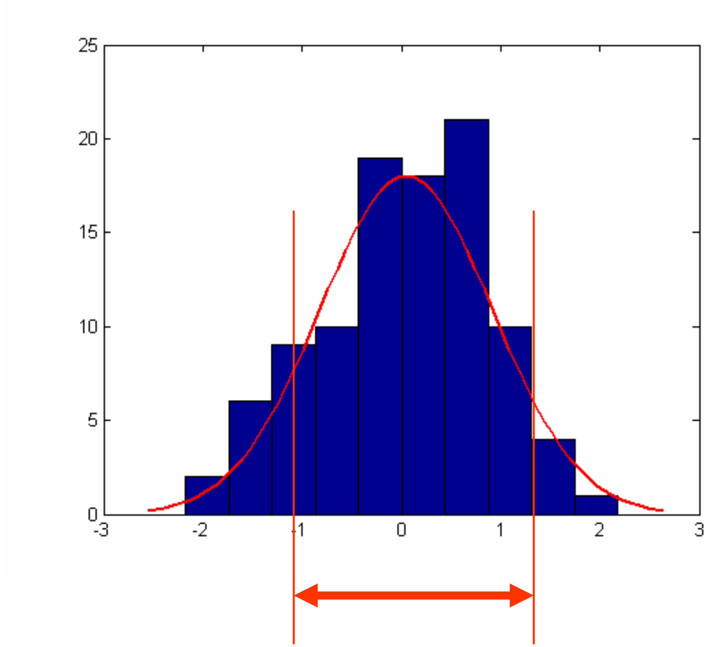
The concept of Dispersion

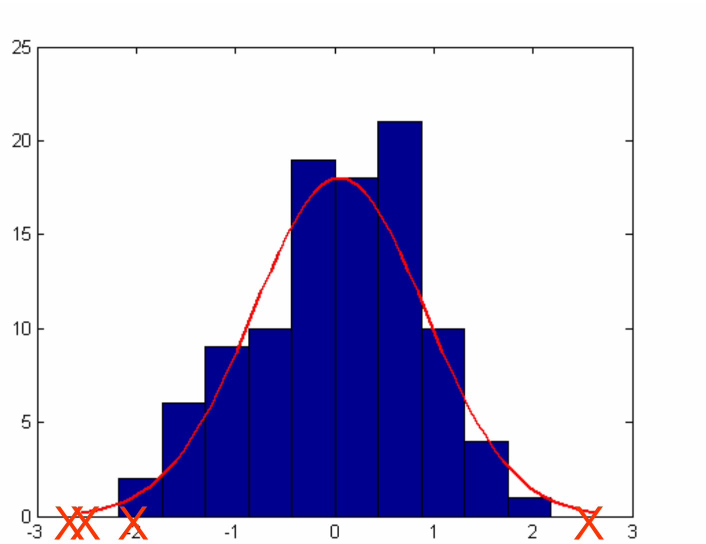
The concept

- Law of large numbers tells us that the average result will -> E(X)
- What can dispersion can we expect arround E(X)
- Does it have a sense to use the concept of dispersion?

The concept

- Given a set of data or a functional distribution we would like to have a number for comparing dispersions.



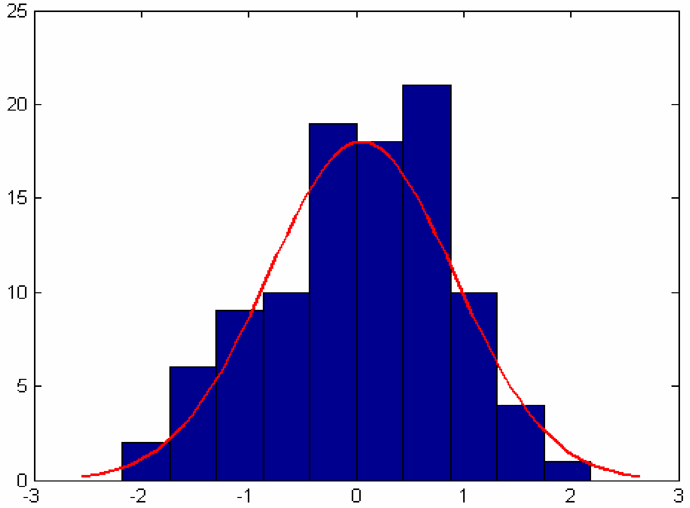


Extreme points

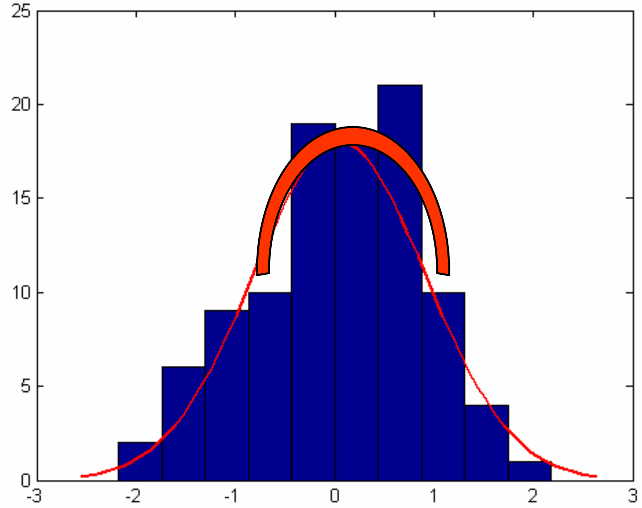
Margin that embraces 50 % of all points

The concept

- Given a set of data or a functional distribution we would like to have a number for comparing dispersions.



Mean distance to the center of mass of the distribution



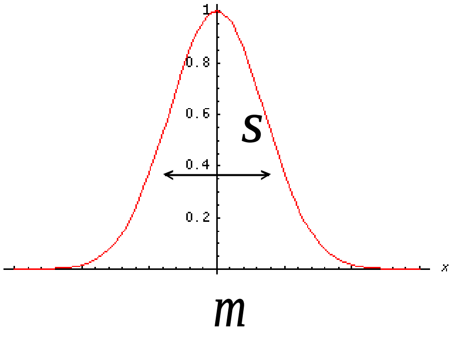
Measure of curvature: Second degree polinomial

The measures

- Standard deviation

$$m(X) = E(X) = \sum_{x \in X} xP(X = x)$$

$$s^2(X) = E\left((X - m)^2\right) = \sum_{x \in X} (x - m)^2 P(X = x)$$



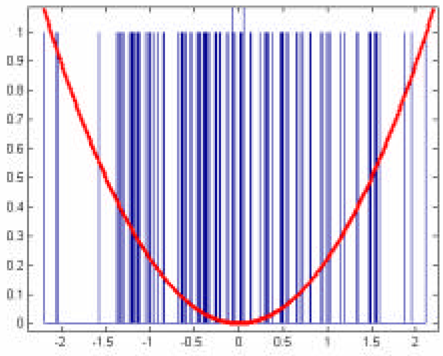
- Mean absolute distance

$$m(X) = E(X) = \sum_{x \in X} xP(X = x)$$

$$MAD(X) = E\left(\left|X - m\right|\right) = \sum_{x \in X} \left|X - m\right| P(X = x)$$

The measures

- Standard deviation
- Justification:
  - Samples far from the mean increase quadratically the value of the standard deviation.



$$s^2(X) = E\left(\left(X - m\right)^2\right) = \sum_{x \in X} \left(x - m\right)^2 P(X = x)$$

  - Is the value of the curvature of the log gausian.

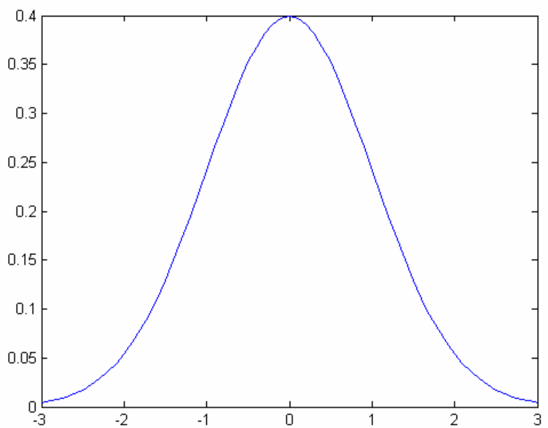
The measures

Standard deviation

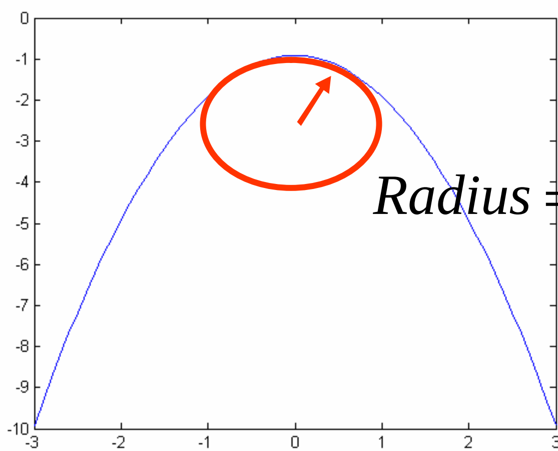
Meaning of the curvature of a plane curve:

