Stat 5101 Lecture Slides: Deck 2

Axioms for Probability and Expectation, Consequences of Linearity of Expectation, Random Vectors, Time Series, Laws of Large Numbers, Bernoulli, Binomial, Hypergeometric, and Discrete Uniform Distributions

Charles J. Geyer
School of Statistics
University of Minnesota

Axioms

An expectation operator is a mapping $X \mapsto E(X)$ of random variables to real numbers that satisfies the following axioms:

$$E(X + Y) = E(X) + E(Y)$$

for any random variables X and Y,

$$E(X) \geq 0$$

for any nonnegative random variable X (one such that $X(s) \geq 0$ for all s in the sample space),

$$E(aX) = aE(X)$$

for any random variable X and any constant a, and

$$E(Y) = 1$$

when Y is the constant random variable $s \mapsto 1$.

Axioms (cont.)

The fourth axiom is usually written, a bit sloppily, as

$$E(1) = 1$$

The reason this is sloppy is that on the left-hand side 1 must indicate a random variable, because the argument of an expectation operator is always a random variable, and on the right-hand side 1 must indicate a real number, because the value of an expectation operator is always a real number.

When we have a constant as an argument of an expectation operator, we always take this to mean a constant random variable.

Axiom Summary

$$E(X+Y) = E(X) + E(Y) \tag{1}$$

$$E(X) \geq 0, \qquad \text{when } X \geq 0 \tag{2}$$

$$E(aX) = aE(X) \tag{3}$$

$$E(1) = 1 \tag{4}$$

(3) and (4) together imply

$$E(a) = a, \qquad \text{for any constant } a$$

It can be shown (but we won't here) that when the sample space is finite these axioms hold if and only if the expectation operator is defined in terms of a PMF as we did before.

Axioms (cont.)

$$E(X+Y) = E(X) + E(Y)$$

says an addition operation can be pulled outside an expectation.

$$X \geq 0 \quad \text{implies} \quad E(X) \geq 0$$

says nonnegativity can be pulled outside an expectation.

$$E(aX) = aE(X)$$

says a constant can be pulled outside an expectation.

Axioms (cont.)

Many students are tempted to overgeneralize, and think *anything* can be pulled outside an expectation. Wrong!

In general

$$E(XY) \neq E(X)E(Y)$$
$$E(X/Y) \neq E(X)/E(Y)$$
$$E\{g(X)\} \neq g\Big(E\{X\}\Big)$$

although we may have equality for certain special cases.

Axioms (cont.)

We do have

$$E(X - Y) = E(X) - E(Y)$$

because

$$E(X - Y) = E\{X + (-1)Y\} = E(X) + (-1)E(Y)$$

by axioms (1) and (3).

Axioms (cont.)

We do have

$$E(a + bX) = a + bE(X)$$

because

$$E(a + bX) = E(a) + E(bX) = a + bE(X)$$

by axioms (1), (3), and (4).

Axiom Summary (cont.)

$$E(X \pm Y) = E(X) \pm E(Y)$$

addition and subtraction come out

$$X \geq 0 \quad \text{implies} \quad E(X) \geq 0$$

nonnegative comes out

$$E(aX) = aE(X)$$

constants come out

$$E(a + bX) = a + bE(X)$$

linear functions come out. But that's all!

Linearity of Expectation

By mathematical induction the "addition comes out" axiom extends to any finite number of random variables.

For any random variables X_1, ..., X_n

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$$

More generally, for any random variables X_1, ..., X_n and any constants a_1, ..., a_n

$$E(a_1X_1 + \cdots + a_nX_n) = a_1E(X_1) + \cdots + a_nE(X_n)$$

This very useful property is called *linearity of expectation*.

Linearity of Expectation (cont.)

Using linearity of expectation, we could have calculated the expectation of a binomial random variable much more simply than we did before.

If $X_1, \ldots, X_n$ are random variables having the same expectation μ, then the random variable $Y = X_1 + \cdots + X_n$ has expectation $n\mu$.

A Bin(n, p) random variable Y is equal in distribution to the sum of n IID Ber(p) random variables having expectation p.

Conclusion: $E(Y) = np$. No calculation necessary.

Why Axioms?

Well, partly because mathematicians like them.

Euclid had axioms for geometry. Every other area of mathematics has them too. So probability theory should have them too.

But also they are very useful, as we just saw.

Variance

Another word for expectation is *mean*. We sometimes say $E(X)$ is the mean of X or the mean of the distribution of X.

The mean of X is our "best guess" value of X before it is observed (for some definition of "best").

How far away will X be from the mean when it is observed?

The *variance* of X

$$\text{var}(X) = E\{(X - \mu)^2\}$$

where $\mu = E(X)$, is one notion that helps answer this question.

Variance (cont.)

A variance is an expectation $E\{(X-\mu)^2\}$, so it is a real number.

Since $(X-\mu)^2 \geq 0$ always, $\text{var}(X) \geq 0$ for any random variable X by the axiom about nonnegativity.

This is another important "sanity check". Probabilities are between zero and one, inclusive. Variances are nonnegative.

Variance (cont.)

When there are many probability distributions under consideration, we decorate the variance operator with a parameter

$$\text{var}_\theta(X) = E_\theta\{(X - \mu)^2\}$$

Example: If X is a Ber(p) random variable, then we already know the mean is p, so the variance is

$$\text{var}_p(X) = E_p\{(X - p)^2\} = p(1 - p)$$

(this was a homework problem, we won't redo it here).

Why Variance?

Variance is expected squared deviation from the mean. Why that for a measure of spread-out-ness of a random variable?

