ORF 245 Fundamentals of Statistics
Great Expectations

Robert Vanderbei

Fall 2015

Slides last edited on November 16, 2015

Definition

The *expectation* of a random variable is a measure of it's “average value”.

Discrete Case:

$$\mathbb{E}(X) = \sum_i x_i p(x_i)$$

Caveat: If it's an infinite sum and the x_i's are both positive and negative, then the sum can fail to converge. We restrict our attention to cases where the sum *converges absolutely*:

$$\sum_i |x_i| p(x_i) < \infty$$

Otherwise, we say that the expectation is *undefined*.

Continuous Case:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Corresponding Caveat: If

$$\int_{-\infty}^{\infty} |x| f(x) dx = \infty$$

we say that the expectation is *undefined*.

Geometric Random Variable

Recall that a geometric random variable takes on positive integer values, $1, 2, \ldots$, and that

$$p(k) = P(X = k) = q^{k-1}p$$

where $q = 1 - p$.

We compute:

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=1}^{\infty} kpq^{k-1} = p\sum_{k=1}^{\infty} kq^{k-1} = p\sum_{k=1}^{\infty} \frac{d}{dq}q^k \\
&= p\,\frac{d}{dq}\sum_{k=1}^{\infty} q^k = p\,\frac{d}{dq}\,q\sum_{k=0}^{\infty} q^k = p\,\frac{d}{dq}\,\frac{q}{1-q} \\
&= p\frac{(1-q)(1) - q(-1)}{(1-q)^2} = p\frac{1}{(1-q)^2} \\
&= \frac{1}{p}
\end{aligned}$$

(Isn't calculus fun!)

Poisson Random Variable

Recall that a Poisson random variable takes on nonnegative integer values, $0, 1, 2, \ldots$, and that

$$p(k) = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

where λ is some positive real number.

We compute:

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{\infty} k \, \frac{\lambda^k}{k!} \, e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} e^{\lambda} \\
&= \lambda
\end{aligned}$$

We now see that λ is the *mean*.

Exponential Random Variable

Recall that an exponential random variable is a continuous random variable with

$$f(x) = \lambda e^{-\lambda x}, \qquad x \geq 0,$$

where $\lambda > 0$ is a fixed parameter.

We compute:

$$\begin{aligned} \mathbb{E}(X) &= \int_0^\infty x \, \lambda e^{-\lambda x} \, dx \\ &= \frac{1}{\lambda} \int_0^\infty u \, e^{-u} \, du \\ &= \frac{1}{\lambda} \end{aligned}$$

(the last integral being done using *integration by parts*).

Normal Random Variable

Recall that a normal random variable is a continuous random variable with

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

We compute:

$$\begin{aligned}
\mathbb{E}(X) &= \int_{-\infty}^{\infty} x \, \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx \\
&= \int_{-\infty}^{\infty} (u+\mu) \, \frac{1}{\sqrt{2\pi}\sigma} e^{-u^2/2\sigma^2} du \\
&= \int_{-\infty}^{\infty} u \, \frac{1}{\sqrt{2\pi}\sigma} e^{-u^2/2\sigma^2} du + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-u^2/2\sigma^2} du \\
&= \qquad\qquad 0 \qquad\qquad + \mu \qquad 1 \\
&= \mu
\end{aligned}$$

The expected value of X is the mean μ.

Theorem

Let $g(\cdot)$ be some given function.

Discrete Case: $$\mathbb{E}(g(X)) = \sum_{x_j} g(x_j)p(x_j)$$

Continuous Case: $$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Derivation (Discrete case): Let $Y = g(X)$. Then

$$\mathbb{E}(g(X)) = \mathbb{E}(Y) = \sum_i y_i p_Y(y_i)$$

Let $A_i = \{x_j \mid g(x_j) = y_i\}$. Then,

$$p_Y(y_i) = \sum_{x_j \in A_i} p(x_j)$$

and so

$$\mathbb{E}(Y) = \sum_i y_i \sum_{x_j \in A_i} p(x_j) = \sum_i \sum_{x_j \in A_i} y_i p(x_j) = \sum_i \sum_{x_j \in A_i} g(x_j)p(x_j) = \sum_{x_j} g(x_j)p(x_j)$$

Note: Usually $\mathbb{E}(g(X)) \neq g(\mathbb{E}(X))$.