– Radius of a circle that aproximates locally the curve. $\kappa = \frac{\frac{d^2y}{dx^2}}{(1+(\frac{dy}{dx})^2)^{3/2}}$ $\kappa \approx \frac{d^2y}{dx^2}$

$$p(x) = \frac{1}{\sqrt{2p}\, s} e^{-\frac{(x-m)^2}{s^2}}$$



$$\log(p(x)) = -\frac{(x-m)^2}{s^2} + \log(\frac{1}{\sqrt{2p}\, s})$$



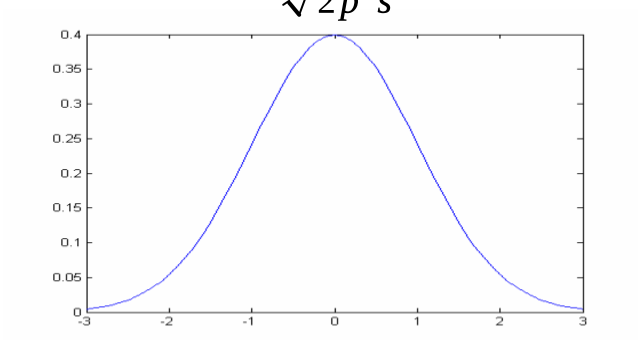
$$Radius = \frac{1}{\frac{d \log(p(x))}{dx}} = s^2$$

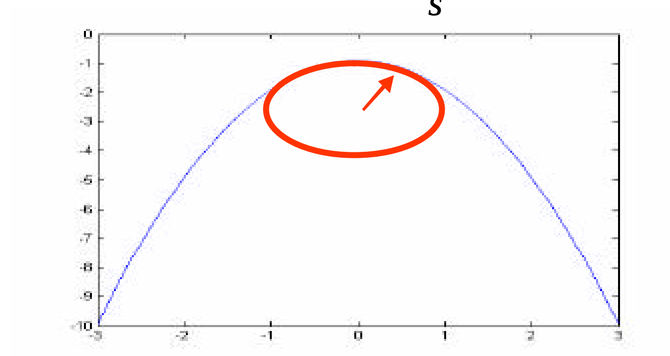
The measures

Standard deviation

Meaning of the curvature of a plane curve:

$$p(x) = \frac{1}{\sqrt{2p}\,s} e^{-\frac{(x-m)^2}{s^2}} \qquad \log(p(x)) = -\frac{(x-m)^2}{s^2} + \log\left(\frac{1}{\sqrt{2p}\,s}\right)$$





– Certainty=low dispersion->Small radius->small variance

– Uncertainty=high dispersion->big radius->big variance

$$Radius = \frac{1}{\frac{d \log(p(x))}{dx}} = s^2$$

The measures

Standard deviation

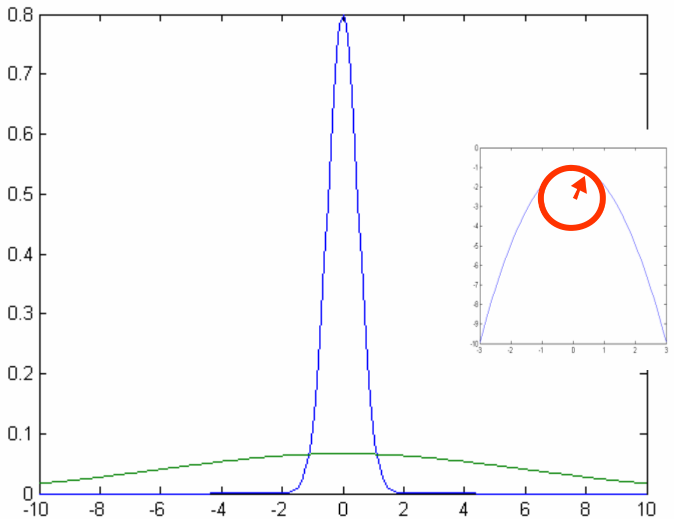
– Certainty=low dispersion->Small radius->small variance

– Uncertainty=high dispersion->big radius->big variance

$$Radius = \frac{1}{\frac{d\log(p(x))}{dx}} = s^{2}$$

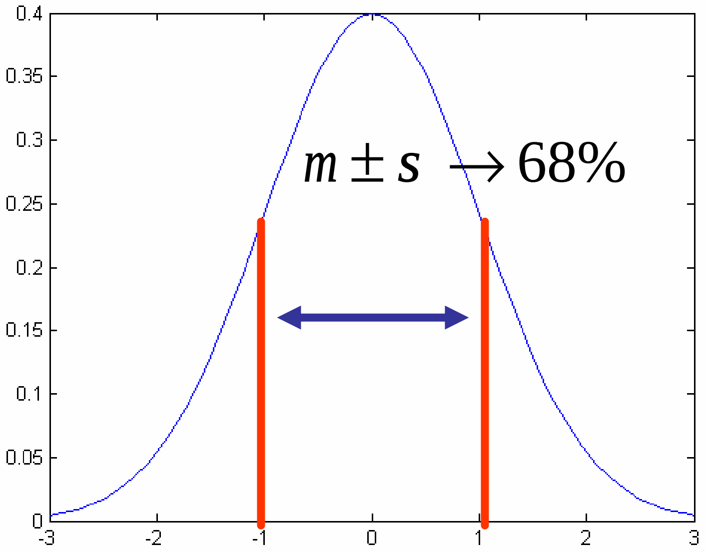
$$s = 0.5$$

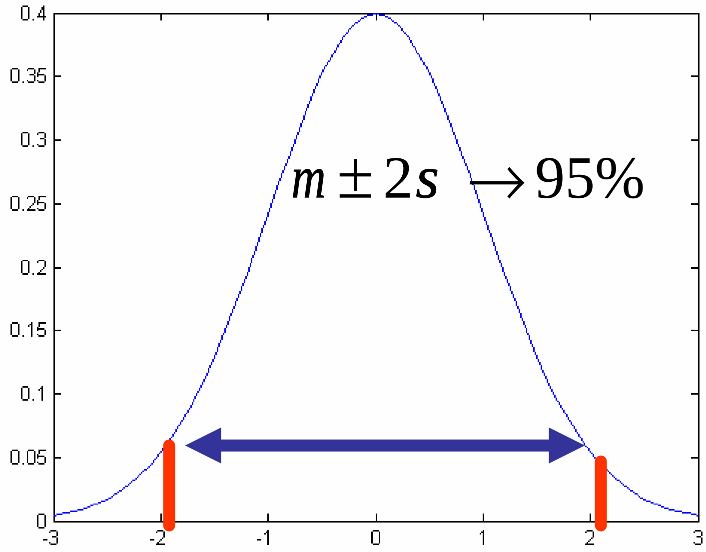
$$s = 6$$



Information about Gausian distributions.

- % of the cases for a given dispersion





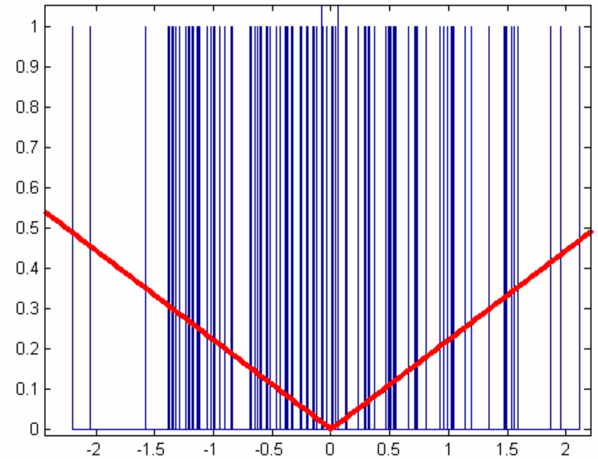
$$m \pm 3s \rightarrow 99\%$$

The measures

- Mean absolute distance

$$m(X) = E(X) = \sum_{x \in X} xP(X = x)$$

$$MAD(X) = E(|X - m|) = \sum_{x \in X} |X - m| P(X = x)$$

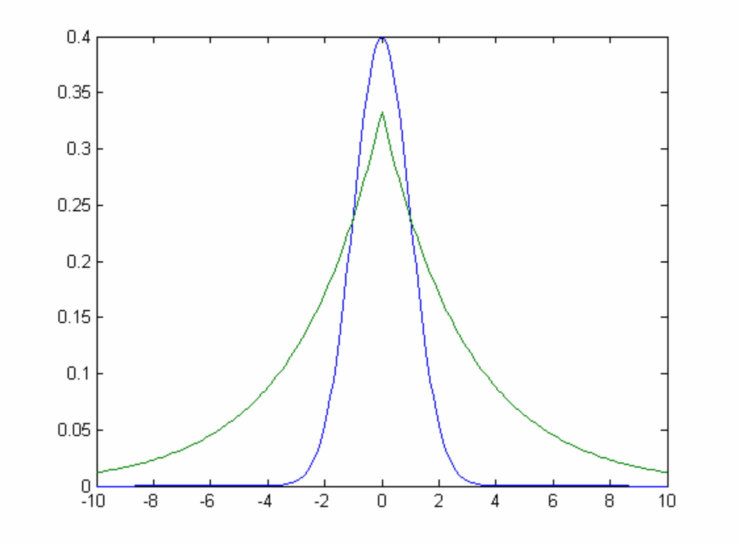


Samples far from the mean increase linearly the value of the standard deviation. Note: outliers do not count as much as in the standard deviation.