Partly mathematical convenience, and partly the deep role it plays in large sample theory (much more on this later).

But other measures are possible and useful in certain contexts, for example, $E\{|X - \mu|\}$, the expected absolute deviation from the mean.

Standard Deviation

One issue with variance is that it has the wrong units. If X is a length with dimension feet (ft). Then $\mu = E(X)$ also has dimension ft. Hence $(X - \mu)^2$ and var(X) have dimension square feet (ft^2).

So var(X) is not comparable to values of X.

Standard Deviation (cont.)

For this reason we introduce

$$\mathrm{sd}(X) = \sqrt{\mathrm{var}(X)}$$

(since variances are nonnegative, the square root, meaning *non-negative* square root is always defined).

If X has dimension feet (ft), then so does $\mathrm{sd}(X)$.

$\mathrm{sd}(X)$ is called the *standard deviation* of X.

Standard Deviation (cont.)

In applications, $\text{sd}(X)$ is preferable to $\text{var}(X)$.

Since $\text{sd}(X)$ has the same units as X, values of X can be directly compared to $\text{sd}(X)$. We say $X-\mu$ is "large" if it is large compared to $\text{sd}(X)$.

In theory, $\text{var}(X)$ is preferable to $\text{sd}(X)$. Generally, $\text{sd}(X)$ can only be calculated by calculating $\text{var}(X)$ first. Moreover, it is variance that appears many theoretical contexts.

When we choose single letters to indicate them, we usually choose the single letter for standard deviation.

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

Standard Deviation (cont.)

Be careful! It is easy to miss the distinction if you are not paying attention.

If you are told the variance is θ and you need the standard deviation, then it is $\sqrt{\theta}$.

If you are told the standard deviation is θ and you need the variance, then it is θ^2.

The "Short Cut" Formula

For any random variable X

$$\text{var}(X) = E(X^2) - E(X)^2$$

because, with $\mu = E(X)$,

$$\begin{aligned}
\text{var}(X) &= E\{(X - \mu)^2\} \\
&= E(X^2 - 2\mu X + \mu^2) \\
&= E(X^2) - 2\mu E(X) + \mu^2 E(1) \\
&= E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - \mu^2
\end{aligned}$$

Variance of the Bernoulli Distribution

For homework we proved that, if X is a Ber(p) random variable, then $E(X^k) = p$ for all positive integers k.

Hence

$$\text{var}(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p)$$

Variance of the Binomial Distribution

If X is a Bin(n, p) random variable, then for homework we proved that $E\{X(X-1)\} = n(n-1)p^2$ and in class we proved that $E(X) = np$.

Since

$$E\{X(X-1)\} = E(X^2 - X) = E(X^2) - E(X)$$

we have

$$E(X^2) = E\{X(X-1)\} + E(X)$$

Variance of the Binomial Distribution (cont.)

Hence

$$\begin{aligned}
\text{var}(X) &= E(X^2) - E(X)^2 \\
&= E\{X(X-1)\} + E(X) - E(X)^2 \\
&= n(n-1)p^2 + np - n^2p^2 \\
&= -np^2 + np \\
&= np(1-p)
\end{aligned}$$

Mean of the Discrete Uniform Distribution

Suppose X is a random variable having the discrete uniform distribution on the set $\{1, \ldots, n\}$. Then

$$E(X) = \frac{1}{n} \sum_{i=1}^{n} i$$

Mean of the Discrete Uniform Distribution (cont.)

There is a story about the famous mathematician Carl Friedrich Gauss. When he was still in elementary school, the teacher gave the class the problem of adding the numbers from 1 to 100, hoping to occupy them for a while, but young Gauss got the answer almost immediately.

Presumably, he had figured out the following argument. Write down the numbers to be added twice and add

$$\begin{array}{rccccc} & 1 & 2 & \ldots & 99 & 100 \\ + & 100 & 99 & \ldots & 2 & 1 \\ \hline & 101 & 101 & \ldots & 101 & 101 \end{array}$$

In general, there are n pairs that sum to $n+1$, so the total is $n(n+1)$, which is twice the desired answer.

Mean of the Discrete Uniform Distribution (cont.)

Hence

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

and

$$E(X) = \frac{n+1}{2}$$

Variance of the Discrete Uniform Distribution

Suppose X is a random variable having the discrete uniform distribution on the set $\{1, \ldots, n\}$.

To do the variance we need to know

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

No cute story for this. We prove by mathematical induction.

To do this we verify the case $n = 1$ is correct.

$$1 = \frac{1 \cdot (1+1)(2 \cdot 1 + 1)}{6} = \frac{1 \cdot 2 \cdot 3}{6}$$

That checks.

Variance of the Discrete Uniform Distribution (cont.)

Then we check that if the $n = k$ case is correct, this implies the $n = k + 1$ case.

$$\begin{aligned}
\sum_{i=1}^{k+1} i^2 &= (k+1)^2 + \sum_{i=1}^{k} i^2 \\
&= (k+1)^2 + \frac{k(k+1)(2k+1)}{6} \\
&= (k+1)\left[(k+1) + \frac{k(2k+1)}{6}\right] \\
&= (k+1)\frac{6k + 6 + 2k^2 + k}{6}
\end{aligned}$$

Variance of the Discrete Uniform Distribution (cont.)

This should equal $n(n+1)(2n+1)/6$ with $k+1$ substituted for n

$$\frac{(k+1)[(k+1)+1][2(k+1)+1]}{6} = (k+1)\frac{(k+2)(2k+3)}{6}$$
$$= (k+1)\frac{2k^2+7k+6}{6}$$

And this is what we got before. So the induction step checks. And we are done.