Theorem

Theorem: $$\mathbb{E}\left(a+\sum_{i=1}^{n} b_i X_i\right) = a + \sum_{i=1}^{n} b_i \mathbb{E}(X_i)$$

Proof: We give the proof for the continuous case with $n = 2$. Other cases are similar.

$$\begin{aligned}
\mathbb{E}(Y) &= \iint (a + b_1 x_1 + b_2 x_2) f(x_1, x_2) dx_1 dx_2 \\
&= a \iint f(x_1, x_2) dx_1 dx_2 + b_1 \iint x_1 f(x_1, x_2) dx_1 dx_2 + b_2 \iint x_2 f(x_1, x_2) dx_1 dx_2 \\
&= a + b_1 \int x_1 \left(\int f(x_1, x_2) dx_2 \right) dx_1 + b_2 \int x_2 \left(\int f(x_1, x_2) dx_1 \right) dx_2 \\
&= a + b_1 \int x_1 f_{X_1}(x_1) dx_1 + b_2 \int x_2 f_{X_2}(x_2) dx_2 \\
&= a + b_1 \mathbb{E}(X_1) + b_2 \mathbb{E}(X_2)
\end{aligned}$$

NOTE: In this class, an integral without limits is an integral from $-\infty$ to ∞.
It's *not an indefinite integral*.

Example

Consider a binomial random variable Y representing the number of successes in n independent trials where each trial has success probability p.

It's expectation is defined in terms of the probability mass function as

$$\mathbb{E}(Y) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

This sum is tricky to simplify.

Here's an easier way. Let X_i denote the Bernoulli random variable that takes the value 1 if the i-th trial is a success and 0 otherwise.

Then

$$Y = \sum_{i=1}^{n} X_i$$

and so

$$\mathbb{E}(Y) = \sum_{i=1}^{n} \mathbb{E}(X_i) = \sum_{i=1}^{n} p = np$$

Variance and Standard Deviation

Definition: The *variance* of a random variable X is defined as

$$\sigma^2 := \mathrm{Var}(X) := \mathbb{E}\left(X - \mathbb{E}(X)\right)^2$$

The *standard deviation*, denoted by σ, is simply the square root of the variance.

Theorem: If $Y = a + bX$, then $\mathrm{Var}(Y) = b^2\,\mathrm{Var}(X)$.

Proof:

$$\begin{aligned}
\mathbb{E}\left(Y - \mathbb{E}(Y)\right)^2 &= \mathbb{E}\left(a + bX - \mathbb{E}(a + bX)\right)^2 \\
&= \mathbb{E}\left(a + bX - a - b\mathbb{E}(X)\right)^2 \\
&= \mathbb{E}\left(bX - b\mathbb{E}(X)\right)^2 \\
&= b^2\,\mathbb{E}\left(X - \mathbb{E}(X)\right)^2 \\
&= b^2\,\mathrm{Var}(X)
\end{aligned}$$

Bernoulli Distribution

Recall that $q = 1 - p$ and

$$\mathbb{E}(X) = 0q + 1p = p$$

Hence

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}(X - \mathbb{E}(X))^2 \\ &= (0-p)^2 q + (1-p)^2 p \\ &= p^2 q + q^2 p \\ &= pq(p+q) \\ &= pq \end{aligned}$$

Important Note: $\mathbb{E}X^2 = \mathbb{E}(X^2) \neq (\mathbb{E}(X))^2$

Normal Distribution

Recall that

$$\mathbb{E}(X) = \mu$$

Hence

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}(X-\mu)^2 \\ &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \end{aligned}$$

Make a change of variables $z = (x-\mu)/\sigma$ to get

$$\text{Var}(X) = \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}}dz$$

This last integral evaluates to $\sqrt{2\pi}$ a fact that can be checked using integration by parts with $u = z$ and $dv =$ "everything else". Hence

