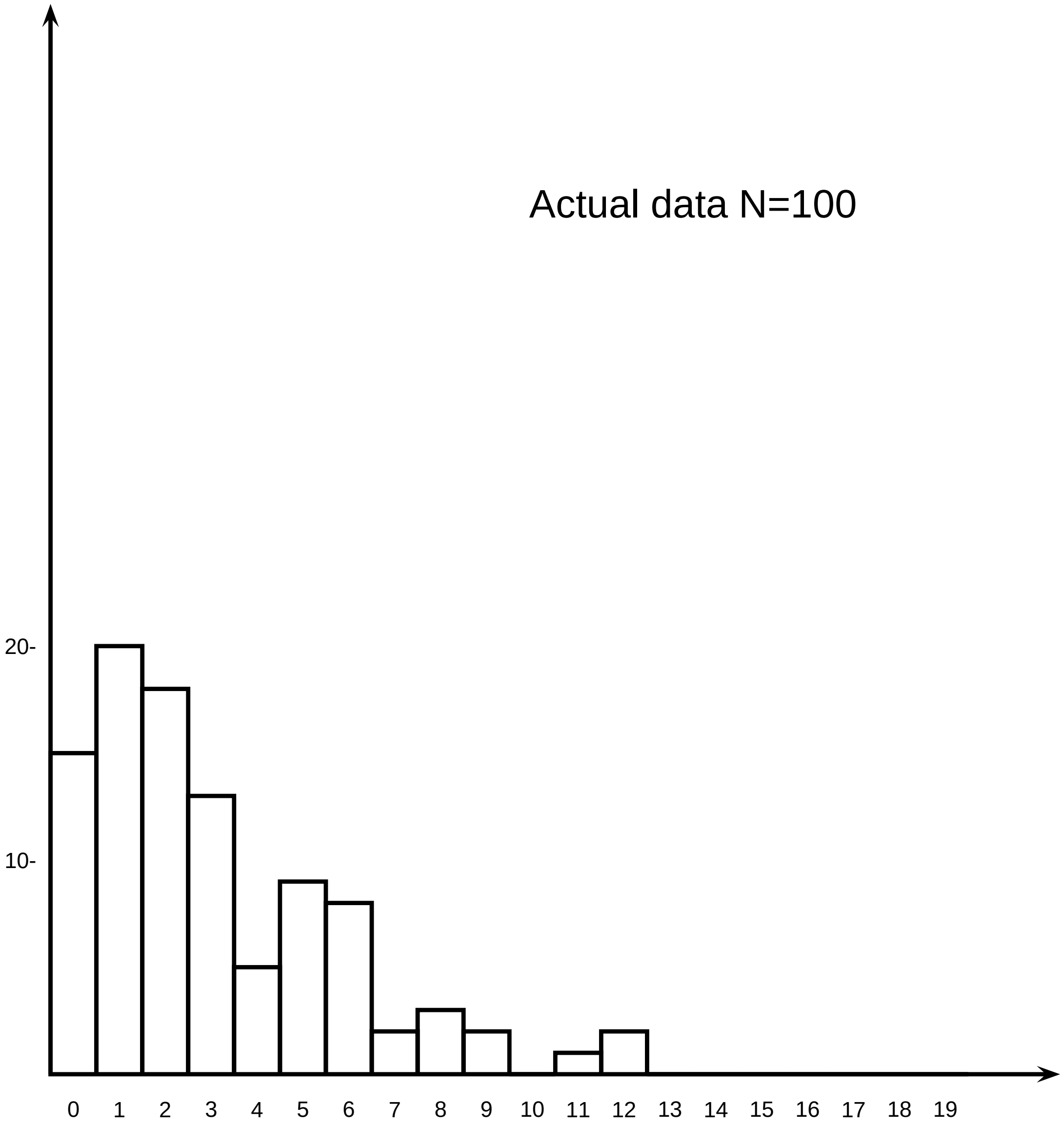
Sample X_{10}	Sample from exp($\overline{X_{10}}$)
0.780191	1.0678
0.81452	1.55077
1.81128	1.70147
2.37752	2.62435
3.11818	2.70093
4.4289	7.20934
5.19915	8.12043
7.63543	8.51629
9.71337	21.4475
92.6205	32.9842

The exponential sample has a central tendency of 8.79231 and a dispersion of 10.4667, which emphasises the obvious fact that **the sample size is way too small**.