Variance of the Discrete Uniform Distribution (cont.)

$$
\begin{aligned}
\text{var}(X) &= E(X^2) - E(X)^2 \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} i\right)^2 \\
&= \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 \\
&= (n+1)\left[\frac{2n+1}{6} - \frac{n+1}{4}\right] \\
&= (n+1)\left[\frac{8n+4-6n-6}{24}\right] \\
&= \frac{(n+1)(n-1)}{12}
\end{aligned}
$$

Variance of the Discrete Uniform Distribution (cont.)

In summary, for the discrete uniform distribution on $\{1, \ldots, n\}$

$$E(X) = \frac{n+1}{2}$$
$$\text{var}(X) = \frac{(n+1)(n-1)}{12}$$

If you are now panicking about how complicated this calculation is for this very simple distribution and wondering how hard this course will be, don't. This is a fairly horrible example, despite the simplicity of the distribution.

The Mean Square Error Formula

The *mean square error* of a number a considered as a prediction of (the value of, when observed) a random variable X is

$$\text{mse}(a) = E\{(X - a)^2\}$$

Write $\mu = E(X)$. Then this can also be calculated

$$\text{mse}(a) = \text{var}(X) + (\mu - a)^2$$

Proof of the Mean Square Error Formula

$$\begin{aligned}
\text{mse}(a) &= E\{(X - a)^2\} \\
&= E\{(X - \mu + \mu - a)^2\} \\
&= E\{(X - \mu)^2 + 2(\mu - a)(X - \mu) + (\mu - a)^2\} \\
&= E\{(X - \mu)^2\} + 2(\mu - a)E(X - \mu) + (\mu - a)^2 \\
&= \text{var}(X) + (\mu - a)^2
\end{aligned}$$

because

$$E(X - \mu) = E(X) - \mu = 0$$

Minimizing Mean Square Error

The number a considered as a prediction of a random variable X that minimizes the mean square error

$$\mathrm{mse}(a) = \mathrm{var}(X) + (\mu - a)^2$$

is clearly $a = \mu$, because the first term on the right-hand side does not contain a, and the second term on the right-hand, being a square, is nonnegative, and minimized when $a = \mu$, which makes it zero.

Conclusion: $E(X)$ is the best prediction of X, where "best" is defined to mean *minimizing mean square error* of the prediction.

Minimizing Mean Square Error (cont.)

Philosophically, this can't serve as a definition of expectation, because the definition would be circular. Mean square error is defined in terms of expectation, and expectation is defined in terms of mean square error.

Practically, this does give a very precise property of expectation that does tell us something important.

Moreover, it makes mathematically precise our blather about thinking of expectation as best prediction. It is, but only when "best" means minimizing mean square error.

Mean and Variance of Linear Functions

If X is a random variable and a and b are constants, then

$$E(a + bX) = a + bE(X)$$
$$\text{var}(a + bX) = b^2 \text{var}(X)$$

These are used often, remember them.

The first we have seen before and is intuitively obvious, the other not so obvious. Write $\mu = E(X)$. Then

$$\begin{aligned} \text{var}(a + bX) &= E\{[(a + bX) - (a + b\mu)]^2\} \\ &= E\{(bX - b\mu)^2\} \\ &= E\{b^2(X - \mu)^2\} \\ &= b^2 E\{(X - \mu)^2\} \\ &= b^2 \text{var}(X) \end{aligned}$$

Mean of a Random Vector

For any random vector $\mathbf{X} = (X_1, \ldots, X_n)$ we define

$$E(\mathbf{X}) = \boldsymbol{\mu}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and

$$\mu_i = E(X_i), \qquad i = 1, \ldots, n$$

The expectation of a random vector is a vector, the components of which are the expectations of the components of the random vector.

Mean of a Random Vector (cont.)

In one sense, $E(\mathbf{X}) = \boldsymbol{\mu}$ is merely a notational convenience. We write one vector equation, rather than n scalar equations

$$\mu_i = E(X_i), \qquad i = 1, \ldots, n$$

In another sense, this is an important concept, trivial though it may be, because it is essential part of treating $\mathbf{X}$ as a single object (rather than n objects, its components).

Variance of a Random Vector

So if the mean of a random vector is an ordinary numeric vector, is the same true for variance? No!

Covariance

The *covariance* of random variables X and Y is

$$\text{cov}(X, Y) = E\{(X - \mu)(Y - \nu)\}$$

where

$$\mu = E(X)$$
$$\nu = E(Y)$$

Covariance (cont.)

Covariance generalizes variance.

$$\mathrm{cov}(X, X) = \mathrm{var}(X)$$

because

$$\mathrm{cov}(X, X) = E\{(X - \mu)(X - \mu)\} = E\{(X - \mu)^2\} = \mathrm{var}(X)$$

The covariance of a random variable with itself is the variance.

Covariance (cont.)

A covariance operator is a symmetric function of its arguments

$$\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$$

because multiplication is commutative.

Covariance (cont.)

The generalization of the "short cut" formula to covariance is

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

Note that in case $X = Y$ we get the "short cut" formula for variance.

The proof is a homework problem.

Covariance (cont.)

The generalization of the formula about taking out linear functions to covariance is

$$\operatorname{cov}(a + bX, c + dY) = bd \operatorname{cov}(X, Y)$$

Note that in case $X = Y$, $a = c$, and $b = d$ we get the formula we already proved for variance.

The proof is a homework problem.

Variance of a Random Vector (cont.)