$$\text{Var}(X) = \sigma^2$$

An Equivalent Alternate Formula for Variance

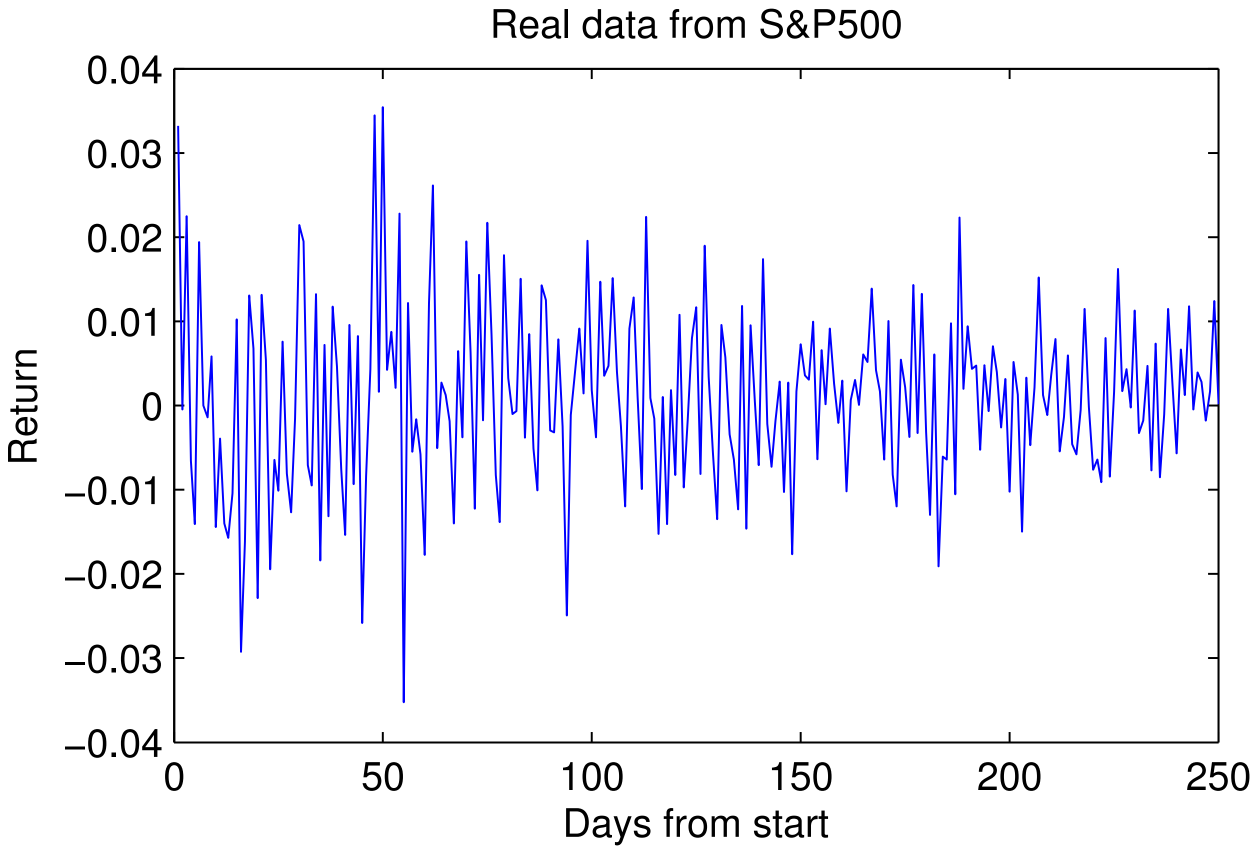
$$\mathsf{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

Let μ denote the expected value of X: $\mu = \mathbb{E}(X)$.

$$\begin{aligned}
\mathsf{Var}(X) &= \mathbb{E}(X-\mu)^2 \\
&= \mathbb{E}(X^2 - 2\mu X + \mu^2) \\
&= \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2 \\
&= \mathbb{E}(X^2) - 2\mu^2 + \mu^2 \\
&= \mathbb{E}(X^2) - \mu^2
\end{aligned}$$