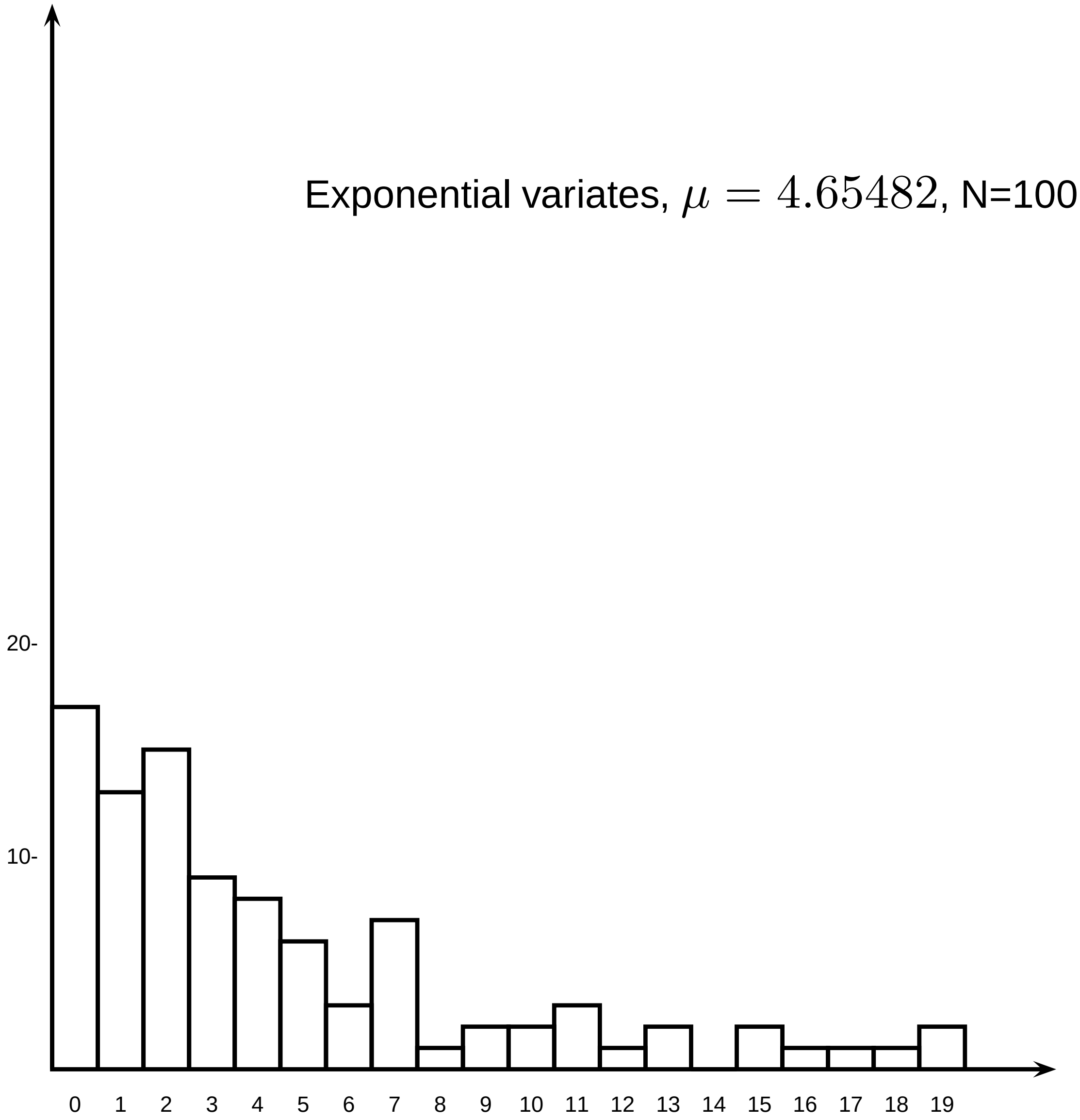
The trust in the validity of the sample increases as the sample is bigger. However, the bigger the sample, the more effort is needed to collect it, an obvious tradeoff.