The *variance* of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is an ordinary numeric *matrix*, the $n \times n$ matrix having components $\mathrm{cov}(X_i, X_j)$

$$\mathrm{var}(\mathbf{X}) = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \cdots & \mathrm{var}(X_n) \end{pmatrix}$$

Why?

Variance of a Random Vector (cont.)

There is no agreement in the literature about what to call the matrix defined to be var($\mathbf{X}$) on the preceding slide.

Some authors call it the *covariance* matrix of $\mathbf{X}$ because its elements are covariances.

Some authors call it the *variance-covariance* matrix of $\mathbf{X}$ because its diagonal elements are variances.

Some authors, disgusted with the terminological confusion, call it the *dispersion* matrix of $\mathbf{X}$.

We will only call it the *variance* matrix of $\mathbf{X}$ in this course.

Notational Conventions

We have a convention X for random variables, x for ordinary variables.

We have another convention A for sets, a for elements.

We have another convention $\mathbf{A}$ for matrices, $\mathbf{a}$ for vectors.

Combining the first and third, we have $\mathbf{X}$ for random vectors, X for random variables (random scalars).

But now we are in trouble. We can't tell whether a boldface capital letter is a random vector or an ordinary matrix.

Similarly, we would be in trouble if we had a random matrix.

Notational Conventions (cont.)

Typographical conventions can't do everything.

Sometimes you just have to read what symbols are defined to mean.

Or sometimes you just have to figure out from the context in which a symbol is used what it could possibly mean.

Notational Conventions (cont.)

If we write

$$\mathbf{M} = \mathrm{var}(\mathbf{X})$$

then you just have to figure out

- the argument of the variance operator must be a random thingummy, presumably a random vector because of the boldface, although the text should make this clear, and

- that makes $\mathbf{M}$ an ordinary (non-random) matrix, which is why it is (also) denoted by a boldface capital letter.

Matrix Multiplication

A course named "linear algebra" is not a prerequisite for this course, but you are assumed to have at least seen matrices and matrix multiplication somewhere. For review, if $\mathbf{A}$ and $\mathbf{B}$ are matrices, and the column dimension of $\mathbf{A}$ is the same as the row dimension of $\mathbf{B}$, then the product $\mathbf{AB} = \mathbf{C}$ is defined by

$$c_{ik} = \sum_j a_{ij} b_{jk}$$

where a_{ij} are components of $\mathbf{A}$ and similarly for the other two.

The row dimensions of $\mathbf{C}$ and $\mathbf{A}$ are the same. The column dimensions of $\mathbf{C}$ and $\mathbf{B}$ are the same.

Matrix Multiplication (cont.)

If $\mathbf{A}$ is $l \times m$ and $\mathbf{B}$ is $m \times n$, then $\mathbf{C}$ is $l \times n$ and

$$c_{ik} = \sum_{j=1}^{m} a_{ij} b_{jk}, \qquad i = 1, \ldots, l \text{ and } k = 1, \ldots, n$$

Matrix Multiplication (cont.)

Multiplication of ordinary numbers (scalars) is *commutative*

$$ab = ba$$

for any numbers a and b.

Matrix multiplication is not. In general,

$$\mathbf{AB} \neq \mathbf{BA}$$

In general, it is not even true that $\mathbf{AB}$ and $\mathbf{BA}$ are both defined. The dimensions may be such that only one is defined.

Multiplying a Matrix and a Vector

If we think of vectors as matrices having one dimension (row or column) equal to one, then we don't have to define a new kind of multiplication involving vectors. If vectors are matrices, then we use the matrix multiplication already defined.

However, we keep the lower case boldface for vectors, even when thinking of them as matrices. So now there are two kinds of vectors, row vectors

$$\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \ldots & x_n \end{pmatrix}$$

and column vectors

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Transpose of a Matrix

If $\mathbf{A}$ has components a_{ij}, then the *transpose* of $\mathbf{A}$, denoted $\mathbf{A}^T$ has components a_{ji}.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

$$\mathbf{A}^T = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix}$$

Note $(\mathbf{A}^T)^T = \mathbf{A}$.

Symmetric Matrices

A matrix $\mathbf{M}$ is *symmetric* if $\mathbf{M}^T = \mathbf{M}$.

Expressed in components, this says $m_{ij} = m_{ji}$ for all i and j.

Note that a symmetric matrix is automatically *square*, meaning the row and column dimensions are the same.

Because $\text{cov}(X_i, X_j) = \text{cov}(X_j, X_i)$, every variance matrix is symmetric.

Transpose of a Vector

The transpose of a row vector is a column vector and vice versa.

Linear Functions

The analog of scalar-to-scalar linear functions

$$x \mapsto a + bx$$

is vector-to-vector linear functions

We have row vector to row vector linear functions

$$\mathbf{x} \mapsto \mathbf{a} + \mathbf{x}\mathbf{B}$$

and have column vector to column vector linear functions

$$\mathbf{x} \mapsto \mathbf{a} + \mathbf{B}\mathbf{x}$$

Almost the same, but slightly different, very confusing.

A Completely Arbitrary Convention

To avoid this confusion, we make a rule

> *Whenever we think of a vector as a matrix, it is always a column vector!*

Almost everybody uses the same convention.

This does not mean vectors are really matrices. They aren't. Treating vectors as matrices is just a stupid mathematician trick that avoids a separate definition for the meaning of $\mathbf{AB}$ and $\mathbf{Bx}$.

Linear Functions (cont.)

Now row vector to row vector linear functions

$$\mathbf{x}^T \mapsto \mathbf{a}^T + \mathbf{x}^T \mathbf{B}$$

and column vector to column vector linear functions

$$\mathbf{x} \mapsto \mathbf{a} + \mathbf{B}\mathbf{x}$$

look different enough so they can't be confused.