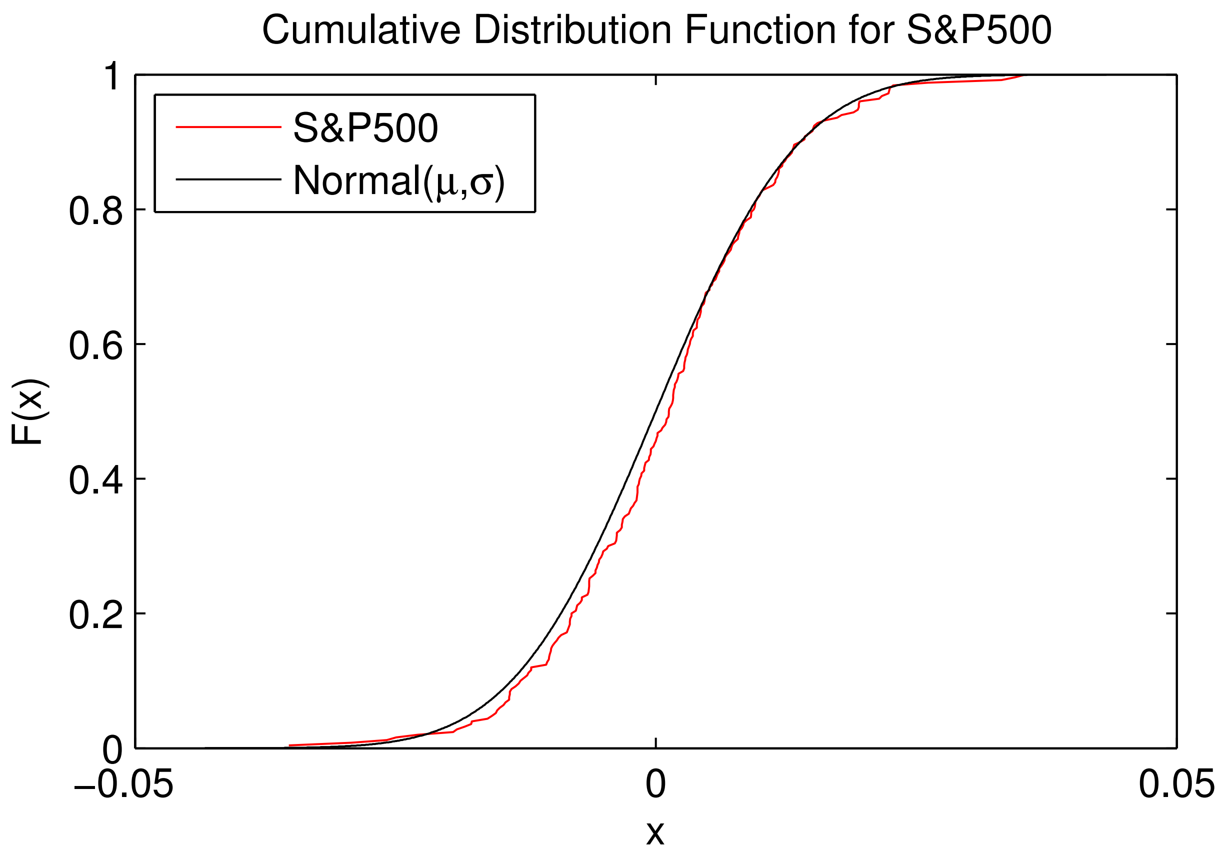
Standard and Poors 500 – Daily Returns

Raw data: R_j, $j = 1, 2, \ldots, n$

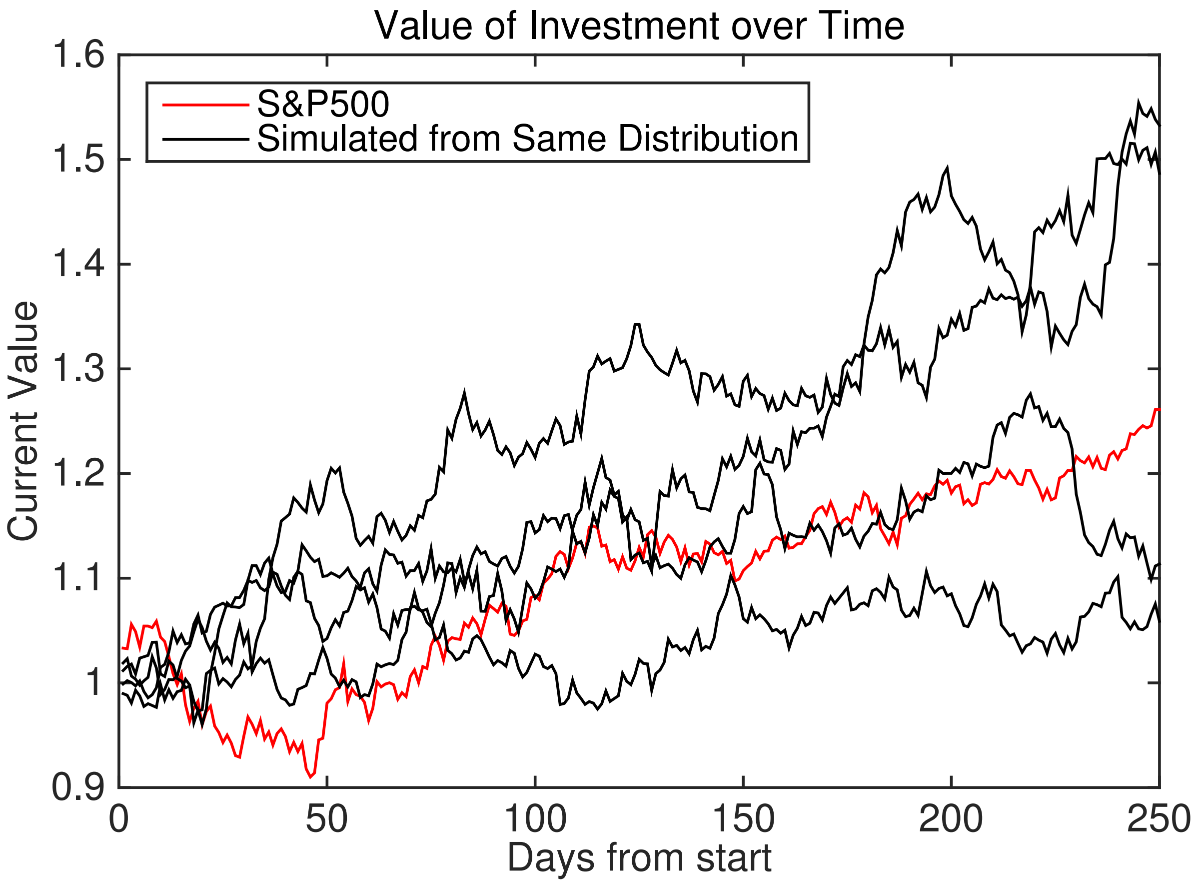


Standard and Poors 500 – Daily Returns

