The concept of Dispersion

The concept
• Law of large numbers tells us that the average result will \( \rightarrow E(X) \)
• What can dispersion can we expect around \( E(X) \)
• 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.

The measures
• Standard deviation
\[
\sigma^2(X) = E[(X - \mu)^2] = \sum (x - \mu)^2 P(X = x)
\]
• Mean absolute distance
\[
\mu(X) = E(X) = \sum x P(X = x)
\]
\[
MAX(D) = E[|X - \mu|] = \sum |x - \mu| P(X = x)
\]
The measures

- Standard deviation
- Meaning of the curvature of a plane curve:
  - Radius of a circle that approximates locally the curve.
  \[
  \kappa = \frac{1}{\left(1 + \frac{2}{3} \frac{\text{d} \alpha}{\text{d}x}\right)^{3/2}} \approx \frac{1}{\text{d} \alpha / \text{d}x}
  \]
  \[
  R = \frac{1}{\kappa}
  \]

- Certainty: low dispersion -> small radius -> small variance
- Uncertainty: high dispersion -> big radius -> big variance

\[
\frac{1}{\text{d} \alpha / \text{d}x} = \sigma^2
\]

Cramer-Rao and statistics

Information about Gaussian distributions.

- % of the cases for a given dispersion

- Uses: measure of dispersion for distributions longer tail than Gaussians.

\[
\text{MAID} X = E|X - \mu| = \sum_x |X - \mu| P(X = x)
\]

Laplace vs. Gaussian

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.

\[ \text{MAE}(X) = E[|X - \mu|] = \sum_{x \in X} |x - \mu| P(X = x) \]

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

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
  Gaussian vs. geometric

Description of a random variable
• Gaussian
  \[ p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \]

• Geometric
  \[ P(X = i) = p^i(1-p) \quad \text{for } i = 1, 2, 3, \ldots \]

• Negative binomial

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 assessing 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/~savage/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.

\[ A = 200,000 \]
\[ r = 14\% \]
\[ \sum_{i=1}^{20} A(1+r)^i = 0 \]

The flaw of averages

• Simulations on real data:
  Start: 1974 Avg. Return 15.4%, Goes the distance.
  Start: 1975 Avg. Return 15.4%, Tanks in 13 yrs.

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 subtracted independently of \( r \)

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

Distribution of the invested capital
Tchebychev Inequality

- A bound on the upper probability.

\[
P(|X - E(X)| > c) \leq \frac{\sigma^2}{c^2}
\]

Tchebychev Inequality

- Geometrical meaning of

\[
P(|X - E(X)| > c) \leq \frac{\sigma^2}{c^2}
\]

Tchebychev Inequality

- Geometrical meaning of

\[
P(|X - E(X)| > c) \leq \frac{\sigma^2}{c^2}
\]

Tchebychev Inequality

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

\[
P(|X - E(X)| > k\sigma) \leq \frac{1}{k^2}
\]

- for any \( k > 0 \) or equivalently

\[
P(|X - E(X)| > c) \leq \frac{\sigma^2}{c^2}
\]

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.

Tchebychev Inequality

- Proof
  - Given an ordered set \( \{x_1, x_2, x_3, x_4, \ldots\} \)
  - We define the subset
  - then

\[
\sigma^2 = \sum_{i=1}^{\infty} (x_i - E(X))^2 P_i \geq \sum_{i=1}^{\infty} (x_i - E(X))^2 P_i \geq c^2 \sum_{i=1}^{\infty} P_i
\]

\[
\sigma^2 \geq c^2 P(|X - E(X)| > c)
\]
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

Syndrom of infinite variance

• Mandelbrot’s paper on long tail densities

\[ \mu(X) = E[X] = \sum_{X} x \cdot P(X = x) \]
\[ \sigma^2(X) = \text{Var}(X) = \sum_{X} (x - \mu)^2 \cdot P(X = x) \]
\[ P(X = s) = \frac{1}{s^\alpha} \]

Note that for \( \alpha < 2 \) the sum diverges.