A large amount of research was done to study the limits on the error that could be made as a function of the size of the sample. Well-known statistical methods allow to calculate the smallest sample size giving a satisfactory level of confidence.

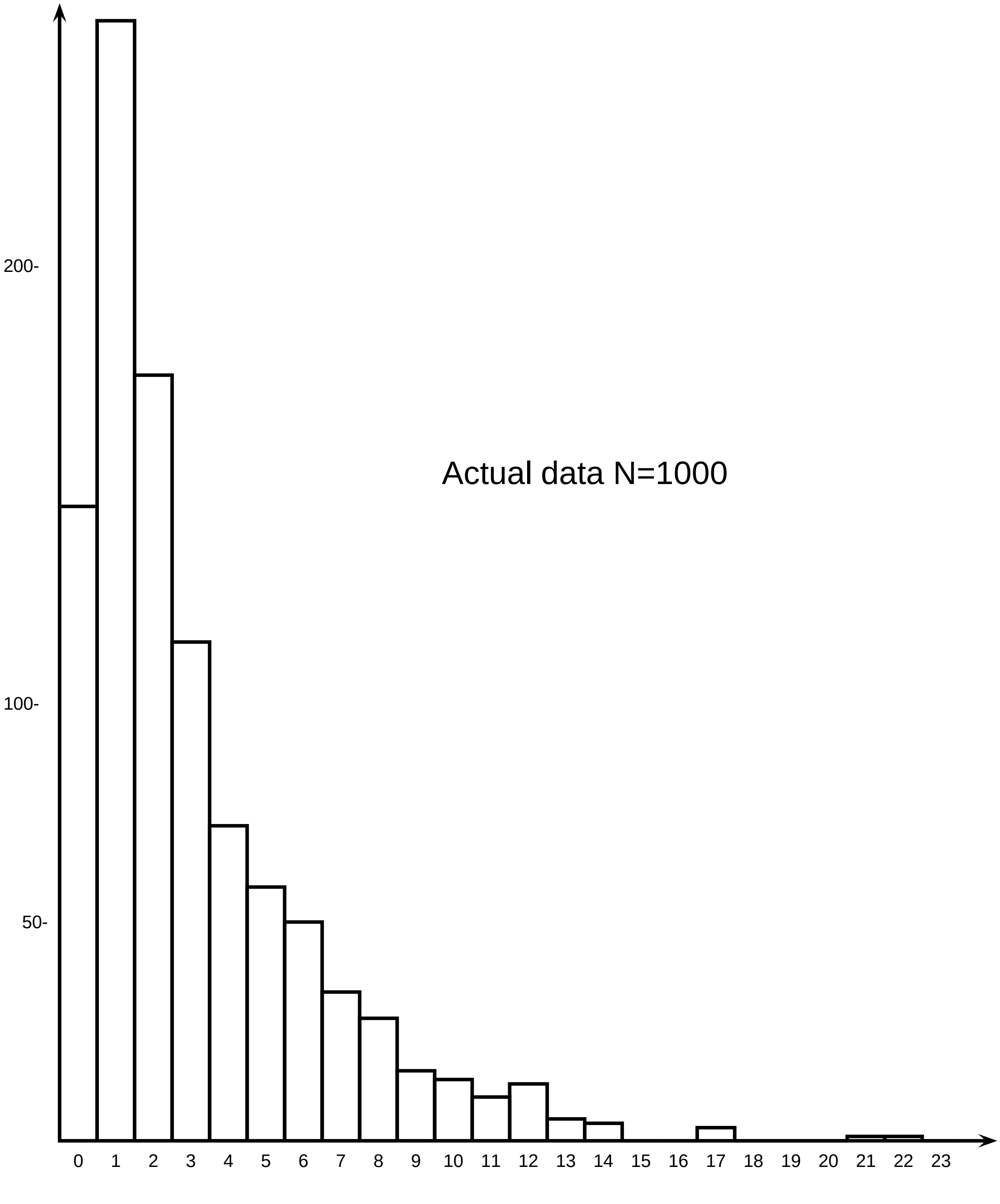
We will compare larger samples (they all are subsets of the same sample, so the sample of size 10 is made of the first 10 values from a sample of 10,000).