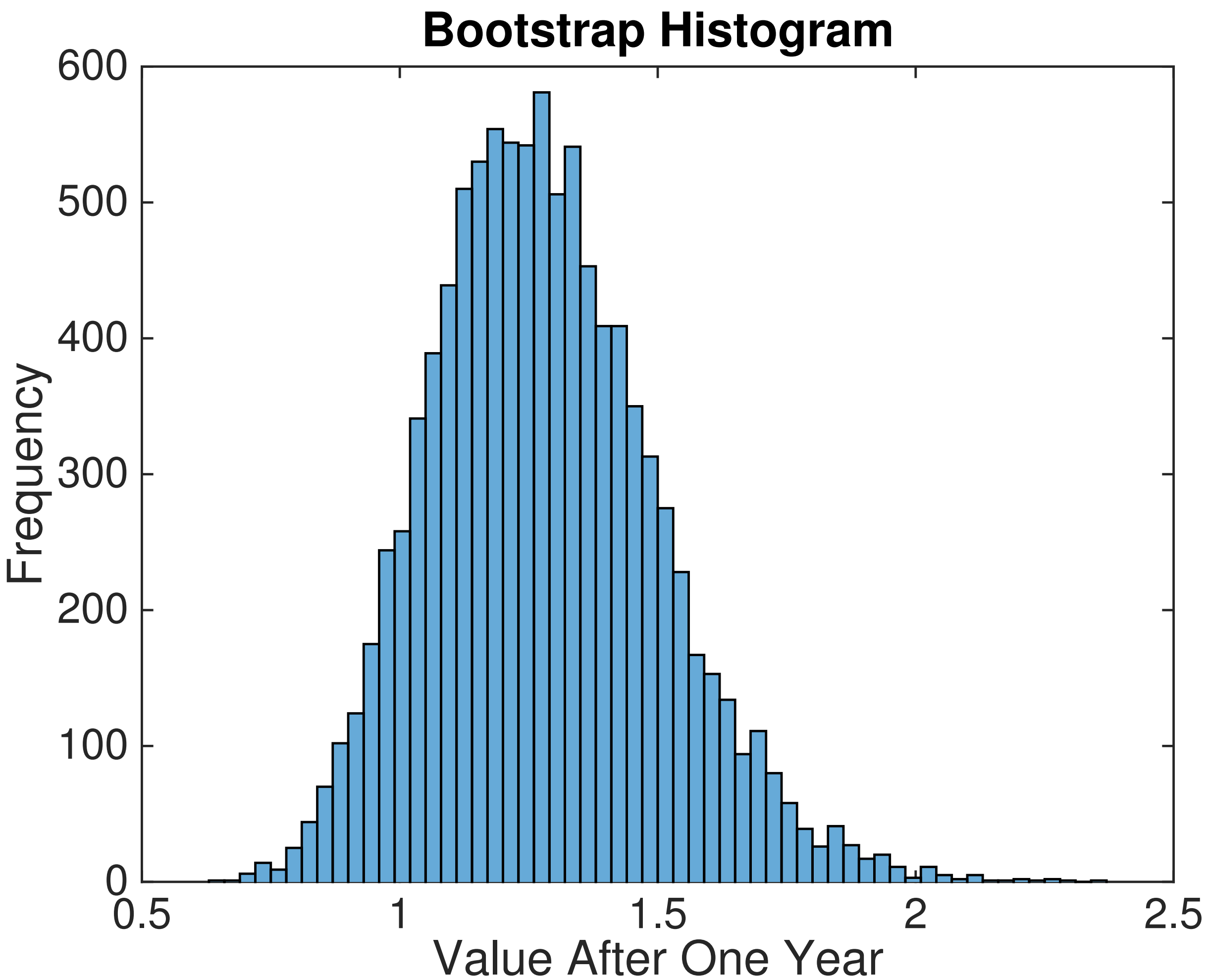
$$\mu = \mathbb{E}(R_i) \approx \sum_j R_j/n = 9.86 \times 10^{-4}, \quad \sigma = \sqrt{\mathsf{Var}(R_i)} \approx \sqrt{\sum_j (R_j - \mu)^2/n} = 0.0108$$