The lower case boldface letters are column vectors unless they are transposed, in which case they are row vectors.

Matrix Multiplication and Transposition

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$
$$(\mathbf{ABC})^T = \mathbf{C}^T\mathbf{B}^T\mathbf{A}^T$$
$$(\mathbf{ABCD})^T = \mathbf{D}^T\mathbf{C}^T\mathbf{B}^T\mathbf{A}^T$$

and so forth. Also

$$(\mathbf{Ax})^T = \mathbf{x}^T\mathbf{A}^T$$
$$(\mathbf{x}^T\mathbf{B})^T = \mathbf{B}^T\mathbf{x}$$

Linear Functions (cont.)

Taking transposes of both sides of the column vector to column vector linear function

$$\mathbf{x} \mapsto \mathbf{a} + \mathbf{B}\mathbf{x}$$

gives the row vector to row vector linear function

$$\mathbf{x}^T \mapsto \mathbf{a}^T + \mathbf{x}^T \mathbf{B}^T$$

so it enough to know about one of these.

The preference in the "completely arbitrary convention" for column vectors means we only need to do one of these.

Linear Functions (cont.)

Finally things get simple (but don't forget all that stuff about transposes).

A general vector-to-vector linear function has the form

$$\mathbf{y} = \mathbf{a} + \mathbf{B}\mathbf{x}$$

where the dimensions have to be such that the expression makes sense

$$\underbrace{\mathbf{y}}_{m\times 1} = \underbrace{\mathbf{a}}_{m\times 1} + \underbrace{\mathbf{B}}_{m\times n}\underbrace{\mathbf{x}}_{n\times 1}$$

for any m and n.

This function maps vectors of dimension n to vectors of dimension m.

Linear Functions (cont.)

If $\mathbf{X}$ is a random vector and if $\mathbf{a}$ and $\mathbf{B}$ are a non-random vector and matrix, respectively, then

$$E(\mathbf{a} + \mathbf{B}\mathbf{X}) = \mathbf{a} + \mathbf{B}E(\mathbf{X})$$
$$\text{var}(\mathbf{a} + \mathbf{B}\mathbf{X}) = \mathbf{B}\,\text{var}(\mathbf{X})\mathbf{B}^T$$

Sometimes we write this a bit more neatly. If $\mathbf{X}$ is a random vector having mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$ and if $\mathbf{a}$ and $\mathbf{B}$ are a non-random vector and matrix, respectively, then

$$E(\mathbf{a} + \mathbf{B}\mathbf{X}) = \mathbf{a} + \mathbf{B}\boldsymbol{\mu}$$
$$\text{var}(\mathbf{a} + \mathbf{B}\mathbf{X}) = \mathbf{B}\mathbf{M}\mathbf{B}^T$$

Linear Functions (cont.)

Since we have no other theorems about mean vectors and variance matrices, we must reduce this to the scalar case by introducing components.

$$E(\mathbf{a} + \mathbf{B}\mathbf{X}) = \mathbf{a} + \mathbf{B}\boldsymbol{\mu}$$

written out in components is

$$E\left(a_i + \sum_j b_{ij} X_j\right) = a_i + \sum_j b_{ij} E(X_j)$$

and this is just linearity of expectation.

Inner and Outer Products

For vectors $\mathbf{x}$ and $\mathbf{y}$ of the same dimension n,

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

is a scalar (1×1 matrix), and

$$\mathbf{x}\mathbf{y}^T$$

is an $n \times n$ matrix with components $x_i y_j$.

The former is called the *inner product* of these two vectors and the latter is called the *outer product*.

Linear Functions (cont.)

Another expression for the variance matrix is

$$\text{var}(\mathbf{X}) = E\{(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\}$$

where $\boldsymbol{\mu} = E(\mathbf{X})$. The argument of the expectation operator is an outer product.

$$\begin{aligned}
\text{var}(\mathbf{a} + \mathbf{B}\mathbf{X}) &= E\{[(\mathbf{a} + \mathbf{B}\mathbf{X}) - (\mathbf{a} + \mathbf{B}\boldsymbol{\mu})][(\mathbf{a} + \mathbf{B}\mathbf{X}) - (\mathbf{a} + \mathbf{B}\boldsymbol{\mu})]^T\} \\
&= E\{(\mathbf{B}\mathbf{X} - \mathbf{B}\boldsymbol{\mu})(\mathbf{B}\mathbf{X} - \mathbf{B}\boldsymbol{\mu})^T\} \\
&= E\{[\mathbf{B}(\mathbf{X} - \boldsymbol{\mu})][\mathbf{B}(\mathbf{X} - \boldsymbol{\mu})]^T\} \\
&= E\{\mathbf{B}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\mathbf{B}^T\} \\
&= \mathbf{B}E\{(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\}\mathbf{B}^T \\
&= \mathbf{B}\,\text{var}(\mathbf{X})\mathbf{B}^T
\end{aligned}$$

Positive Definite Matrices

What is the property of variance matrices analogous to the property $\text{var}(X) \geq 0$ for random variables?

Consider the linear function $\mathbf{x} \mapsto \mathbf{b}^T\mathbf{x}$. This is a special case of the the general linear function formula with $\mathbf{a} = 0$ and $\mathbf{b}^T = \mathbf{B}$. Hence

$$0 \leq \text{var}(\mathbf{b}^T\mathbf{X}) = \mathbf{b}^T\mathbf{M}\mathbf{b}$$

where $\mathbf{M} = \text{var}(\mathbf{X})$.

