SYSM 6303: Quantitative Introduction to Risk and Uncertainty in Business
Lecture 4: Fitting Data to Distributions

M. Vidyasagar

Cecil & Ida Green Chair
The University of Texas at Dallas
Email: M.Vidyasagar@utdallas.edu

September 27, 2014

Outline

Overview

Given data, we wish to fit it with an appropriate probability distribution.

Issues:

- Which distribution should we choose?
- How do we estimate the parameters in the distribution?
- How can we quantify how well the distribution fits the data?
- How can we use distributions to test some simple hypotheses?

Remember: You can fit *any distribution to any data*. But that doesn't mean that the fitted distribution is a *good fit!*

Outline

Outline

Sample Mean

Given independent samples $x_1, \ldots, x_n$ of a r.v. X, the quantity $\hat{\mu}$ defined by

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

is called the **sample mean**. It is just the average of the observations.

Note: The sample mean is itself a random quantity, because it depends on the random samples x_i – if we repeat the experiment we will in general get a different value.

Sample Variance

Given independent samples $x_1, \ldots, x_n$ of a r.v. X, the quantity $\hat{V}$ defined by

$$\hat{V} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2.$$

is called the **sample variance**.

Like the sample mean, the sample variance is also a random quantity.

The question of interest is: How well do $\hat{\mu}$ and $\hat{V}$ represent the *actual* mean and variance?

Unbiased Estimator of the Mean

An estimator is said to be **unbiased** if its expected value is equal to its true value.

Fact: The sample mean $\hat{\mu}$ is an unbiased estimate of the true mean.

Observe that

$$E(\hat{\mu}) = \frac{1}{n} \sum_{i=1}^{n} E(x_i) = E(X).$$

Unbiased Estimator of the Variance

Fact: The sample variance $\hat{V}$ is *not* an unbiased estimator.

We can compute the expected value of $\hat{V}$ as

$$E\left[\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_i\right)^2\right] = \frac{1}{n^2}\left[\sum_{i=1}^{n} E\left(nx_i - \sum_{j=1}^{n} x_j\right)^2\right].$$

This simplifies to

$$E(\hat{V}) = \frac{1}{n^2}\sum_{i=1}^{n}\left[(n-1)x_i - \sum_{j\neq i} x_j\right]^2 = \frac{n-1}{n}V(X),$$

where $V(X)$ is the true variance of X.

Unbiased Estimator of the Variance (Cont'd)

Therefore the sample variance is a *biased estimate* and is too low by a factor of $(n-1)/n$.

The unbiased estimate of the variance is

$$\frac{n}{n-1}\hat{V} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

Obviously, when the number of samples n is very large, $\hat{V}$ is very close to the unbiased estimate.

Unbiased Estimator of the Standard Deviation

Similarly, the quantity

$$\left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \right]^{1/2}$$

is an unbiased estimate of the standard deviation $\sigma(X)$.

The Matlab commands `mean`, `var`, and `std` compute the unbiased estimates defined above.

Consistent Estimators

An estimate is said to be **consistent** if it converges to the true value as the number of samples $n \to \infty$. It is obvious that every unbiased estimator is consistent, but not every consistent estimator has to be unbiased.

Probably the best-known consistent, but biased, estimator is the sample variance $\hat{V}$. We have already seen that

$$E(\hat{V}) = \frac{n-1}{n} V(X).$$

So as $n \to \infty$, the sample variance $\hat{V}$ converges to the right value, though it is a biased estimate.

Outline

Parametric Representation of Probability Distributions

As the name implies, a 'maximum likelihood estimator' is one that best explains the observed data *from within a chosen class of models*.

To make the idea more precise, suppose we have a class of probability distributions, call them $\phi(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector. The parameter vector is supposed to capture *every feature* of the probability distribution.

Some Examples on Finite Sets

Example 1: If there is a two-sided coin, then its probability distribution can be captured with just one parameter: the probability of 'Heads'. We need not specify the probability of 'Tails' since it is just one minus the probability of 'Heads'.

Example 2: For a six-sided die, its probability distribution can be captured by specifying the probabilities of any five of the six outcomes. (The sixth one need not be specified.)

Some Examples on Countable Sets

Example 3: Recall that the geometric distribution on the nonnegative integers $\mathbb{N} \cup \{0\} = \{0, 1, \ldots\}$ is given by

$$\Pr\{X = n\} = (1 - p)p^n, n = 0, 1, \ldots$$

So it is completely described by the single parameter p.

Example 4: Recall that the Poisson distribution on the nonnegative integers $\mathbb{N} \cup \{0\} = \{0, 1, \ldots\}$ is given by

$$\Pr\{X = n\} = e^{-\lambda} \frac{\lambda^n}{n!},$$

where $\lambda > 0$ is called the 'rate' of the process. So it is completely described by the single parameter λ.