Standard and Poors 500 – Value Over Time



Standard and Poors 500 – Bootstrap



Standard and Poors 500 – Matlab Code

```
load -ascii 'sp500.txt'
n = length(sp500);
R = sp500;
mu = sum(R)/n
sigma = std(R)


figure(1);
plot(R);
xlabel('Days from start');
ylabel('Return');
title('Real data from S&P500');


figure(2);
Rsort = sort(R);
x = (-400:400)/10000;
y = cdf('norm', x, mu, sigma);
plot(Rsort, (1:n)/n, 'r-'); hold on;
plot(x,y,'k-');            hold off;
xlabel('x');
ylabel('F(x)');
title('Cumulative Dist. Func. for S&P500');
legend('S&P500', 'Normal(\mu,\sigma)');
```

```
figure(3);
P = cumprod(1+R);
plot(P,'r-'); hold on;
for i=1:4
    RR = R(randi(n,[n 1]));
    PP = cumprod(1+RR);
    plot(PP,'k-');
end
xlabel('Days from start');
ylabel('Current Value');
title('Value of Investment over Time');
legend('S&P500', 'Simulated from Same Distribution');
hold off;


figure(4);
P = prod(1+R);
for i=1:10000
    RR = R(randi(n,[n 1]));
    PP(i) = prod(1+RR);
end
histogram(PP);
xlabel('Value After One Year');
ylabel('Frequency');
title('Bootstrap Histogram');
```

Standard and Poors 500 – The Data

The data file is called `sp500.txt`.
It is 250 lines of plain text.
Each line contains one number R_i.
Here are the first 15 lines...

```
 .033199973
-.00048403243
 .022474383
-.0065553654
-.014074893
 .019397096
-1.0780741e-05
-.0014122923
 .0058298966
-.014425864
-.0039424103
-.014017057
-.015702278
-.010432392
 .010223599
```

Covariance

Given two random variables, X and Y, let $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$.

The *covariance* between X and Y is defined as:

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \\ &= \mathbb{E}(XY) - \mu_X \mu_Y \end{aligned}$$

Proof of equality.

$$\begin{aligned} \mathbb{E}((X - \mu_X)(Y - \mu_Y)) &= \mathbb{E}(XY - X\mu_Y - \mu_X Y + \mu_X \mu_Y) \\ &= \mathbb{E}(XY) - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y \\ &= \mathbb{E}(XY) - \mu_X \mu_Y \end{aligned}$$

Comment: If X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ and so $\mathrm{Cov}(X, Y) = 0$. The converse is not true.