Positive Definite Matrices (cont.)

An arbitrary symmetric matrix $\mathbf{M}$ is *positive semidefinite* if

$$\mathbf{b}^T\mathbf{M}\mathbf{b} \geq 0, \qquad \text{for all vectors } \mathbf{b}$$

and is *positive definite* if

$$\mathbf{b}^T\mathbf{M}\mathbf{b} > 0, \qquad \text{for all nonzero vectors } \mathbf{b}$$

(the zero vector is the vector having all components zero).

Every variance matrix is *symmetric* and *positive semidefinite*.

Variance of a Sum

We now want to look at the special case where the linear function is the sum of the components. If $\mathbf{u}$ is the vector whose components are all equal to one, then

$$\mathbf{u}^T\mathbf{X} = \sum_{i=1}^{n} X_i$$

where $\mathbf{X} = (X_1, \ldots, X_n)$.

Variance of a Sum (cont.)

If $\mathbf{M}$ is the variance matrix of $\mathbf{X}$, then

$$\begin{aligned}
\text{var}\left(\sum_{i=1}^{n} X_i\right) &= \mathbf{u}^T\mathbf{M}\mathbf{u} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \text{cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} \text{var}(X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \text{cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} \text{var}(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \text{cov}(X_i, X_j)
\end{aligned}$$

Variance of a Sum (cont.)

In short the variance of a sum is the sum of all the variances and all the covariances.

Variance is more complicated than expectation.

The expectation of a sum is the sum of the expectations, but the analog is not — in general — true for variance.

Uncorrelated

Random variables X and Y are *uncorrelated* if $\text{cov}(X, Y) = 0$.

Then

$$0 = \text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

so "multiplication comes out of expectation"

$$E(XY) = E(X)E(Y)$$

Note that this holds, not in general, but if and only if X and Y are uncorrelated.

Variance of a Sum (cont.)

We say a sequence of random variables $X_1, \ldots, X_n$ is *uncorrelated* if X_i and X_j are uncorrelated whenever $i \neq j$. Then

$$\text{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{var}(X_i)$$

(the variance of the sum is the sum of the variances).

Note this holds if the random variables are **uncorrelated**, not in general.

Expectation and Variance of a Sum

To review: the expectation of a sum is the sum of the expectations,

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

whether or not the random variables are independent or uncorrelated.

But the variance of a sum is the sum of the variances

$$\text{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{var}(X_i)$$

only if the random variables are uncorrelated or independent (since independent implies uncorrelated), not in general.

Axiomatic Characterization of Independence

A sequence $X_1, \ldots, X_n$ is *independent* if and only if

$$E\left(\prod_{i=1}^{n} h_i(X_i)\right) = \prod_{i=1}^{n} E\{h_i(X_i)\}$$

for any functions $h_1, \ldots, h_n$.

"Multiplication comes out of expectation" not just for the variables themselves, but for any functions of them.

From this is clear that, if $X_1, \ldots, X_n$ are independent, then so are $g_1(X_1), \ldots, g_n(X_n)$ for any functions $g_1, \ldots, g_n$.

Axiomatic Characterization of Independence (cont.)

We prove the two notions of independent are the same. Suppose the PMF factors

$$f(\mathbf{x}) = \prod_{i=1}^{n} f_i(x_i)$$

and the support is a product $S_1 \times \cdots \times S_n$, then

$$E\left(\prod_{i=1}^{n} h_i(X_i)\right) = \sum_{x_1 \in S_1} \cdots \sum_{x_n \in S_n} \prod_{i=1}^{n} h_i(x_i) f_i(x_i) = \prod_{i=1}^{n} E\{h_i(X_i)\}$$

Axiomatic Characterization of Independence (cont.)

Conversely, suppose the "multiplication comes out of expectation" property holds. Consider the case $h_i = I_{\{x_i\}}$ — that is, each h_i is the indicator function of the point x_i — so

$$\begin{aligned} f(\mathbf{x}) &= \Pr(X_1 = x_1 \text{ and } \cdots \text{ and } X_n = x_n) \\ &= E\left(\prod_{i=1}^{n} I_{\{x_i\}}(X_i)\right) \\ &= \prod_{i=1}^{n} E\left(I_{\{x_i\}}(X_i)\right) \\ &= \prod_{i=1}^{n} f_i(x_i) \end{aligned}$$

Independent versus Uncorrelated

Independent implies uncorrelated. The converse is, in general, false.

If X and Y are independent, then

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$$

Independent versus Uncorrelated (cont.)

Conversely, let X be any nonconstant random variable such that X and $-X$ have the same distribution. An example is the uniform distribution on $\{-2, -1, 0, 1, 2\}$.

Then X and $Y = X^2$ are uncorrelated

$$\begin{aligned} \operatorname{cov}(X, Y) &= \operatorname{cov}(X, X^2) \\ &= E(X^3) - E(X)E(X^2) \\ &= E(-X^3) - E(-X)E(X^2) \\ &= -\operatorname{cov}(X, Y) \end{aligned}$$

the third equals sign being that X and $-X$ have the same distribution.

Since the only number x that satisfies $x = -x$ is zero, we have $\operatorname{cov}(X, Y) = 0$.

Independent versus Uncorrelated (cont.)

Conversely, the PMF of the random vector (X, Y) is given by the table

		$y = x^2$		
		0	1	4
	-2	0	0	$1/5$
	-1	0	$1/5$	0
x	0	$1/5$	0	0
	1	0	$1/5$	0
	2	0	0	$1/5$

The support is not a Cartesian product, so the variables are not independent.