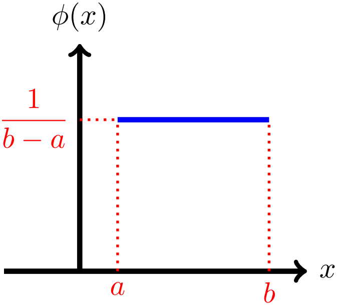
Some Examples on Continuous R.V.s

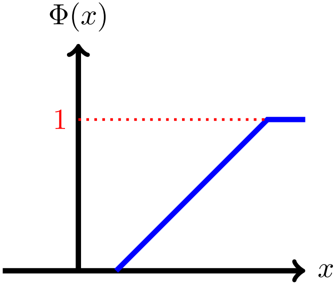
Example 5: Recall that the uniform distribution on $\mathbb{R}$ is described by the density

$$\phi(x) = \begin{cases} 1/(b-a) & \text{if } x \in [a,b], \\ 0 & \text{if } x \notin [a,b]. \end{cases}$$

The depiction of the uniform density and distribution are shown on the next slide. This distribution is completely characterized by two parameters, for example: (i) the two numbers a and b, or (ii) the starting point b and the height $1/(b-a)$, or other equivalent representations.

Uniform Distribution: Density and CDF





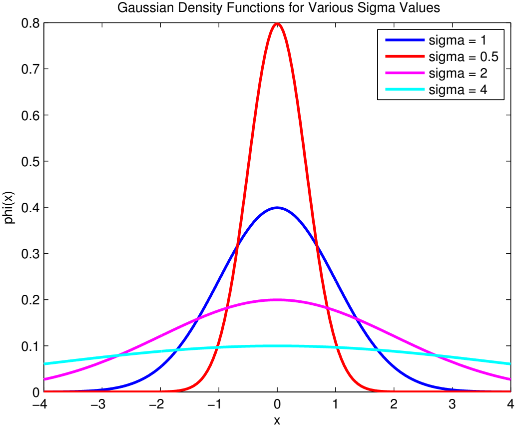
Some Examples on Continuous R.V.s (Cont'd)

Example 6: Recall that the Gaussian density function is defined by

$$\phi(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp[-(x-\mu)^2/2\sigma^2].$$

So it is completely characterized by the two parameters (μ, σ).

Gaussian Density Function



The Likelihood and Log-Likelihood Functions

Given independent samples $x_1, \ldots, x_n$, we can compute the *likelihood* of this particular sequence being observed as

$$L(x_1^n; \boldsymbol{\theta}) = \prod_{i=1}^{n} \phi(x_i; \boldsymbol{\theta}),$$

where we use x_1^n as a shorthand for $x_1, \ldots, x_n$.

The **log-likelihood function** is the logarithm of $L(x_1^n)$. Clearly, it is given by

$$LL(x_1^n; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \phi(x_i; \boldsymbol{\theta}).$$

Likelihood Functions for Continuous R.V.s

Suppose X is a continuous r.v. with density $\phi(x;\boldsymbol{\theta})$, and we have observed i.i.d. samples $x_1, \ldots, x_n$. For a continuous r.v., the likelihood of observing *precisely* a specified value is zero. So the likelihood and log-likelihood functions in this case are defined as

$$L(x_1^n;\boldsymbol{\theta}) = \prod_{i=1}^{n} \phi(x_i;\boldsymbol{\theta}),$$

$$LL(x_1^n;\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \phi(x_i;\boldsymbol{\theta}).$$

Maximum Likelihood Estimators

A **maximum likelihood estimator (MLE)** is one that chooses the parameter vector θ so as to maximize the likelihood of observing that particular sample.

Since $\log$ is a monotonic function, maximizing the likelihood function is equivalent to maximizing the log-likelihood function.

In symbols,

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}} LL(x_1^n; \boldsymbol{\theta}),$$

where the log-likelihood function is defined above, for discrete-valued or continuous r.v.s as appropriate.

Simple Example of Maximum Likelihood Estimation

Suppose we wish to estimate the probability of "Heads", call it p, from n coin tosses out of which k turn out to be heads. Here p is the variable of optimization. The likelihood of getting k Heads out of n coin tosses, as a function of p, is

$$L(p) = \binom{n}{k} p^k (1-p)^{n-k},$$

$$LL(p) = \log \binom{n}{k} + k \log p + (n-k) \log(1-p),$$

$$\frac{dLL(p)}{p} = \frac{k}{p} - \frac{n-k}{1-p} = 0 \text{ if } p = \frac{k}{n}.$$

So the MLE for p is k/n.

Properties of MLE

We briefly mention a few attractive properties of MLEs.

- MLEs are *consistent* - as $m \to \infty$, the estimated parameter vector approaches the 'true' value, if the data is generated by a 'true' model.
- The estimated parameter vector $\boldsymbol{\theta}^*$ is itself a random vector, as it depends on the random samples. It can be shown that, as $l \to \infty$, the distribution of $\boldsymbol{\theta}^*$ around the 'true' vector is asymptotically the multivariate normal distribution.
- The MLE is also *efficient* – no other consistent estimator has lower asymptotic mean squared error than the MLE.