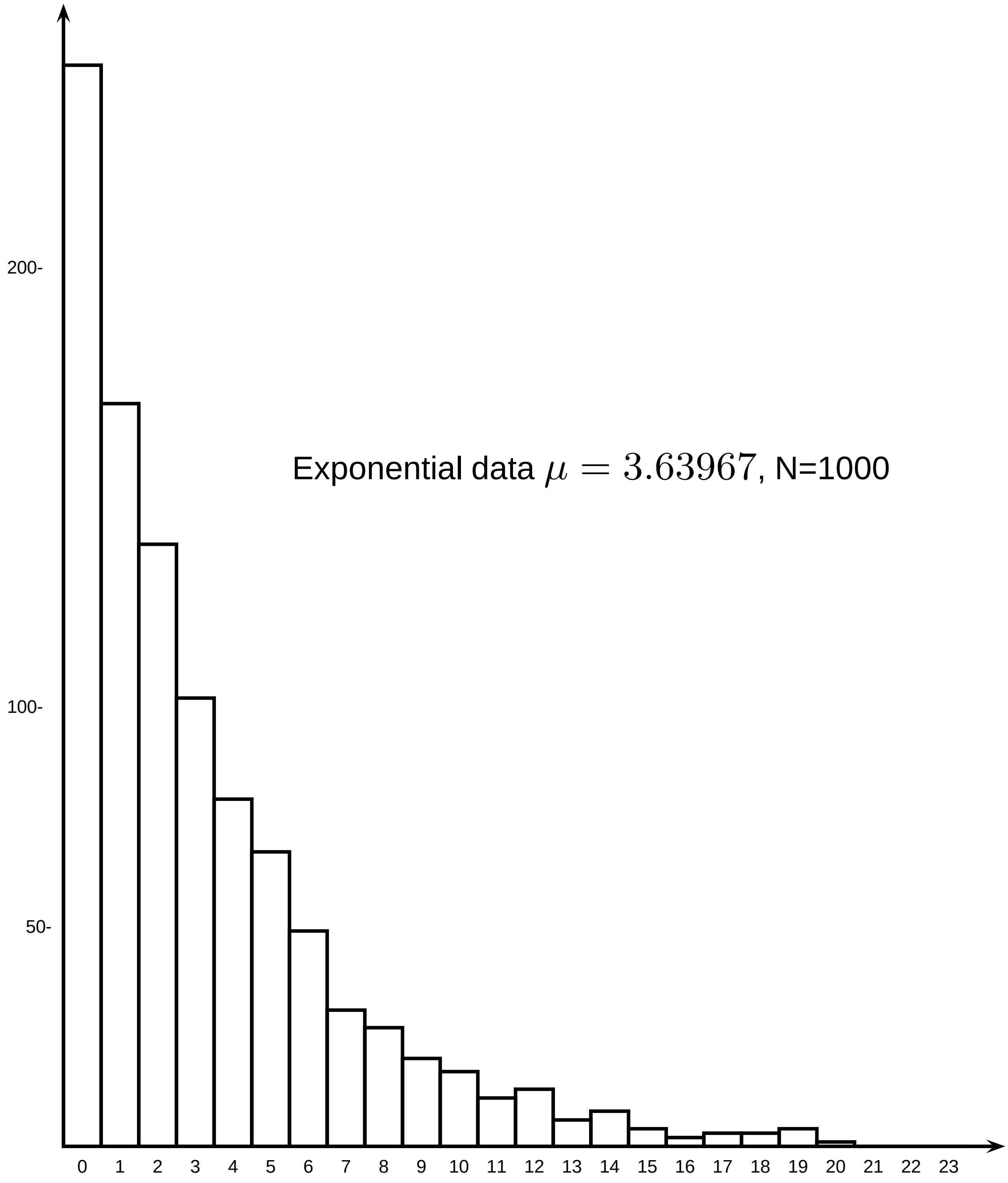
$$\overline{X_{100}} = 4.65482 \, S_{100} = 9.49841$$