Independent versus Uncorrelated (cont.)

Independent implies uncorrelated.

Uncorrelated **does not** imply independent.

Uncorrelated is a pairwise property: $\text{cov}(X_i, X_j) = 0$ only looks at two variables at a time.

Independent **is not** a pairwise property: this was a homework problem.

Exchangeability

We want to consider some dependent sequences of random variables.

The simplest form of dependence is exchangeability.

A function $f : S \to T$ is invertible if it has an inverse $g : T \to S$, which satisfies $g[f(x)] = x$ for all $x \in S$ and $f[g(y)] = y$ for all $y \in T$ (deck 1, slide 14).

An invertible function $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is called a *permutation* because it produces a reordering of 1, ..., n.

We know there are $n!$ such functions (the number of permutations of n things).

Exchangeability (cont.)

A sequence of random variables X_1, ..., X_n is *exchangeable* if the random vectors $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (X_{\pi(1)}, \ldots, X_{\pi(n)})$ have the same distribution for any permutation π.

In particular, X_1, ..., X_n are identically distributed but may be dependent.

Every variable has the same variance as any other, and every pair of different variables has the same covariance as any other. Thus

$$\text{var}\left(\sum_{i=1}^{n} X_i\right) = n\,\text{var}(X_1) + n(n-1)\,\text{cov}(X_1, X_2)$$

Sampling With and Without Replacement

When the random variables $X_1, \ldots, X_n$ are uniformly distributed on a set S of size N and we define

$$Y_i = g(X_i)$$

for any function $g : S \to \mathbb{R}$, then we say $X_1, \ldots, X_n$ are a *sample* of size n from a *population* S of size N.

Note: the sample $X_1, \ldots, X_n$ is random, but the population $S = \{x_1, \ldots, x_N\}$ is not.

We can think of the Y_i as being measurements of one real-valued quantity on the individuals in the sample.

Sampling With and Without Replacement (cont.)

If X_1, ..., X_n are independent random variables, then we say we are *sampling with replacement*.

The picture is that we have written the names on the individuals on slips of paper, put all of the slips in a urn, mixed well, and drawn one which is X_1.

Then we draw another, but in order for the situation to be exactly the same as before, we need the same slips in the urn, well mixed, as before. Thus we put the slip with the name of X_1 back in the urn (replace it) and mix well. Then we draw another which is X_2. And so on for the rest.

Sampling With and Without Replacement (cont.)

Sampling with replacement is not the way real surveys are done. They are done *without replacement* which means the slips are not put back in the urn after draws.

In sampling with replacement, the same individual may appear multiple times in the sample. In sampling without replacement, the individuals in the sample are all different.

In sampling with replacement, X_1, ..., X_n are IID. In sampling without replacement, X_1, ..., X_n are exchangeable.

Sampling With and Without Replacement (cont.)

In sampling with replacement, the random vector $(X_1, \ldots, X_n)$ is uniformly distributed on the N^n possible assignments of values x_1, ..., x_N to variables X_1, ..., X_n.

In sampling without replacement, the random vector $(X_1, \ldots, X_n)$ is uniformly distributed on the $(N)_n$ ways to choose n things from N things if order matters or on the $\binom{N}{n}$ ways to choose n things from N things if order doesn't matter.

Sampling With and Without Replacement (cont.)

We are interested in the mean and variance of

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

In case $Y_1, \ldots, Y_n$ are independent, which is the case in sampling with replacement, we know

$$E\left(\sum_{i=1}^{n} Y_i \right) = \sum_{i=1}^{n} E(Y_i)$$

$$\text{var}\left(\sum_{i=1}^{n} Y_i \right) = \sum_{i=1}^{n} \text{var}(Y_i)$$

Sampling With and Without Replacement (cont.)

In case $Y_1, \ldots, Y_n$ are IID with mean μ and variance σ^2, which is the case in sampling with replacement, this becomes

$$E\left(\sum_{i=1}^{n} Y_i\right) = n\mu$$

$$\text{var}\left(\sum_{i=1}^{n} Y_i\right) = n\sigma^2$$

Hence

$$E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \mu$$

$$\text{var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{\sigma^2}{n}$$

Sampling With and Without Replacement (cont.)

Thus we have proved the following. If Y_1, ..., Y_n are IID with mean μ and variance σ^2, then

$$E(\overline{Y}_n) = \mu$$
$$\text{var}(\overline{Y}_n) = \frac{\sigma^2}{n}$$

In particular, this holds for sampling with replacement.

Sampling With and Without Replacement (cont.)

One of these formulas, $E(\overline{Y}_n) = \mu$, is the same for sampling without replacement because the expectation of a sum is the sum of the expectations regardless of whether the variables are independent or dependent.

The variance formula changes. Write $c = \text{cov}(Y_i, Y_j)$, $i \neq j$. Then

$$\text{var}\left(\sum_{i=1}^{n} Y_i\right) = n\sigma^2 + n(n-1)c$$

Sampling With and Without Replacement (cont.)

In the case of sampling without replacement, c is determined by the fact that in case $n = N$ where the sample is the population, there is no randomness: $\sum_i Y_i$ is just the sum over the population (the order of individuals in the sample does not matter in the sum).

Hence

$$0 = \text{var}\left(\sum_{i=1}^{N} Y_i\right) = N\sigma^2 + N(N-1)c$$

and

$$c = -\frac{\sigma^2}{N-1}$$

Sampling With and Without Replacement (cont.)