Outline

Outline

MLE for Finite Outcomes

Suppose a random variable X has m possible outcomes (example: a six-sided die). Then the parametrized family of distributions is

$$\phi(x; \boldsymbol{\theta}) = (\theta_1, \ldots, \theta_m),$$

with the proviso that each θ_i is nonnegative and that all of them must add up to one.

Suppose $k_1, \ldots, k_n$ are the number of times that the various outcomes are observed. Then the MLE estimator for the probability of each outcome is

$$\theta_i^* = \frac{k_i}{n},$$

the fraction of times that particular outcome is actually observed.

MLE for Finite Outcomes (Cont'd)

Example: If 100 tosses of a coin result in 63 heads and 37 tails, then the maximum likelihood estimates are

$$\hat{P}(H) = 0.63, \hat{P}(T) = 0.37.$$

Example: If 1000 rolls of a six-sided die result in the following results: $k_1 = 177$, $k_2 = 165$, $k_3 = 155$, $k_4 = 161$, $k_5 = 170$, $k_6 = 172$. Then the MLE of the probability distribution is the vector

$$\theta^* = [\ 0.177 \quad 0.165 \quad 0.155 \quad 0.161 \quad 0.170 \quad 0.172\ ].$$

MLE for Poisson Distribution

Suppose that we have observations of an integer-valued random variable (for example a histogram), and we wish to fit a Poisson distribution to it.

Recall that the Poisson distribution on the nonnegative integers $\mathbb{N} \cup \{0\} = \{0, 1, \ldots\}$ is given by

$$\Pr\{X = n\} = e^{-\lambda} \frac{\lambda^n}{n!},$$

where $\lambda > 0$ is called the 'rate' of the process. Therefore the distribution is completely described by the single parameter λ.

MLE for Poisson Distribution – 2

Suppose there are a total of m observations, and that the value n occurs exactly m_n times. Then the MLE for the rate λ is

$$\lambda^* = \frac{1}{m} \sum_{n=0}^{N} m_n n = \sum_{n=0}^{N} n \phi_n,$$

where $\phi_n = m_n / m$ is the fraction of observations that equal m_n.

MLE for Poisson Distribution – Example

Suppose the time to fabricate a part has been observed for 100 samples. All 100 samples are completed within 20 days. The observed number of days (from 1 to 20) for each of the 100 parts is shown below:

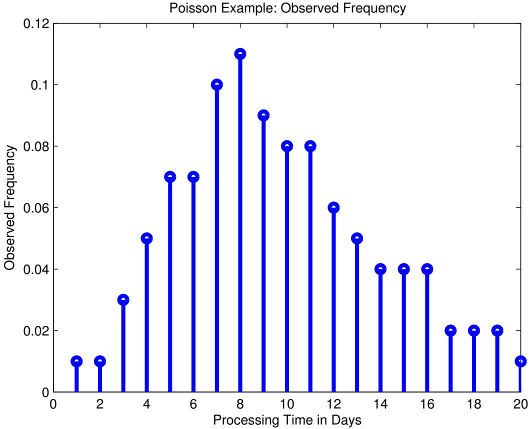
$$D = [\ 1 \quad 1 \quad 3 \quad 5 \quad 7 \quad 7 \quad 10 \quad 11 \quad 9 \quad 8 \quad 8 \quad 6 \quad 5 \quad 4 \quad 4 \quad 4 \quad 2 \quad 2 \quad 2 \quad 1\ ].$$

The next slide shows the plot of the observed frequencies, which are the above entries divided by 100.

We will fit a Poisson distribution to the data.

Poisson Distribution Example (Cont'd)

The figure below shows the observed frequencies.



Poisson Distribution Example (Cont'd)

Recall the formula for the maximum likelihood estimate of the rate of a Poisson process given the observe frequencies:

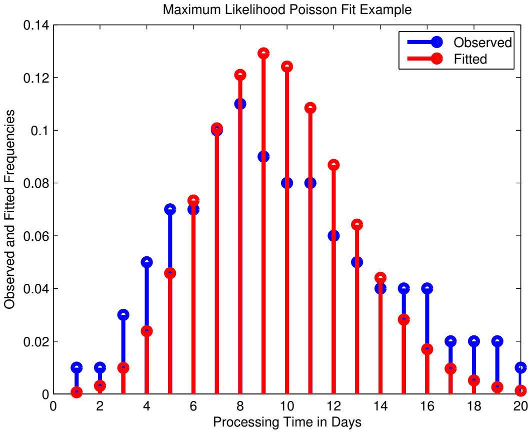
$$\lambda^* = \frac{1}{m} \sum_{n=0}^{N} m_n n = \sum_{n=0}^{N} n \phi_n ,$$

We apply this formula with $m = 20$s, which leads to $\lambda^* = 9.6100$.

The next slide shows the observed frequency and the fitted Poisson distribution with the rate λ^*.

Poisson Distribution Example (Cont'd)

The figure below shows the observed and fitted frequencies.