Covariance/Variance of Linear Combinations

If $U = a + \sum_{i=1}^{n} b_i X_i$ and $V = c + \sum_{j=1}^{m} d_j Y_j$, then

$$\mathrm{Cov}(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{m} b_i d_j \mathrm{Cov}(X_i, Y_j)$$

If X_i's are independent, then $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$ and so

$$\begin{aligned} \mathrm{Var}\left(\sum_i X_i\right) &= \mathrm{Cov}\left(\sum_i X_i, \sum_j X_j\right) \\ &= \sum_i \mathrm{Cov}(X_i,\ X_i) \\ &= \sum_i \mathrm{Var}(X_i) \end{aligned}$$

Variance of a Binomial RV

Recall our representation of a Binomial random variable Y as a sum of independent Bernoulli's:

$$Y = \sum_{i=1}^{n} X_i$$

From this we see that

$$\mathsf{Var}(Y) = \sum_i \mathsf{Var}(X_i) = np(1-p).$$

Correlation Coefficient

The *correlation coefficient* between two random variables X and Y is denoted by ρ and defined as

$$\rho = \frac{\mathsf{Cov}(X,\ Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Let's talk about "units". Suppose that X represents a random spatial length measured in meters (m) and that Y representst a random time interval measured in seconds (s). Then, the units of Cov(X, Y) are meter-seconds, Var(X) is measured in meters-squared and Var(Y) has units of seconds-squared. Hence, ρ is unitless—the units in the numerator cancel with the units in the denominator.

One can show that

$$-1 \leq \rho \leq 1$$

always holds.

Conditional Expectation

The following formulas seem self explanatory...

Discrete case:

$$\mathbb{E}(Y \mid X = x) = \sum_y y p_{Y|X}(y|x)$$

Continuous case:

$$\mathbb{E}(Y \mid X = x) = \int y f_{Y|X}(y|x) dy$$

Arbitrary function of Y:

$$\mathbb{E}(h(Y) \mid X = x) = \int h(y) f_{Y|X}(y|x) dy$$

Prediction

Let Y be a random variable. We'd like to give a single deterministic number to represent "where" this random variable sits on the real line. The expected value, $\mathbb{E}(Y)$ is one choice that is quite reasonable if the distribution of Y is symmetric about this mean value. But, many distributions are skewed and in such cases the expected value might not be the best choice. The real question is: how do we quantify what we mean by *best choice*? One answer to that question involves the *mean squared error* (MSE):

$$\text{MSE}(\alpha) = \mathbb{E}(Y - \alpha)^2$$

To find a good estimator, pick the value of α that minimizes the MSE. To find this minimizer, we differentiate and set the derivative to zero:

$$\frac{d}{d\alpha}\text{MSE}(\alpha) = \frac{d}{d\alpha}\mathbb{E}(Y - \alpha)^2 = \mathbb{E}\left(\frac{d}{d\alpha}(Y - \alpha)^2\right) = \mathbb{E}\left(2(Y - \alpha)(-1)\right)$$

Hence, we pick α such that

$$0 = \mathbb{E}(\alpha - Y) = \alpha - \mathbb{E}(Y)$$

i.e.,

$$\alpha = \mathbb{E}(Y)$$

Conclusion: the *mean* minimizes the *mean squared error*.

Least Squares

Suppose we know from some underlying fundamental principle (say physics for example) that some parameter y is related linearly to another parameter x:

$$y = \alpha + \beta x$$

but we don't know α and β. We'd like to do experiments to determine them. A probabilistic model of the experiment has X and Y as random variables. Let's imagine we do the experiment over and over many times and have a good sense of the joint distribution of X and Y. We want to pick α and β so as to minimize

$$\mathbb{E}(Y - \alpha - \beta X)^2$$

Again, we take derivatives and set them to zero. This time we have two derivatives:

$$\frac{\partial}{\partial \alpha}\mathbb{E}(Y - \alpha - \beta X)^2 = \mathbb{E}\left(\frac{\partial}{\partial \alpha}(Y - \alpha - \beta X)^2\right) = -2\mathbb{E}(Y - \alpha - \beta X) = -2(\mu_Y - \alpha - \beta \mu_X) = 0$$

and

$$\frac{\partial}{\partial \beta}\mathbb{E}(Y - \alpha - \beta X)^2 = \mathbb{E}\left(\frac{\partial}{\partial \beta}(Y - \alpha - \beta X)^2\right) = -2\mathbb{E}\left((Y - \alpha - \beta X)X\right) = -2\left(\mathbb{E}(XY) - \alpha\mathbb{E}(X) - \beta\mathbb{E}(X^2)\right) = 0$$