The measures

- Mean absolute distance
  - Uses: measure of dispersion for distributions with longer tail than gausians.

$$MAD(X) = E\left(\left|X - m\right|\right) = \sum_{x \in X} \left|X - m\right| P(X = x)$$

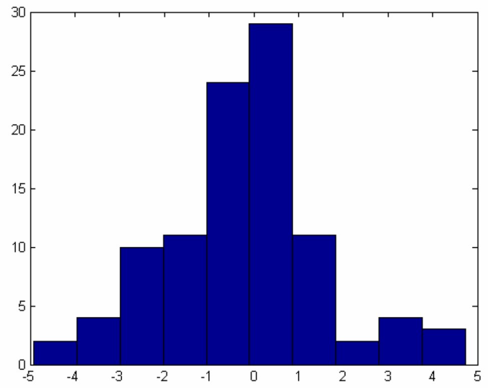


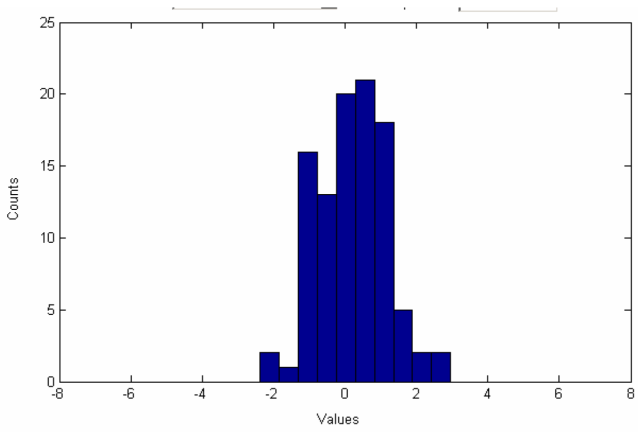
Laplace vs. Gausian

Note: That high values are much more probable in the case of an exponential random variable

The measures

- Mean absolute distance
  – Outliers do not change so much the value of the estimation.

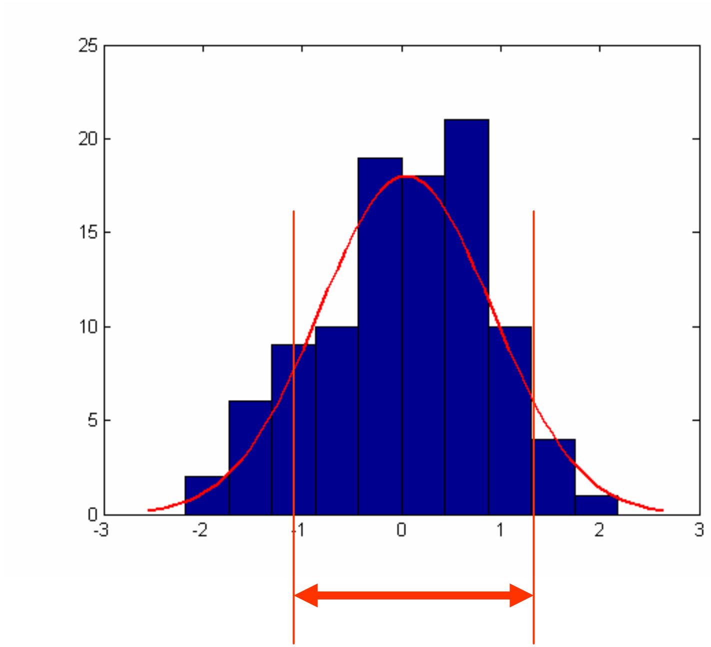




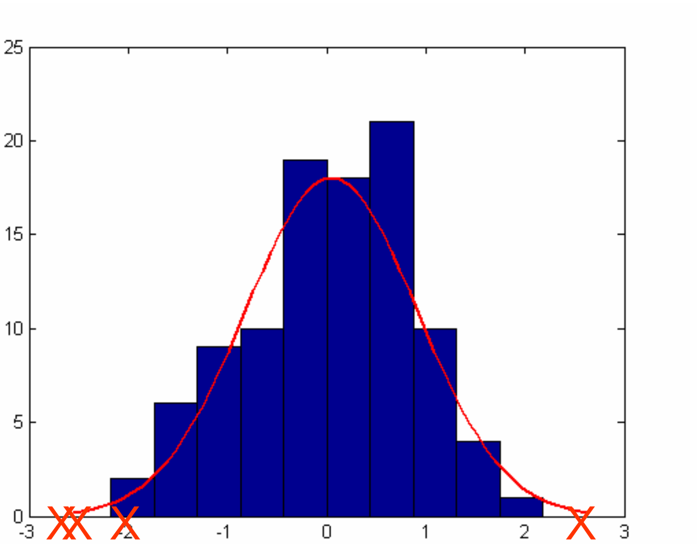
$$MAD(X) = E(|X - m|) = \sum_{x \in X} |X - m| P(X = x)$$

The concept

- Given a set of data or a functional distribution we would like to have a number for comparing dispersions.



Margin that embraces 50 % of all points



Extreme points

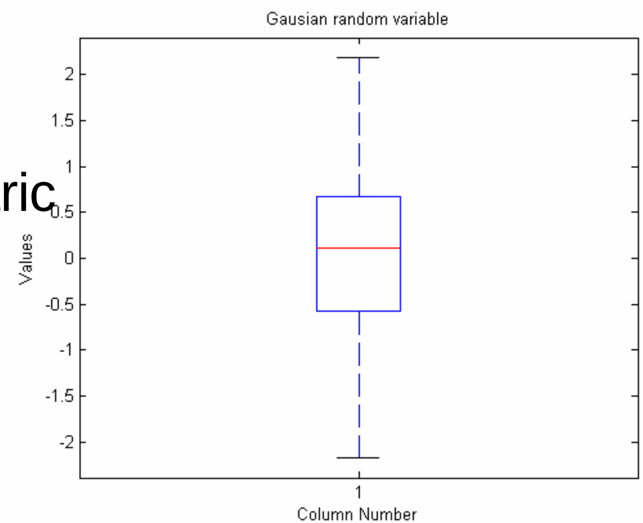
The Box Plot

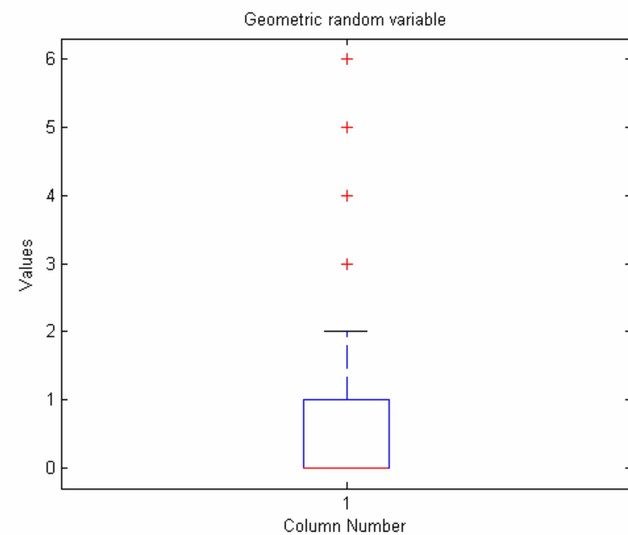
- Gives information about the dispersion summarizing the information about:
  - The interquantile size. x_{25}-x_{75}
  - Median
  - Sample nearest to the 1.5 times the interquantile margin
  - The outliers   (points >1.5 IQR)

- Examples:

  100 points

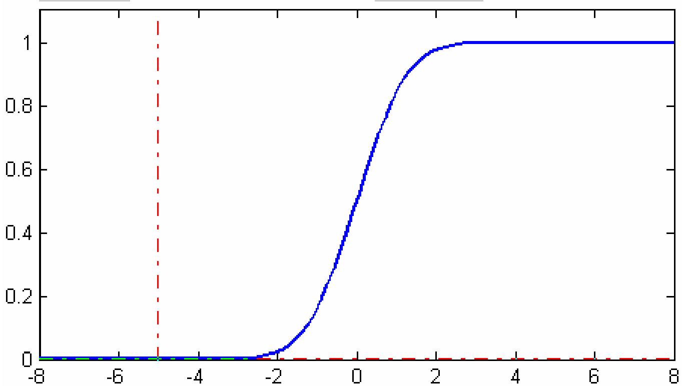
Gausian vs. geometric

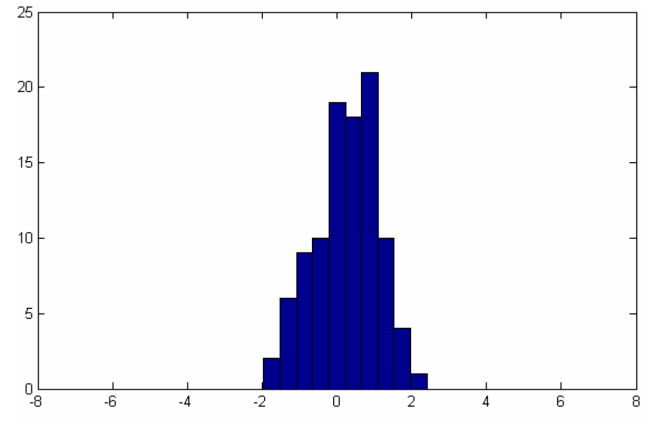


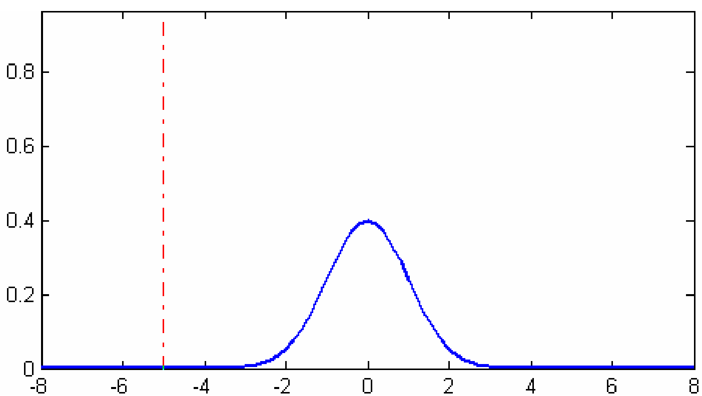


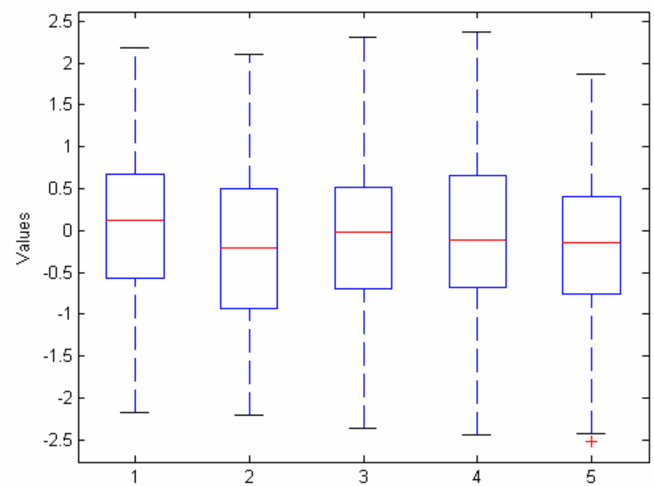
Description of a random variable

- Gausian $p(x) = \frac{1}{\sqrt{2p\,s}} e^{-\frac{(x-m)^2}{s^2}}$



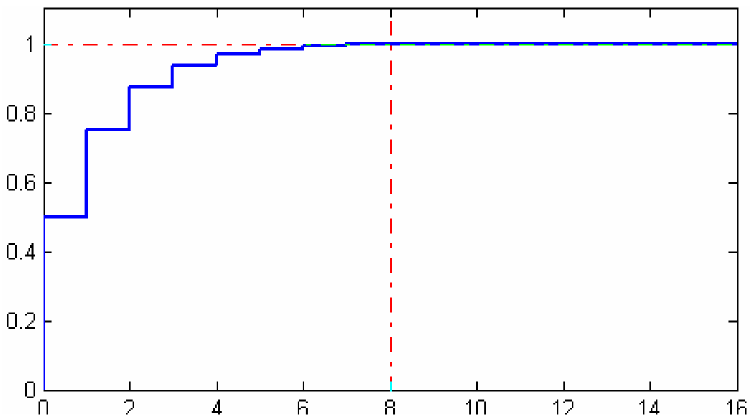


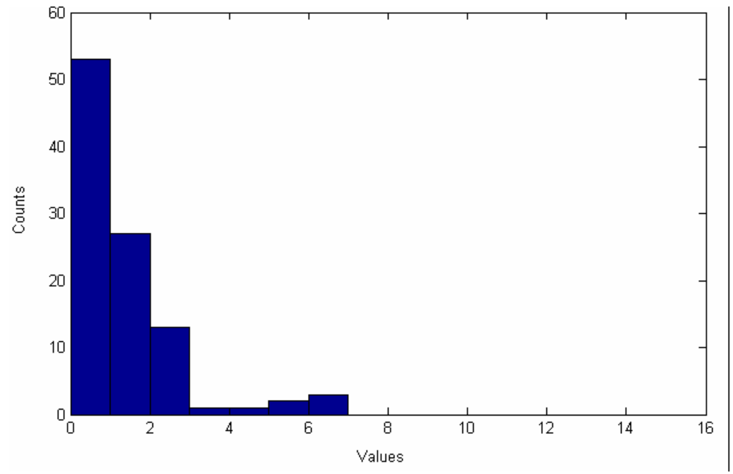


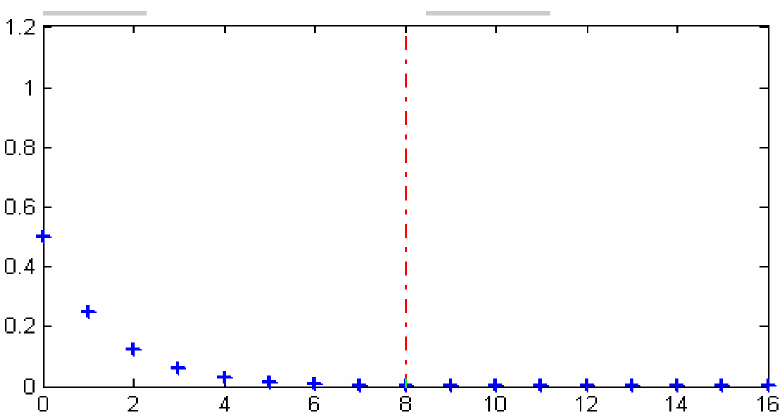


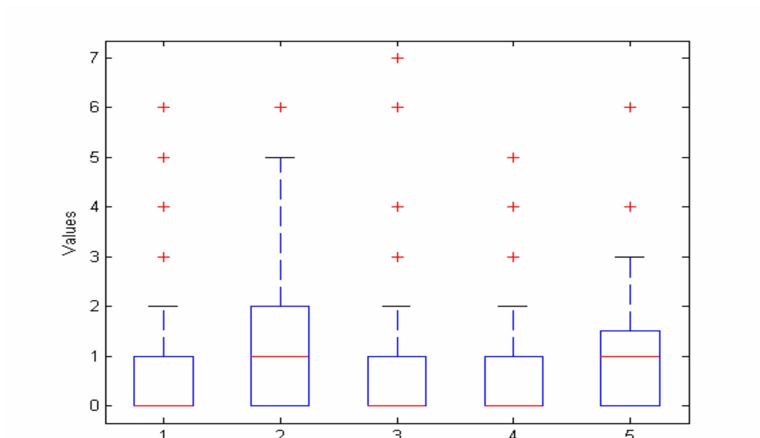
Description of a random variable

- Geometric $$P(X=i)=p^{i-1}(1-p) \quad \text{for } i=1,2,3,\cdots$$



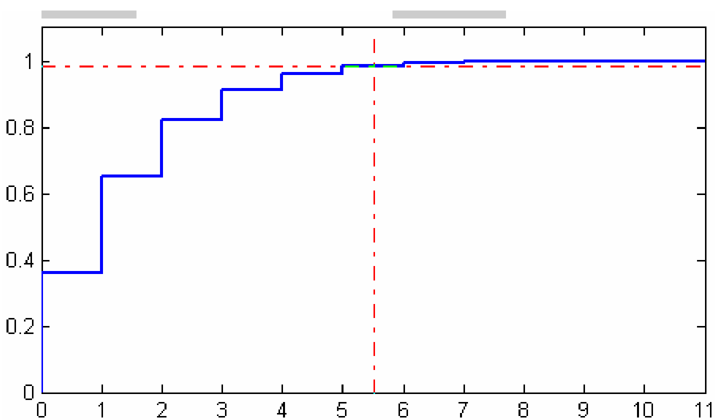


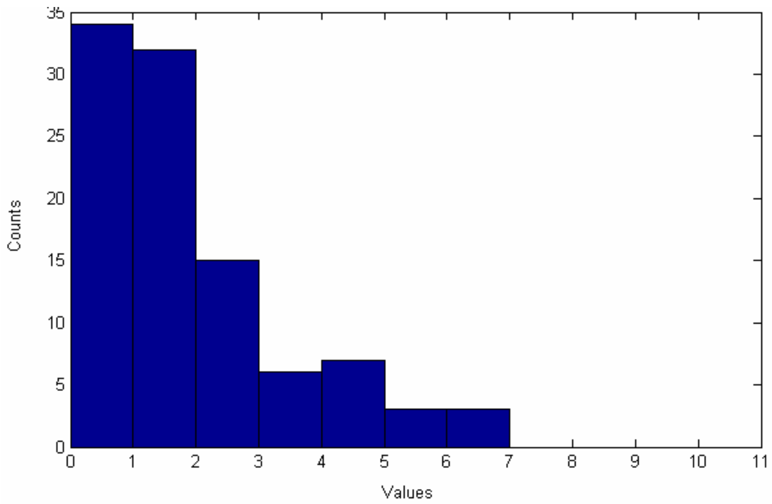


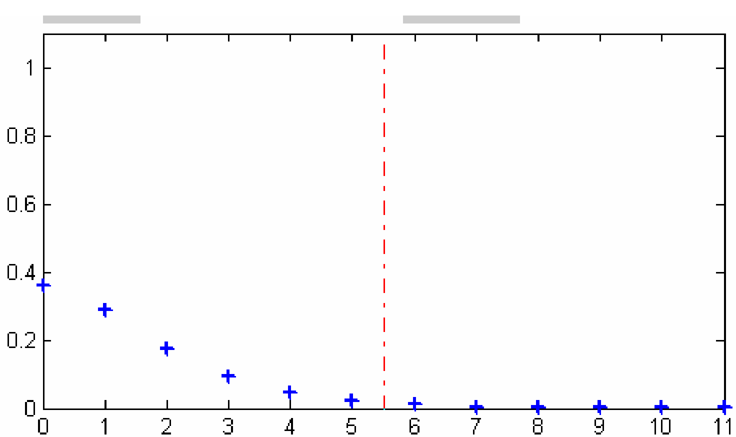


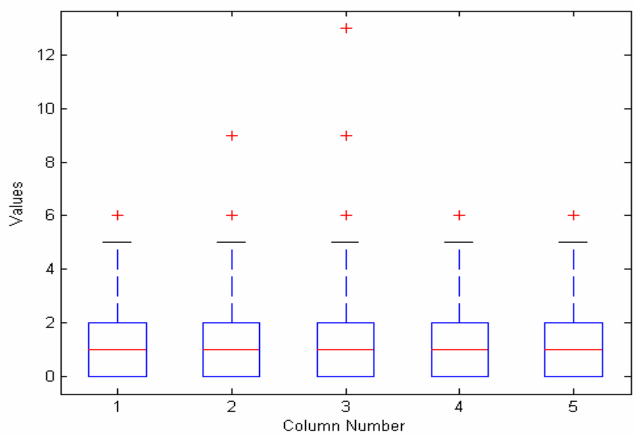
Description of a random variable

- Negative binomial



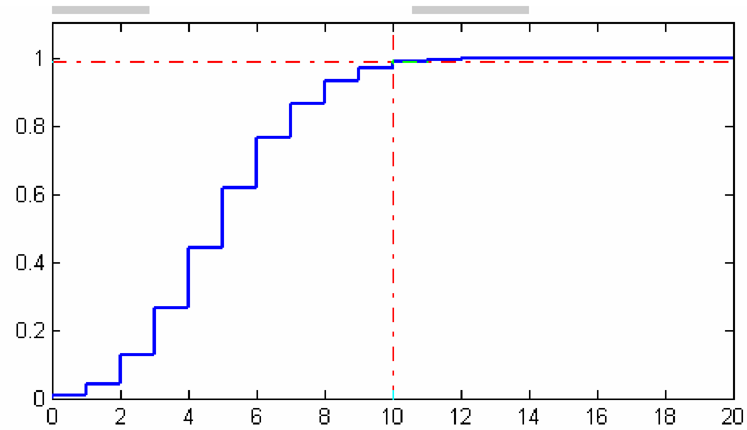


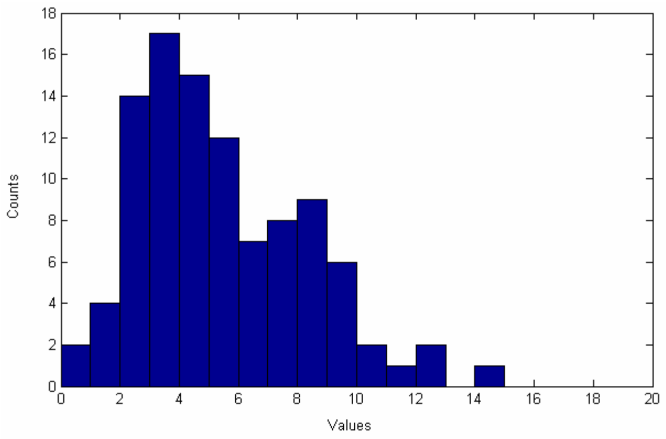