Poisson Distribution Example (Cont'd)

It can be seen that the fit is not particularly good.

What this means is that the Poisson distribution with rate $\lambda^* = 9.6100$ is the *best possible fit within the class of Poisson distributions* – that is all!

We still need other ways to determine whether or not this is a good fit. This is provided by "goodness of fit tests" to be discussed later.

Outline

MLE for Gaussian Distributions

Recall that

$$\phi(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp[-(x-\mu)^2/2\sigma^2].$$

This distribution is completed characterized by the parameter pair (μ, σ).

MLE for Gaussian Distributions – 2

Now suppose we have observations $x_1, \ldots, x_n$. Then the MLE turn out to be the sample mean and sample variance; that is

$$\mu^* = \frac{1}{n} \sum_{i=1}^{n} x_i = \hat{\mu},$$

$$\sigma^* = \left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu^*)^2 \right]^{1/2} = \sqrt{\hat{V}}.$$

As we have seen before, the variance estimator is *biased* but it is the most likely estimate.

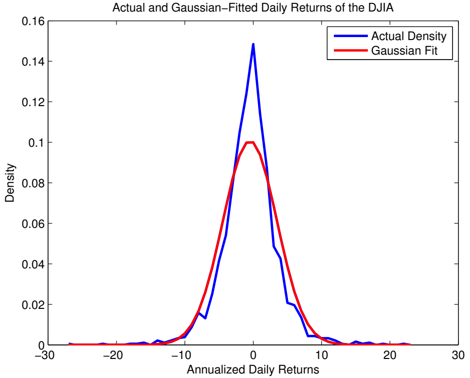
Modeling Asset Returns

Think of a stock price as a sequence of random variables $\{X_t\}$. The ratio $\log(X_{t+1}/X_t)$ is called the "return" at time t (or $t+1$).

Traditional methods of option pricing assume that the returns are Gaussian (or that asset prices are "log-normal").

Recent studies cast doubts on this theory. Some people suggest using stable distributions instead. (More on this later.)

Gaussian Fit to Returns on Dow-Jones Industrial Average



Multivariate Gaussian Distributions

Suppose $\mathbf{X}$ is a random variable that assumes values in $\mathbb{R}^k$. So $\mathbf{X} = (X_1, \ldots, X_k)$ where each X_i is a real-valued r.v. The **multivariate Gaussian Density function** has the form

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{k/2} \det(\Sigma)^{1/2}} \exp[(\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})/2],$$

where $\boldsymbol{\mu} \in \mathbb{R}^k$ is the vector of means, and $\Sigma \in \mathbb{R}^{k \times k}$ is the covariance matrix.

If $k = 1$, $\boldsymbol{\mu}$ is a scalar μ, and Σ is a scalar σ, we get back the (univariate) Gaussian density.

MLE for Multivariate Gaussian Distributions

Suppose have independent samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of the k-dimensional r.v. $\mathbf{X}$. Then the MLE are given by

$$\boldsymbol{\mu}^* = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i, \Sigma^* = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^t (\mathbf{x}_i - \boldsymbol{\mu}^*).$$

Note that if $n < k$, the matrix Σ is singular. Note that unless the number of observations n exceeds the dimension of the random variable k, the matrix Σ will be singular and cannot be inverted.

Example: Daily Returns on Five Stocks

Daily prices are obtained on five stocks: Apple, Merck, Nike, Yahoo and Google. These daily *prices* are then converted into annualized daily *returns* by taking the log of the ratio of successive daily prices and multiplying by 365.

The mean of each return and the covariance of the returns are computed using the Matlab commands `mean` and `cov`.

Again, because the number of samples is large $n = 1341$, the fact that Matlab computes the unbiased estimate of the covariance, whereas the maximum likelihood estimate is the (biased) sample covariance can be ignored.

Example: Daily Returns on Five Stocks (Cont'd)

The 1×5 vector of mean annualized returns is

$$\boldsymbol{\mu}^* = [\ 0.3559 \quad -0.0547 \quad -0.0239 \quad -0.0508 \quad 0.0818\ ],$$

while the 5×5 covariance matrix of annualized returns is

$$\Sigma^* = \begin{bmatrix} 73.3480 & 22.5712 & 31.7058 & 35.0791 & 34.4119 \\ 22.5712 & 49.9162 & 22.4836 & 20.8499 & 20.6514 \\ 31.7058 & 22.4836 & 109.6736 & 30.5985 & 25.7874 \\ 35.0791 & 20.8499 & 30.5985 & 108.1180 & 29.5364 \\ 34.4119 & 20.6514 & 25.7874 & 29.5364 & 145.9925 \end{bmatrix}.$$

Example: Daily Returns on Five Stocks (Cont'd)

The maximum likelihood estimate *among the class of multivariate Gaussian distributions* is given by

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{k/2} \det(\Sigma^*)^{1/2}} \exp[(\mathbf{x} - \boldsymbol{\mu}^*)^t \Sigma^{*-1} (\mathbf{x} - \boldsymbol{\mu}^*)/2],$$

where $\boldsymbol{\mu}^*, \Sigma^*$ are shown on the previous slide.