Central tendency: 5.75857 Dispersion: 5.89394

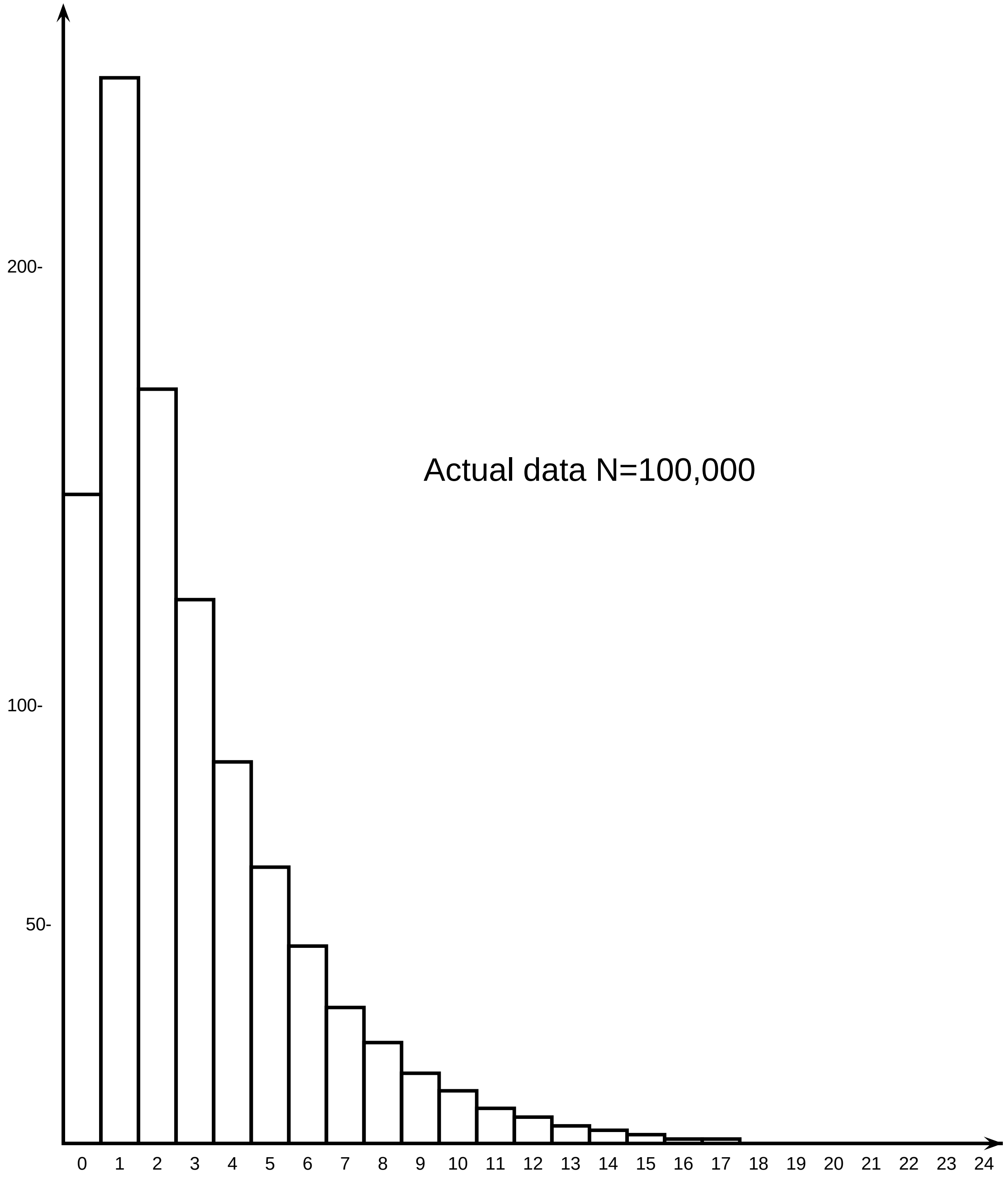


C.t. $\overline{X_{1000}} = 3.63967$ Dispersion = 4.19736

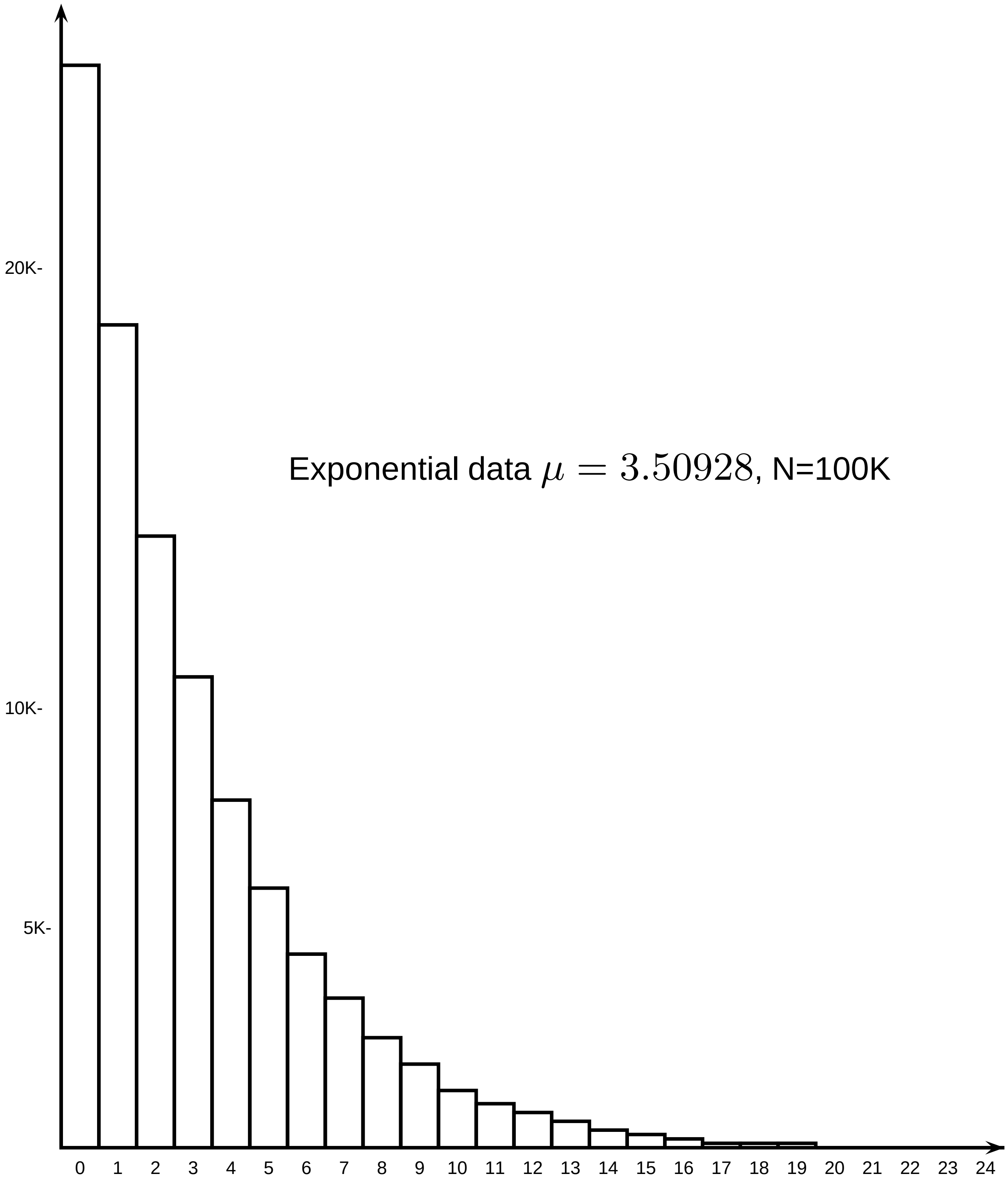


Exponential data $\mu = 3.63967$, N=1000

Central tendency: 3.74801 Dispersion: 3.77764



$\overline{X_{100K}} = 3.50928$ Dispersion = 3.0451



Exponential data $\mu = 3.50928$, N=100K

Central tendency: 3.51309 Dispersion: 3.50344

Parameter estimation

When we have a sample X and firmly believe that it comes from a specific distribution $\mathcal{X}$, we can estimate the parameters of $\mathcal{X}$ using data computed from the sample. A value derived from a sample that can be used to estimate a distribution parameter is called an **estimator**.