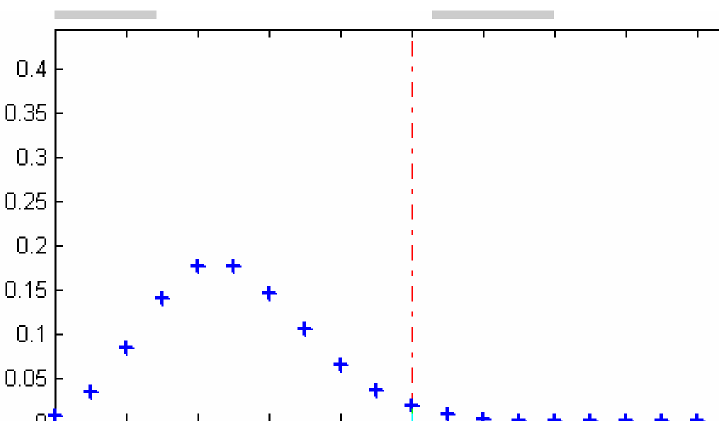


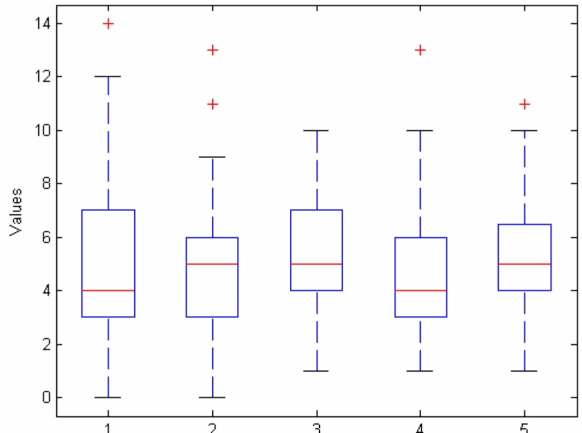
Description of a random variable

- Poisson



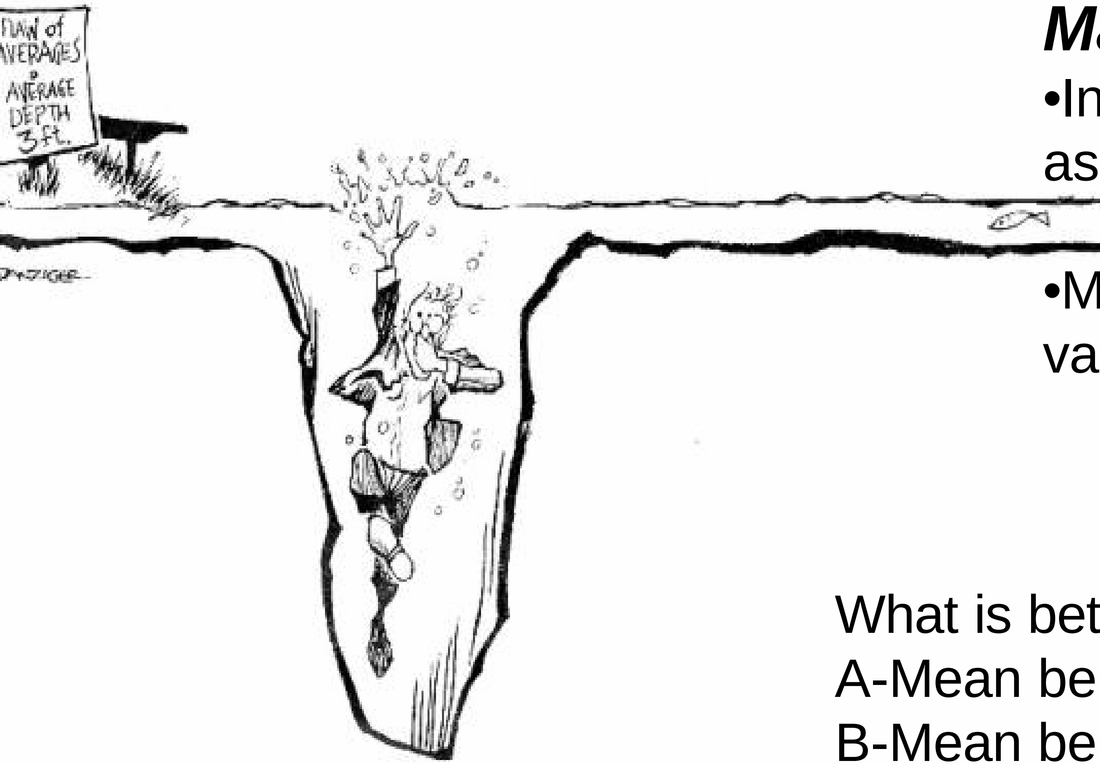






Some intuitions

- The ***flaw*** of averages



Markowitz's Idea:

•Introduce variability when assesing the value of an asset.

•Maximize mean, minimizing variance.

What is better?:

A-Mean benefit of 500 plus minus 400

B-Mean benefit of 200 plus minus 50

http://www.stanford.edu/%7Esavage/flaw/Article.htm

The flaw of averages

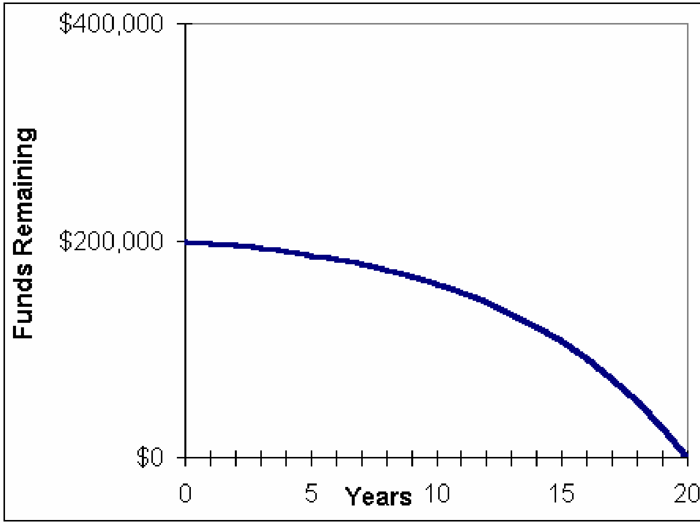
- An investment problem
  - Suppose you want your \$200,000 retirement fund invested in the Standard & Poor's 500 index to last 20 years. How much can you withdraw per year?
  - The return of the S&P has varied over the years but has averaged about 14 percent per year since 1952.
  - If you do this you will be pleased to find that you can withdraw \$32,000 per year.