Outline

Outline

Justification for Using the Gaussian Distribution

The main theoretical justification for using the Gaussian distribution to model r.v.s comes from the central limit theorem. First we state the law of large numbers.

Law of Large Numbers

Note: This is *not* the most general version of the law of large numbers!

Suppose $X_1, \ldots, X_l$ are independent real-valued r.v.s with finite mean μ and finite variance V (or standard deviation $\sigma = \sqrt{V}$); they could even be discrete-valued random variables. Let A_l denotes their average, that is

$$A_l = \frac{1}{l} \sum_{i=1}^{l} X_i.$$

Then the sample average A_n converges "in probability" to the true mean μ.

Law of Large Numbers (Cont'd)

This means that, for every fixed small number ϵ, the tail probability

$$q(n, \epsilon) = \Pr\{|A_n - \mu| > \epsilon\} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

In words, the density function of the average A_n gets concentrated around the true value μ.

Central Limit Theorem (Cont'd)

Things are different if we "center" and "normalize" the average by defining

$$G_l = \frac{A_l - \mu}{\sqrt{l}\sigma}.$$

Theorem: As $l \to \infty$, the distribution function of G_l converges to that of the normal random variable, that is, a Gaussian r.v. with zero mean and standard deviation of one.

Note that the theorem is true even if X is a *discrete-valued* r.v., such as the payoff associated with a coin toss.

Illustrative Example

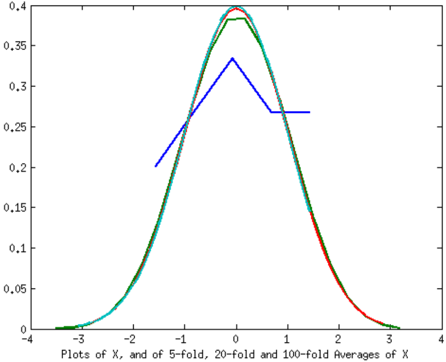
Suppose X assumes values in the discrete set $\{-2, -1, 0, 1, 2\}$ with distribution vector

$$\phi = [\ 0.15 \quad 0.20 \quad 0.25 \quad 0.20 \quad 0.20\ ].$$

Then, for each integer l, the l-fold average A_l also assumes values in the interval $[-2, 2]$. While A_l is also discrete-valued, the number of possible values increases as l increases.

The next slides show the densities of 5-fold, 20-fold, and 100-fold averages of independent copies of X.

Depictions of Densities



Blue curve is the original density, green is the 5-fold average, red is the 20-fold average, and taupe is the 100-fold average.

Outline

Stable Distributions: Motivation

We have seen that the normal distribution is the limit distribution of the "centered" and "scaled" averages of independent samples of random variables with finite variance.

What happens if we average r.v.s that don't necessarily have finite variance? What can the limit distributions look like?

Answer: The *only possible limits* are the stable distributions!

The Gaussians are the only stable distributions with finite variance; the rest all have infinite variance.

Stable Distributions: Motivation

Stable distributions provide a better fit to real-world data compared to the Gaussian, because the Gaussian is a special case of a stable distribution.

This is shown by various examples.

The theory is very advanced, and only a few necessary details are given here.

This section can be skipped at first reading and revisited.

Characteristic Function of a R.V.

If X is a real-valued r.v., then its **characteristic function** ψ_X is defined by

$$\psi_X(u) = E[\exp(\mathbf{i}uX)] = \int_{-\infty}^{\infty} e^{\mathbf{i}ux}\phi(x)dx,$$

where $\phi(x)$ is the density of the r.v. X, and $\mathbf{i} = \sqrt{-1}$.

In other words, the c.f. is the *Fourier transform* of the density.

If X is Gaussian with mean μ and variance σ^2, then its c.f. is also Gaussian.

$$\psi_X(u) = \exp[\mathbf{i}u\mu - u^2\sigma^2/2].$$

Parameters of a Stable Distribution

A Gaussian r.v. is completely specified by just two parameters, namely its mean μ and its standard deviation σ. A Gaussian is also "stable" with "exponent" $\alpha = 2$ (and we will see why in later lectures) but the discussion below is for non-Gaussian stable r.v.s.

Every non-Gaussian stable r.v. X is completely specified by four parameters:

- An **exponent** $\alpha \in (0, 2)$.
- A **skew** $\beta \in [-1, 1]$.
- A **scale** $\gamma \in \mathbb{R}_+$.
- A **location** $\delta \in \mathbb{R}$.

Characteristic Function of a Stable Random Variable

The c.f. of a stable r.v. has two distinct forms, depending on whether the exponent α equals 1 or not.

$$\psi_X(u) = \exp(\mathbf{i}\delta u - \gamma^\alpha |u|^\alpha [1 - \mathbf{i}\beta \tan\left(\frac{\pi\alpha}{2}\right) \frac{|u|}{u}]), \text{ if } \alpha \neq 1,$$

$$\psi_X(u) = \exp(\mathbf{i}\delta u - \gamma u[1 + \mathbf{i}\beta \frac{2}{\pi} \frac{|u|}{u} \log|u|]) \text{ if } \alpha = 1.$$

Unfortunately *there is no closed form expression* for the density ϕ – only for the characteristic function.