We can always compute two values from the sample: $\overline{x}$ and S (corresponding to the mean and standard deviation).

Here are some estimators:

Distribution	Parameter(s)	Estimator
Poisson	λ	$\hat{\lambda} = \overline{X}$
Exponential	μ	$\hat{\mu} = \frac{1}{\overline{X}}$
Geometric	p	$\hat{p} = \frac{1}{\overline{X}}$
Uniform	μ, σ	$\hat{\mu} = \overline{X}$
		$\hat{\sigma} = S$
Gamma	β, θ	$\hat{\beta}$ from table
		$\hat{\theta} = \frac{1}{\overline{X}}$
Normal	μ, σ^2	$\hat{\mu} = \overline{X}$
		$\hat{\sigma} = S$

Note that for the uniform distribution, estimating μ and σ allows to estimate the interval ends a and b.

Fitting

When we try to match a sample X to a random variable (i.e. distribution) $\mathcal{X}$, we start with a **hypothesis** called *“h null”*:

$$H_0\text{: } X \text{ was drawn from } \mathcal{X}$$

Obviously, this H_0 is true or it is false, but we don’t know which.

Statistics offers a way out: to make a “probabilistic” statement about H_0 such as:

> *I claim that H_0 is true with probability 0.95*

Statistics offer methods of proving a claim like this; such “proof” does not make H_0 true or false.

The possibility of making a mistake is expressed in terms of the **significance level** α:

$$\alpha = P(reject\, H_0 \mid H_0\, is\, true)$$

The **confidence** in our claim that H_0 is true clearly is $1 - \alpha$.

Note that a confidence level of 0.5 implies no confidence at all (probability of making a mistake being 50%).

A confidence level below 50% implies some degree of confidence that H_0 is false. In that case, the opposite hypothesis, H_1 appears to be likely to be true. H_1 is defined as:

H_1: X was not drawn from $\mathcal{X}$

χ^2 test

The most popular test for assessing the significance level of an H_0 hypothesis is the χ^2 test.

At the beginning we have a sample X made of n observations $x_i, i = 1, ..., n$ and a hypothesis:

$$H_0\text{: } X \text{ was drawn from } \mathcal{X}$$

Before starting, we need to determine s, the number of parameters of $\mathcal{X}$ that we can estimate using the sample X.

Some simple examples: for the uniform distribution, the mean and the variance can be estimated (or, with same result, a and b), hence $s = 2$. On the other hand, for the exponential distribution, $s = 1$ (this distribution has only 1 parameter to begin with).

The χ^2 test consists of several steps. Their description will be matched by the following example:

Sample X			
0.3	0.15	0.9	0.55
0.44	0.77	0.81	0.21
0.78	0.65	0.4	0.13
0.57	0.33	0.75	0.61
0.95	0.19	0.53	0.1

The sample is clearly too small; a much larger should be used in practice.

Our hypothesis is:

H_0: X was drawn from U(0,1)