$$A(1+r)^{20} - \sum_{k=0}^{19} x(1+r)^k = 0$$

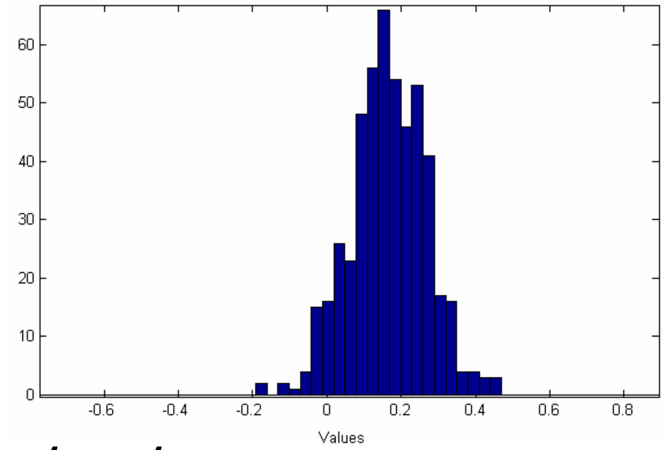
$$A = 200.000$$

$$r = 14\%$$

$$x = \frac{rA(1+r)^{20}}{(1+r)^{20} - 1}$$



The flaw of averages

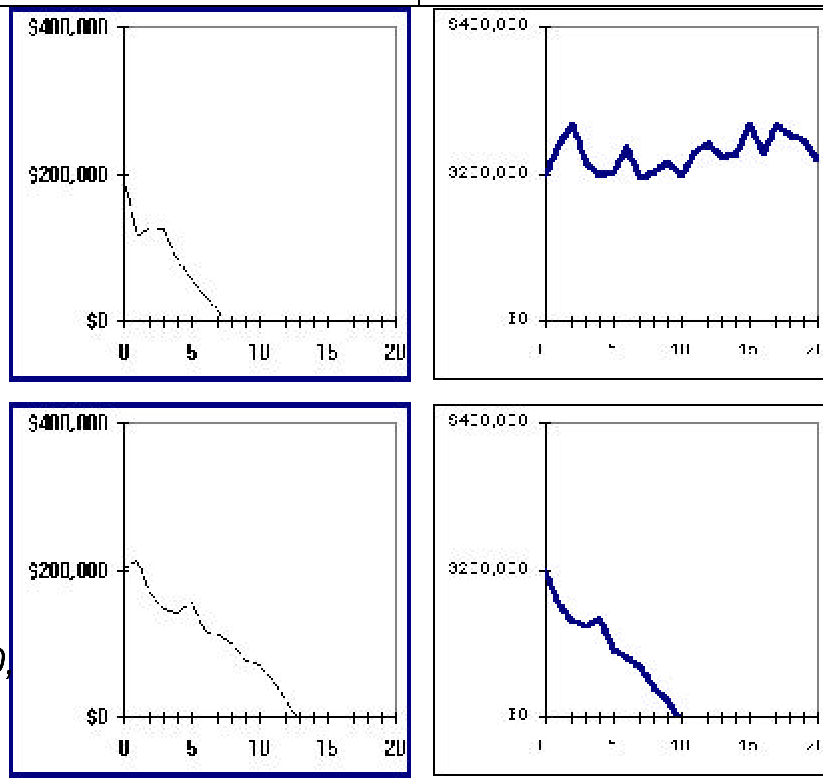


- Model of the return:
  - *r* % with probability p
  - *Histogram of the return*
  - *Average value r, but can fluctuate.*
    - Sometimes gives benefits
    - Sometimes losses
  - Note that each month a fixed quantity is substracted independently of *r*

The flaw of averages

- Simulations on real data:

Start: 1973 Avg. Return 14% Tanks in 8 yrs.	Start:1974 Avg Return 15.4% Goes the distance.
Start: 1975 Avg. Return 15.4% Tanks in 13 yrs.	Start: 1976 Avg. Return 15.3% Tanks in 10 yrs



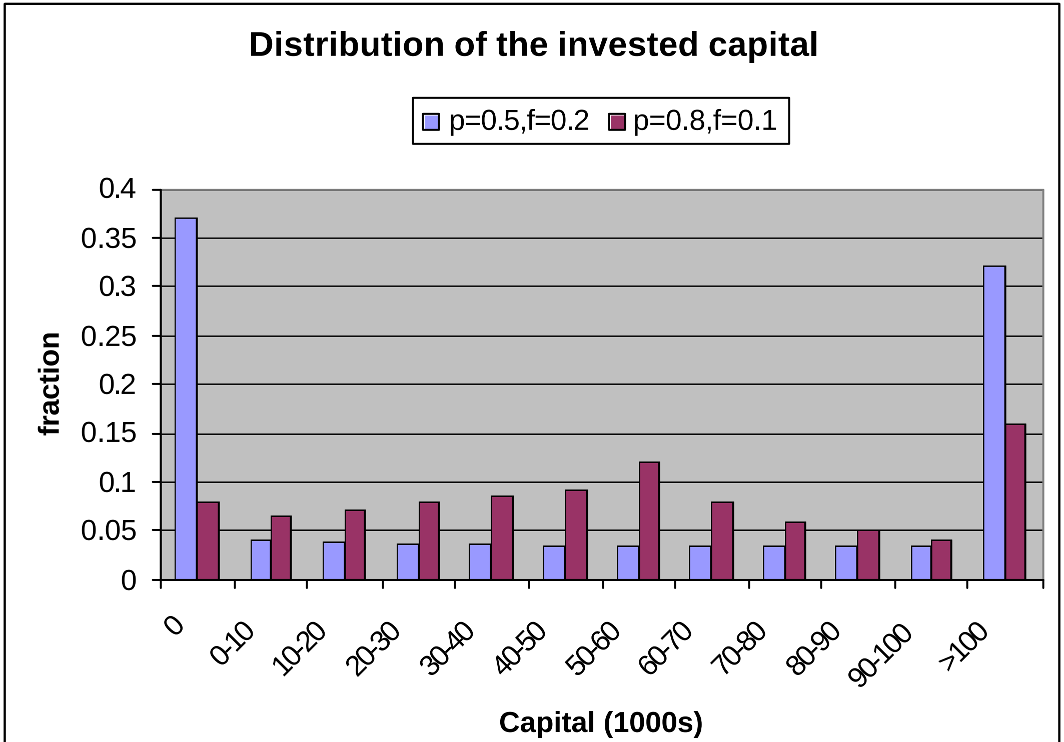
The Flaw of Averages

BY SAM SAVAGE

Published Sunday, October 8, 2000, in the San Jose Mercury News

The flaw of averages

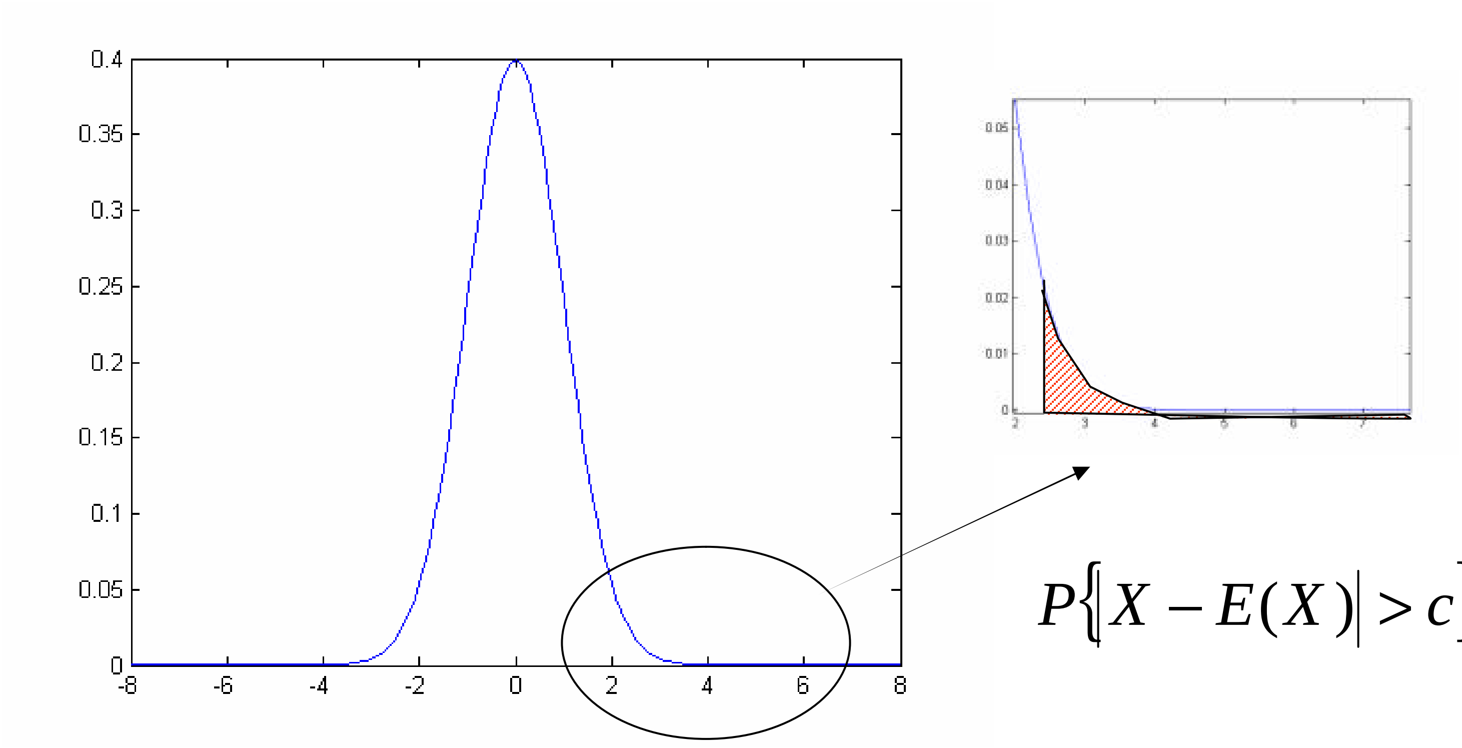
- Model of the return:
  - *r* % with probability p
  - *(1+f)r* % with probability (1-p)/2
  - *(1-f)r* % with probability (1-p)/2
- Simulation.
  - 4.000.000 runs.