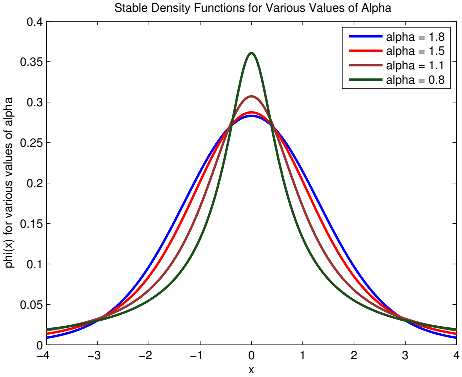
Interpretation of Parameters

The four parameters mean what the names suggest:

- The exponent α controls how slowly the complementary distribution function $\bar{\Phi}(u)$ decays as $u \to \infty$. As α gets smaller, the densities get flatter and wider.
- The skew β is zero if the density function is symmetric, and nonzero otherwise.
- The scale γ is the spread on the u-axis. As γ is decreased, the density function gets spread out.
- The location δ centers the location of the distribution.

The next several slides illustrate the role of these constants.

Varying Alpha



As α is decreased, the peaks get higher and the tails get flatter.

Heavy-Tailed Behavior of Stable Distributions

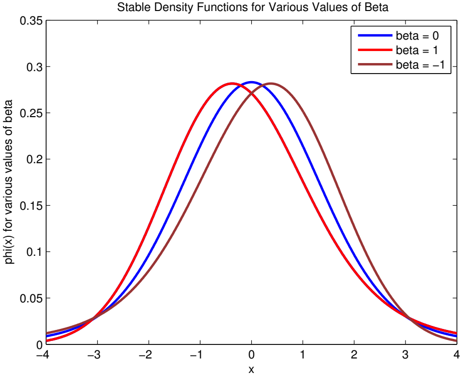
The Gaussian distribution can be thought of as a special case of a stable distribution with $\alpha = 2$.

If $\alpha < 2$, then the r.v. is "heavy-tailed" in that its variance is infinite.

If $\alpha < 1$, then even the mean is infinite.

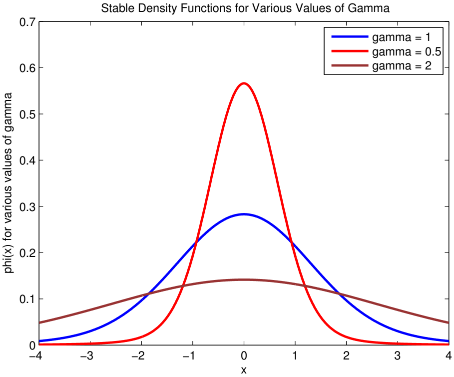
Despite this, stable distributions with $\alpha < 2$ often provide a far better fit to real-world data than Gaussian distributions.

Varying Beta



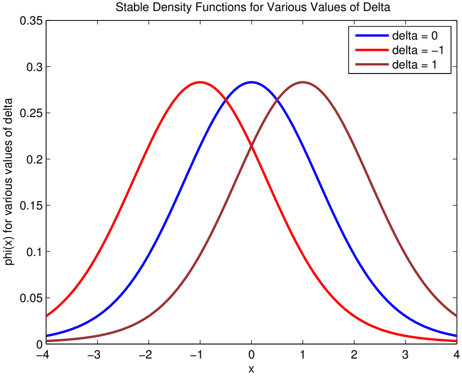
The brown and red curves are asymmetric, though this is hard to see.

Varying Gamma



Smaller values of γ spread out the density function.

Varying Delta



Nonzero values of δ shift the curve to the left or the right but do not otherwise change the shape.

Outline

Fitting Stable Distributions to Data

The utility stblfit.m can be used to fit a stable fit to any data set. It returns a four-dimensional vector consisting of $[\ \alpha \quad \beta \quad \gamma \quad \delta\ ]$.

The next several slides illustrate the application of this utility.

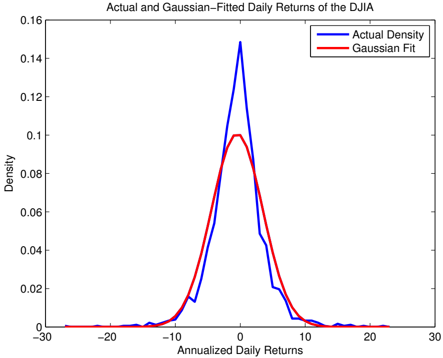
Example: Dow-Jones Industrial Average

Daily closing values of the Dow-Jones Industrial Average (DJIA) are taken for roughly seven years.