Least Squares – Continued

We get two linear equations in two unknowns

$$\begin{aligned} \alpha + \beta\mu_X &= \mu_Y \\ \alpha\mu_X + \beta\mathbb{E}(X^2) &= \mathbb{E}(XY) \end{aligned}$$

Multiplying the first equation by μ_X and subtracting it from the second equation, we get

$$\beta\mathbb{E}(X^2) - \beta\mu_X^2 = \mathbb{E}(XY) - \mu_X\mu_Y$$

This equation simplifies to

$$\beta\sigma_X^2 = \sigma_{XY}$$

and so

$$\beta = \frac{\sigma_{XY}}{\sigma_X^2} = \rho\frac{\sigma_Y}{\sigma_X}$$

Finally, substituting this expression into the first equation, we get

$$\alpha = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X$$

Regression to the Mean

Suppose that a large statistics class has two midterms. Let X denote the score that a random student gets on the first midterm and let Y denote the same student's score on the second midterm. Based on prior use of these two exams, the instructor has figured out how to grade them so that the average and variance of the scores are the same

$$\mu_X = \mu_Y = \mu, \qquad \sigma_X = \sigma_Y = \sigma$$

But, those students who do well on the first midterm tend to do well on the second midterm, which is reflected in the fact that $\rho > 0$. From the calculations on the previous slide, we can estimate how a student will do on the second midterm based on his/her performance on the first one. Our estimate, denoted $\hat{Y}$, is

$$\hat{Y} = \mu - \rho\mu + \rho X$$

We can rewrite this as

$$\hat{Y} - \mu = \rho(X - \mu)$$

In words, we expect the performance of the student on the second midterm to be closer by a factor of ρ to the average than was his/her score on the first midterm. This is a famous effect called *regression to the mean*.

Sum of Poissons

Let X and Y be independent Poisson random variables with parameter λ and μ, respectively. Let $Z = X + Y$. Let's compute the probability mass function:

$$
\begin{aligned}
P(Z = n) &= P(X + Y = n) = \sum_{k=0}^{n} P(X = k,\ Y = n - k) \\
&= \sum_{k=0}^{n} P(X = k)\ P(Y = n - k) \\
&= \sum_{k=0}^{n} \frac{\lambda^k}{k!} e^{-\lambda} \, \frac{\mu^{n-k}}{(n-k)!} e^{-\mu} = e^{-(\lambda+\mu)} \sum_{k=0}^{n} \frac{\lambda^k}{k!} \, \frac{\mu^{n-k}}{(n-k)!} \\
&= e^{-(\lambda+\mu)} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} \lambda^k \mu^{n-k} = e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^n}{n!}
\end{aligned}
$$

Conclusion: The sum is Poisson with parameter $\lambda + \mu$. The result can be extended to a sum of any number of *independent* Poisson random variables:

$$
X_k \sim \text{Poisson}(\lambda_k) \qquad \Longrightarrow \qquad \sum_k X_k \sim \text{Poisson}\left(\sum_k \lambda_k\right)
$$

Sum of Normals

Let X and Y be independent Normal(0,1) r.v.'s and $Z = X + Y$. Compute Z's cdf:

$$P(Z \le z) = P(X + Y \le z) = \int_{-\infty}^{\infty} f(x) P(Y \le z - x) dx$$
$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2/2} \int_{-\infty}^{z-x} e^{-y^2/2} dy dx$$

Differentiating, we compute the density function for Z:

$$\begin{aligned}
f_Z(z) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2/2} e^{-(z-x)^2/2} dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2 + xz - z^2/2} dx \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x-z/2)^2 + z^2/4 - z^2/2} dx = \frac{1}{2\pi} e^{-z^2/4} \int_{-\infty}^{\infty} e^{-(x-z/2)^2} dx \\
&= \frac{1}{2\pi} e^{-z^2/4} \int_{-\infty}^{\infty} e^{-x^2} dx = \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-z^2/4}
\end{aligned}$$

Conclusion: The sum is Normal with mean 0 and variance 2. The result can be extended to a sum of any number of *independent* Normal random variables:

$$X_k \sim \text{Normal}(\mu_k,\ \sigma_k^2) \quad \Longrightarrow \quad \sum_k X_k \sim \text{Normal}\left(\sum_k \mu_k,\ \sum_k \sigma_k^2\right)$$