aken fromTijms, Understanding probability

Tchebychev Inequality

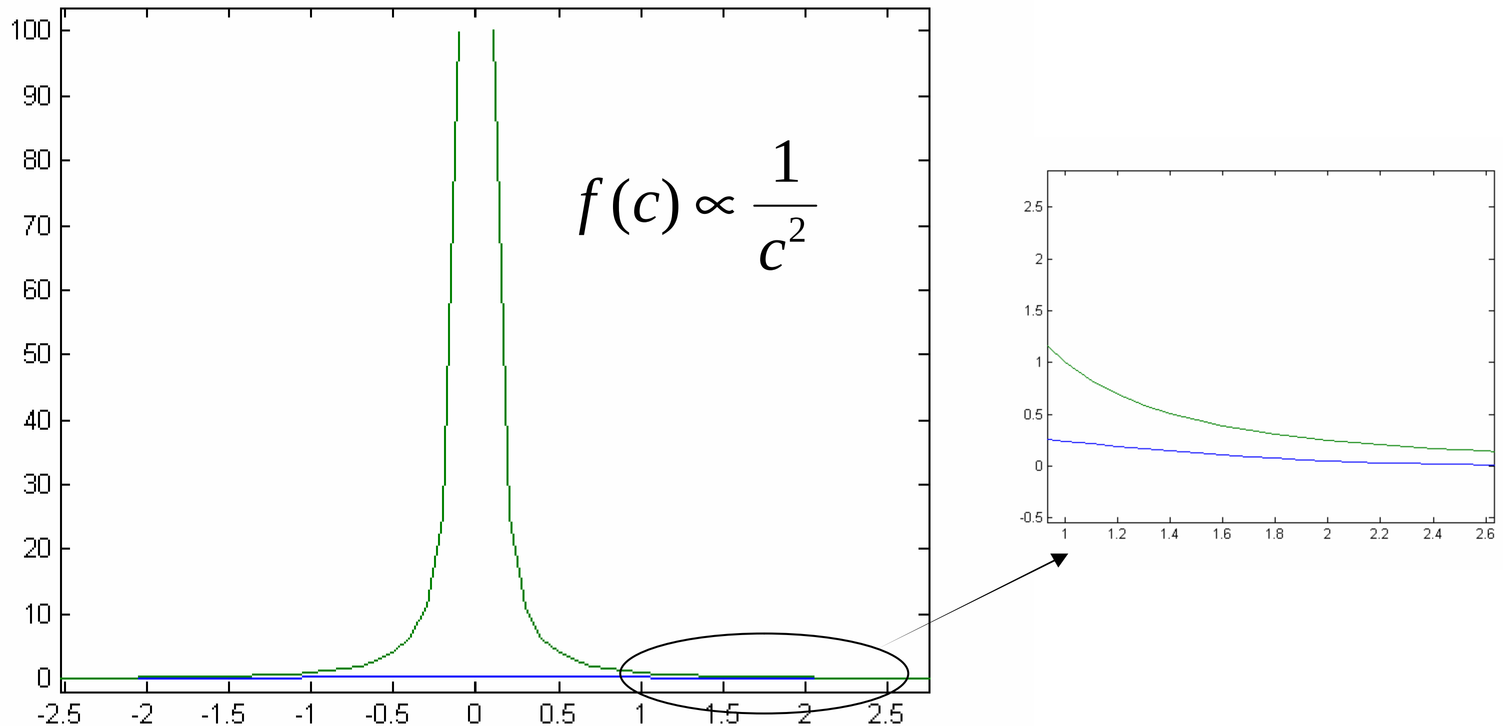
- A bound on the upper probability.



$$P\{|X - E(X)| > c\} \leq \frac{s^2}{c^2}$$

Tchebychev Inequality

- Geometrical meaning of $P\{|X - E(X)| > c\} \leq \frac{s^2}{c^2}$



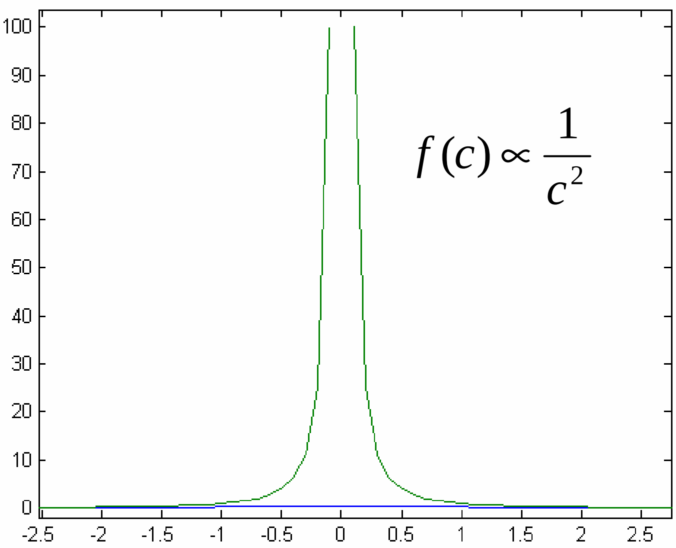
Tchebychev Inequality

- Geometrical meaning of $P\{|X - E(X)| > c\} \leq \frac{s^2}{c^2}$

- What happens with?

$$P(X = x) = \frac{1}{p\left(1 + x^2\right)}$$

$$P(X = x) = \frac{1}{x^a}$$



Only valid on distributions that have finite variance!

Tchebychev Inequality

- For a given probability distribution of a random variable X, with finite variance we have:

$$P\{|X - E(X)| > ks\} \leq \frac{1}{k^2}$$

- for any k>0 or equivalently

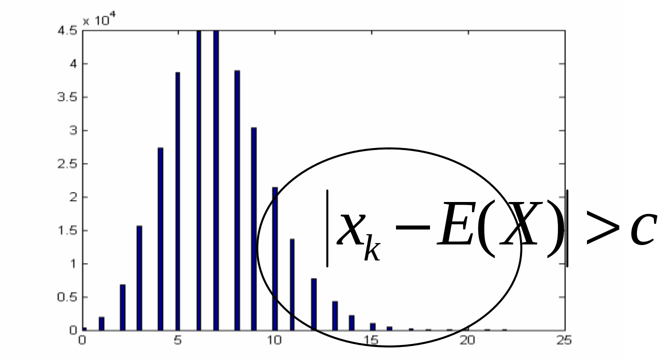
$$P\{|X - E(X)| > c\} \leq \frac{s^2}{c^2}$$

Tchebychev Inequality

- Proof
  - Given an ordered set $\{x_1, x_2, x_3, x_4, x_5, \cdots\}$
  - We define the subset $A = \{k \mid \; |x_k - E(X)| > c\}$
  - then

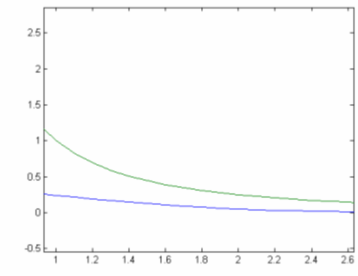
$$s^2 = \sum_k (x_k - E(X))^2 p_k \geq \sum_{k \in A} (x_k - E(X))^2 p_k \geq c^2 \sum_{k \in A} p_k$$

$$s^2 \geq c^2 P\{|X - E(X)| > c\}$$

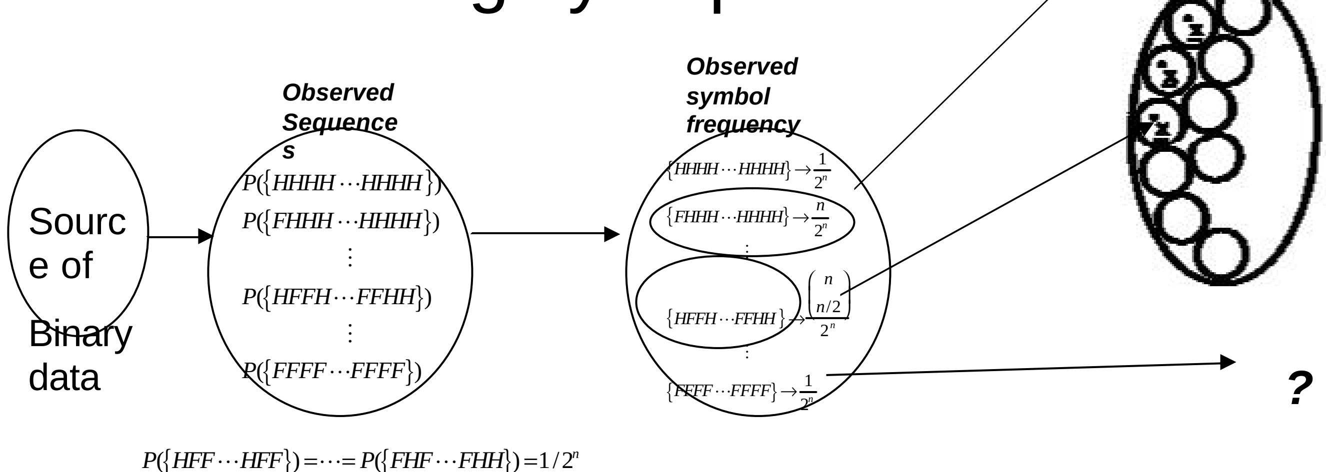


Tchebychev Inequality

- Note that the inequality can be rough, and highly inexact for high values of *c*



- Uses:
  - Information theory, bounds on probabilities and events highly unprobable.