The logarithm of the ratio of successive closing averages, multiplied by 365, gives the **annualized daily returns** of the DJIA.

The Gaussian gives a very poor fit to the data, whereas a stable distribution with $\alpha = 1.6819$ gives an excellent fit, as shown in the next slides.

DJIA Daily Returns: Gaussian Fit



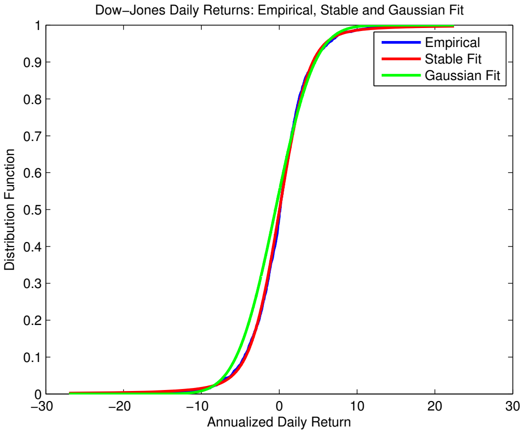
The fit of the density function (histogram) and the Gaussian density with the same mean and standard deviation.

DJIA Daily Returns: Stable Fit

The next slide shows the fit of the cumulative distribution function – observed, stable and Gaussian fitted.

The stable distribution provides an excellent fit, whereas there is a large gap between the Gaussian fit and the observed.

DJIA Daily Returns: Stable Fit (Cont'd)



Outline

Empirical Distributions

Suppose X is a random variable for which we have generated n i.i.d. samples, call them $x_1, \ldots, x_n$.

Then we define the **empirical distribution** of X, based on these observations, as follows:

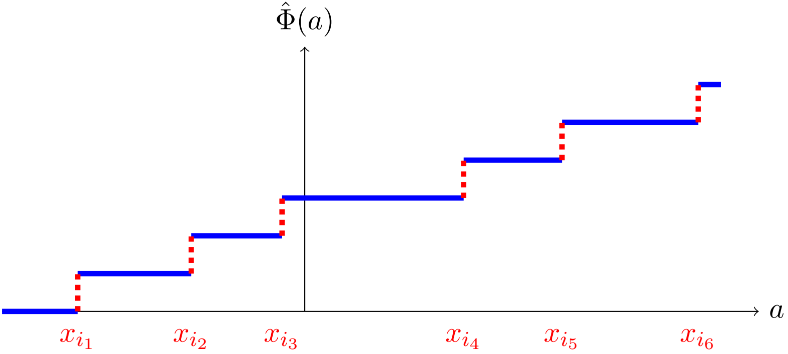
$$\hat{\Phi}(a) = \frac{1}{n} \sum_{i=1}^{n} I_{\{x_i \leq a\}},$$

where I denotes the **indicator function**: $I = 1$ if the condition below is satisfied and $I = 0$ otherwise.

So in this case $\hat{\Phi}(a)$ is just the fraction of the n samples that are $\leq a$. The diagram on the next slide illustrates this.

Empirical Distribution Depicted

Arrange samples $x_1, \ldots, x_n$ in increasing order of magnitude; call them $x_{i_1}, \ldots, x_{i_n}$.



Glivenko-Cantelli Lemma

Theorem: As $n \to \infty$, the empirical distribution $\hat{\Phi}(\cdot)$ approaches the true distribution $\Phi(\cdot)$.

Specifically, if we define the **Kolmogorov-Smirnov distance**

$$d_n = \max_u |\hat{\Phi}(u) - \Phi(u)|,$$

then $d_n \to 0$ as $n \to \infty$.

At *what rate* does the convergence take place?

One-Sample Kolmogorov-Smirnov Statistic

Fix a ‘confidence level’ $\delta > 0$ (usually δ is taken as 0.05 or 0.02). Define the threshold

$$\theta(n, \delta) = \left(\frac{1}{2n} \log \frac{2}{\delta} \right)^{1/2}.$$

Then with probability $1 - \delta$, we can say that

$$\max_u |\hat{\Phi}(u) - \Phi(u)| =: d_n \leq \theta_n.$$

One-Sample Kolmogorov-Smirnov Test

Given samples $x_1, \ldots, x_n$, fit it with some distribution $F(\cdot)$ (e.g. Gaussian). Compute the K-S statistic

$$d_n = \max_u |\hat{\Phi}(u) - F(u)|.$$

Compare d_n with the threshold $\theta(n, \delta)$. If $d_n > \theta(n, \delta)$, we 'reject the null hypothesis' at level δ. In other words, if $d_n > \theta(n, \delta)$, then we are $1 - \delta$ sure that the data was *not* generated by the distribution $F(\cdot)$.

If $d_n \leq \theta(n, \delta)$ then we cannot reject the hypothesis. Usually in this case we accept the hypothesis.