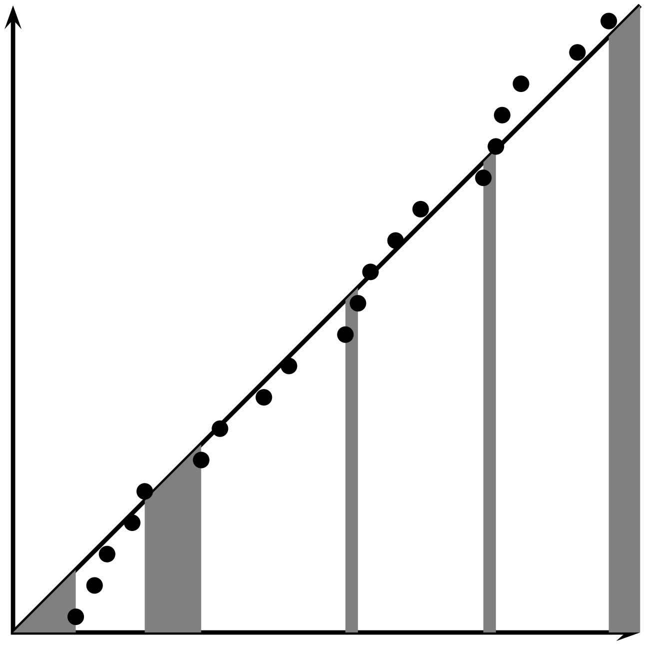
1. X is sorted, so that $\forall_{i<n}\, x_i \leq x_{i+1}$

Sorted sample X				
0.1	0.13	0.15	0.19	0.21
0.3	0.33	0.4	0.44	0.53
0.55	0.57	0.61	0.65	0.75
0.77	0.78	0.81	0.9	0.95

2. The sample is divided into k groups of adjacent values, each group O_i consisting of at least 5 observations. Each group is made of adjacent values, so it defines an interval (not necessarily unique) which can then be marked on the x–axis of the cdf curve of $\mathcal{X}$.

In the example, there is no choice but to divide X into 4 intervals with 5 observations in each.

4 intervals			
i	O_i	low	high
1	5	0.1	0.21
2	5	0.3	0.53
3	5	0.55	0.75
4	5	0.77	0.95



The interval ends are extended to cover the x–axis of the cdf of $U(0,1)$:

4 intervals			
i	O_i	low	high
1	5	0.0	0.25
2	5	0.25	0.54
3	5	0.54	0.76
4	5	0.76	1.0

The intervals are so "perfect" only by accident.

3. Now that we have the observed number of values in each interval, we compute the expected number of values in the same interval.

   For example if interval #2 is made of values lying in the interval (l_2, h_2) then the expected number of values in that interval will be:

$$n \times (F(h_2) - F(l_2))$$

   where F is the cdf of $\mathcal{X}$. This gives the values E_i for $i = 1, .., k$.

   In the example, $F(x) = x$

i	O_i	low	high	E_i
1	5	0.0	0.25	5
2	5	0.25	0.54	6
3	5	0.54	0.76	4
4	5	0.76	1.0	5

4. Compute the value:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Note that the denominator is E_i, so that the value of χ^2 is unbounded.

$$\frac{0}{5} + \frac{1}{6} + \frac{1}{4} + \frac{0}{5} = \frac{5}{12} = 0.42$$

5. Check in the χ^2 table the value of $\chi^2_{\alpha,k-s-1}$. If $\chi^2 > \chi^2_{\alpha,k-s-1}$ the hypothesis H_0 can be rejected with a probability of making an error no greater than α. In the example, $k - s - 1 = 4 - 2 - 1 = 1$. $\chi^2_{0.005,1} = 7.88$ and H_0 cannot be rejected with significance 0.005, hence we accept H_0 as true.

Comments on χ^2

The main purpose of the χ^2 test is to reject hypotheses that are not sufficiently likely. Hence several different variations of χ^2 exist. Some variations:

- Always use $s = 1$ (Knuth). This actually reduces the ability to reject H_0.
- Instead of insisting that all $O_i \geq 5$, insist that all $E_i \geq 5$. This method avoids rejecting H_0 because of a division by 0; it is not clear if this is good or bad.
- Variation of the previous method: insist that all E_i are equal. Very convenient for the uniform distribution, not so for many other.

The recommended value of k and a function of the sample size n.

1. Never use χ^2 for $n < 25$.
2. For $n \geq 25$ use $\sqrt{n} \leq k \leq n/5$.

Finally, goodness–to–fit tests may be approached from another angle. Similarly to α, called “Type I error” and defined as:

$$\alpha = P(reject\, H_0 \mid H_0\, is\, true)$$

we define a “Type II error” β as

$$\beta = P(fail\, to\, reject\, H_0 \mid H_0\, is\, false)$$