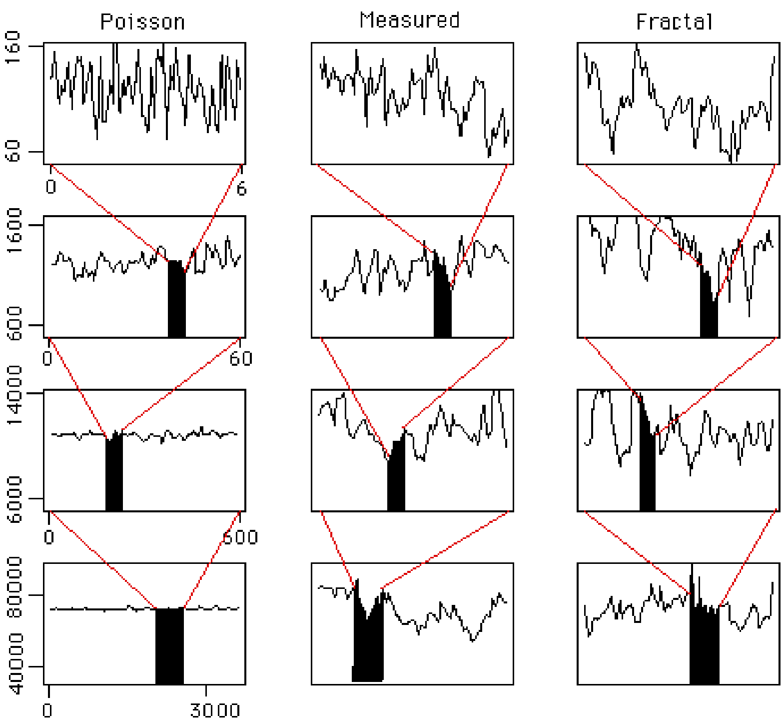


Random variables without variance

- Family known as
  - Pareto Stable or Mandelbrot Levy
- Models:
  - Internet traffic
  - Processes in unix systems
  - Speculative prices/Pluviometric Data

Speculative Prices

- Mandelbrot’s paper on long tail densities
  - An interesting result



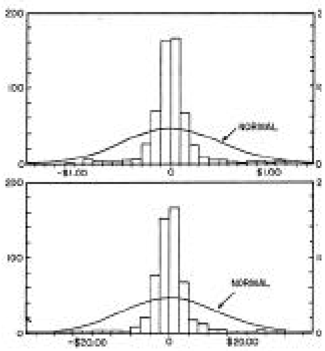


FIG. 1.—Two histograms illustrating departure from normality of the fifth and tenth difference of monthly wool prices, 1890–1937. In each case, the continuous bell-shaped curve represents the Gaussian "interpolate" based upon the sample variance. Source: Gerhard Tintner, *The Variate-Difference Method* (Bloomington, Ind., 1940).

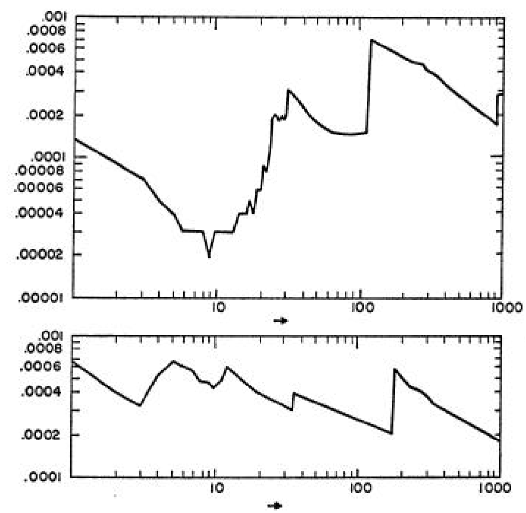


FIG. 2.—Both graphs are relative to the sequential sample second moment of cotton price changes. Horizontal scale represents time in days, with two different origins T^o: on the upper graph, T^o was September 21, 1900; on the lower graph T^o was August 1, 1900. Vertical lines represent the value of the function

$$(T - T^o)^{-1} \sum_{t=T^o}^{t=T} [L(t, 1)]^2 ,$$

where $L(t, 1) = \log_e Z(t + 1) - \log_e Z(t)$ and $Z(t)$ is the closing spot price of cotton on day t, as privately reported by the United States Department of Agriculture.

http://classes.yale.edu/fractals/Panorama/ManuFractals/Internet/Internet4.html

Syndrom of infinite variance

- Mandelbrot’s paper on long tail densities

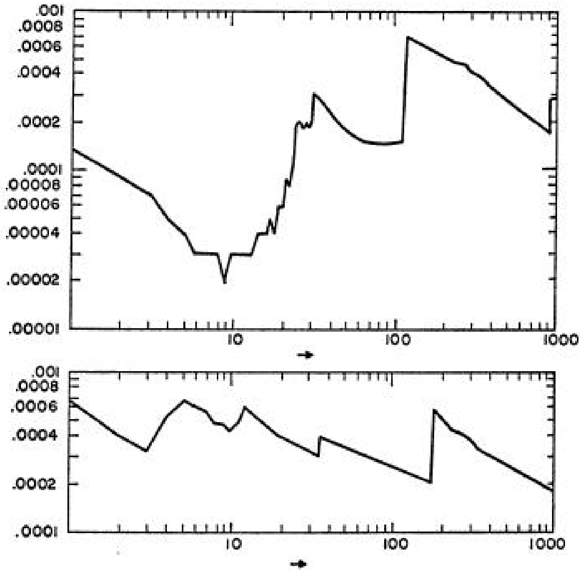


FIG. 2.—Both graphs are relative to the sequential sample second moment of cotton price changes. Horizontal scale represents time in days, with two different origins T°: on the upper graph, T° was September 21, 1900; on the lower graph T° was August 1, 1900. Vertical lines represent the value of the function

$$(T - T^\circ)^{-1} \sum_{t=T^\circ}^{t=T} [L(t,1)]^2 ,$$

where $L(t, 1) = \log_e Z(t+1) - \log_e Z(t)$ and $Z(t)$ is the closing spot price of cotton on day t, as privately reported by the United States Department of

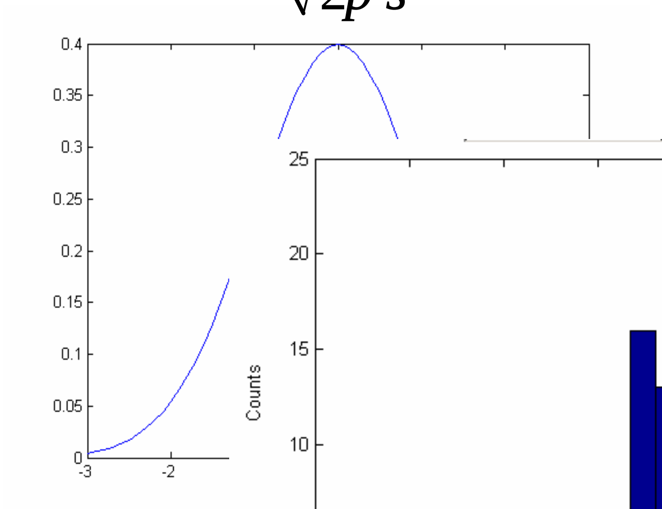
$$m(X) = E(X) = \sum_{x \in X} x P(X = x)$$

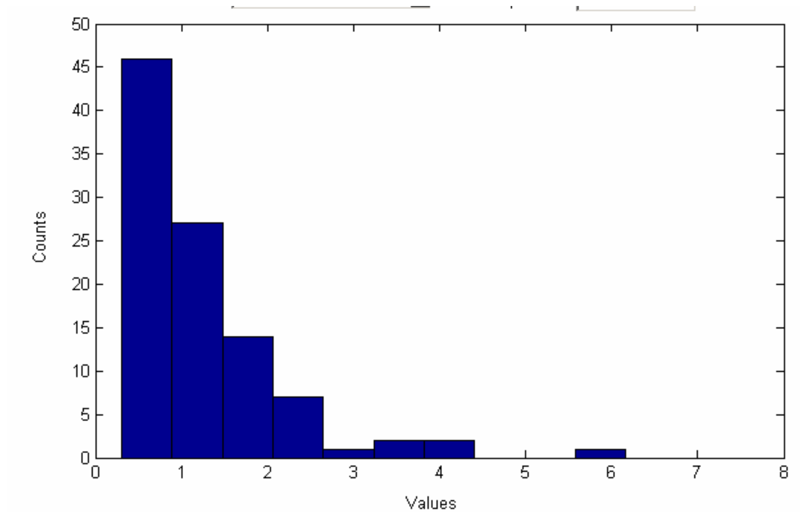
$$s^2(X) = E\left((X - m)^2\right) = \sum_{x \in X} (x - m)^2 P(X = x)$$

$$P(X = x) = \frac{1}{x^a}$$

Note that for alfa<2 the sum diverges

$$p(x)=\frac{1}{\sqrt{2p}\,s}e^{-\frac{(x-m)^2}{s^2}}$$





$\frac{1}{2p\,s})$