Example: Daily Returns on the DJIA

There are $n = 1832$ daily returns. By using the utility `stblfit.m`, we obtain a stable distribution with the parameters

$$\begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} = \begin{bmatrix} 1.6819 \\ -0.0651 \\ 2.2345 \\ -0.0150 \end{bmatrix}$$

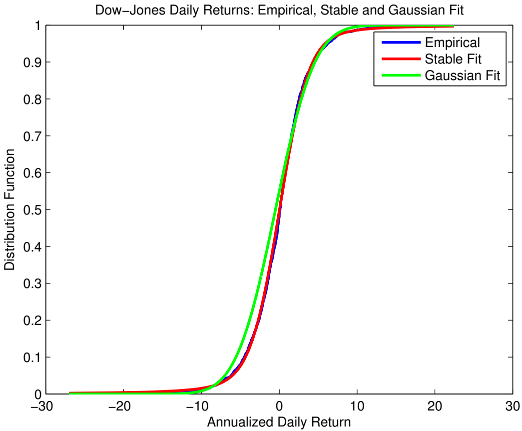
as the best stable fit.

The best Gaussian fit is obtained using the mean $\mu = -0.4798$ and the standard deviation $\sigma = 3.9629$.

The empirical, stable-fitted and Gaussian-fitted CDFs are shown in the next slide.

Question: Is either fit "acceptable"?

DJIA Daily Returns: Stable and Gaussian Fits



Analysis Using One-Sample K-S Test

If we use a confidence level of 95%, then $\delta = 0.05$. The corresponding K-S threshold is

$$\theta(n, \delta) = \left(\frac{1}{2n} \log \frac{2}{\delta} \right)^{1/2} = 0.0317.$$

The actual K-S test statistic, namely the maximum difference between the empirical and stable fitted cumulative distribution function is 0.0266, while the K-S test statistic for the Gaussian fit is 0.0995.

Because the K-S distance for the Gaussian is more than $\theta(n, \delta)$, we can assert with 95% confidence that daily returns are *not Gaussian*. For the stable fit, the K-S distance is less than $\theta(n, \delta)$, so we accept that the stable fit is acceptable.

Application of K-S Test to Discrete Random Variables

Recall the study of the time needed to process a part, measured in days. A total of 100 samples were observed, with the following distribution:

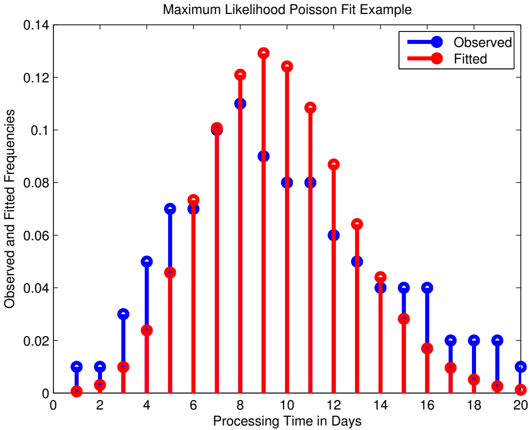
$$D = [\ 1 \quad 1 \quad 3 \quad 5 \quad 7 \quad 7 \quad 10 \quad 11 \quad 9 \quad 8 \quad 8 \quad 6 \quad 5 \quad 4 \quad 4 \quad 4 \quad 2 \quad 2 \quad 2 \quad 1\ ].$$

A Poisson distribution was fit and the maximum likelihood estimate for the rate is $\lambda^* = 9.61$.

The empirical and fitted rates are shown on the next slide. However, to test whether the fit is acceptable, we need to plot the empirical and fitted *cumulative distribution functions.* These are shown in the next two slides.

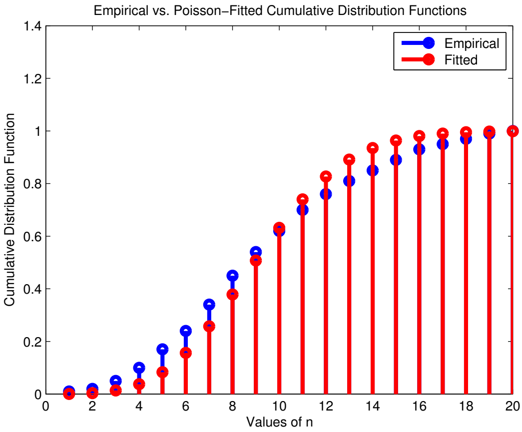
Empirical and Fitted Poisson Frequencies

The figure below shows the observed and fitted frequencies.



Empirical and Fitted CDFs

The figure below shows the observed and fitted frequencies.



Analysis Using K-S Test Statistic

Using $N = 100$ and $\delta = 0.05$, we can compute the K-S test threshold as

$$\theta(n,\delta) = \left(\frac{1}{2n} \log \frac{2}{\delta} \right)^{1/2} = 0.1358.$$

The K-S test statistic is the maximum disparity between the empirical and fitted CDFs and equals 0.0866. Because its value is less than the threshold, we cannot reject the hypothesis, and therefore accept it, perhaps reluctantly.

This is caused by having too few samples, which makes $\theta(n,\delta)$ too large.