Plugging this value for c back into the general formula gives

$$\begin{aligned}
\text{var}\left(\sum_{i=1}^{n} Y_i\right) &= n\sigma^2 + n(n-1)c \\
&= n\sigma^2 - n(n-1)\cdot\frac{\sigma^2}{N-1} \\
&= n\sigma^2\left[1 - \frac{n-1}{N-1}\right] \\
&= n\sigma^2\cdot\frac{N-n}{N-1}
\end{aligned}$$

Sampling With and Without Replacement (cont.)

In sampling with replacement (and for IID)

$$E(\overline{Y}_n) = \mu$$
$$\text{var}(\overline{Y}_n) = \frac{\sigma^2}{n}$$

In sampling without replacement

$$E(\overline{Y}_n) = \mu$$
$$\text{var}(\overline{Y}_n) = \frac{\sigma^2}{n} \cdot \frac{N - n}{N - 1}$$

The factor $(N - n)/(N - 1)$ is negligible when N is much larger than n and is often ignored. The area of statistics that is careful about such things is called *finite population sampling*. The factor $(N - n)/(N - 1)$ is called the *finite population correction*.

The Hypergeometric Distribution

If Y_1, ..., Y_n are IID Bernoulli, then $Z = Y_1 + \cdots + Y_n$ is binomial. What is the analog for sampling without replacement?

More precisely, suppose we have a sample without replacement of size n from a finite population of size N, and we "measure" on each individual a zero-or-one-valued variable Y_i. What is the probability of observing x ones? This depends on how many ones are in the population, say r.

$$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, \qquad x \in \mathbb{N} \text{ and } 0 \le x \le r \text{ and } 0 \le n - x \le N - r$$

This is called the *hypergeometric distribution* with parameters N, n, and r. We won't use it enough to require an abbreviation.

The Hypergeometric Distribution Theorem

The fact that probabilities sum to one for the hypergeometric distribution gives us the following highly nonobvious theorem

$$\sum_{x=\max(0,n+r-N)}^{\min(r,n)} \binom{r}{x}\binom{N-r}{n-x} = \binom{N}{n}$$

The Hypergeometric Distribution (cont.)

Just as with the binomial distribution, the hypergeometric distribution is the distribution of a random variable $Z = Y_1 + \cdots + Y_n$ where the Y_i are identically distributed and Bernoulli, but not independent.

In this case, Y_1, ..., Y_n are exchangeable and arise from sampling without replacement. Thus we can use our formulas for the mean and variance of Z derived for general sampling without replacement.

$$E(Z) = nE(Y_1)$$
$$\text{var}(Z) = n\,\text{var}(Y_1) \cdot \frac{N-n}{N-1}$$

The Hypergeometric Distribution (cont.)

Since Y_1 is zero-or-one valued, it is Ber(p) for some p, and our formulas become

$$E(Z) = np$$
$$\text{var}(Z) = np(1-p) \cdot \frac{N-n}{N-1}$$

The only thing remaining to do is figure out that $p = r/N$, the fraction of ones in the population. This follows from the fact that X_1 is uniformly distributed on the population.

Time Series

"Time series" is just another name for a sequence $X_1, \ldots, X_n$ of random variables. The index, the i in X_i is called "time" whether it really is or not.

Examples, would be the price of a stock on consecutive days, the cholesterol level of a patient in consecutive lab tests, or the number of ears on consecutive corn plants along a row in a cornfield.

Note that in the last example, "time" is not time.

Time Series (cont.)

A fancier name for a time series is *stochastic process*.

Statistics books generally say "time series".

Probability books generally say "stochastic process".

Time Series (cont.)

Without some structure, we can't say anything about time series.

If a time series is an arbitrary bunch of random variables, then anything about it is arbitrary.

Stationary Time Series

A time series is *strictly stationary* if the distribution of a block of length k of consecutive variables

$$(X_{i+1}, X_{i+2}, \ldots, X_{i+k})$$

does not depend on i (every block of length k has the same distribution).

A time series is *weakly stationary* if

$$E(X_i)$$

and

$$\text{cov}(X_i, X_{i+k})$$

do not depend on i, and the latter holds for every $k \geq 0$.

Weakly Stationary Time Series

Every variable has the same mean: $E(X_i) = \mu$ for all i.

Every pair of variables separated by the same distance has the same covariance. Define

$$\gamma_k = \text{cov}(X_i, X_{i+k})$$

(the definition makes sense because the right-hand side does not depend on i). The function $\mathbb{N} \to \mathbb{R}$ defined by $k \mapsto \gamma_k$ is called the *autocovariance function* of the time series.

Weakly Stationary Time Series (cont.)

By the same argument as for IID and exchangeable random variables, we have $E(\overline{X}_n) = \mu$.

Using the autocovariance function

$$\begin{aligned} \text{var}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} \text{var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{cov}(X_i, X_j) \\ &= n\gamma_0 + 2 \sum_{k=1}^{n-1} (n-k)\gamma_k \end{aligned}$$

Weakly Stationary Time Series (cont.)

Summary: If $X_1, \ldots, X_n$ is a weakly stationary time series

$$E(\overline{X}_n) = \mu$$

$$\text{var}(\overline{X}_n) = \frac{1}{n}\left(\gamma_0 + 2\sum_{k=1}^{n-1}\frac{n-k}{n}\cdot\gamma_k\right)$$

AR(1) Time Series

A simple example of a weakly stationary time series is the *autoregressive order one*, AR(1) for short, time series.

Let Z_1, Z_2, ... be IID random variables with mean zero and variance τ^2. Let X_0 be any random variable having mean zero and independent of the Z_i, and recursively define

$$X_n = \rho X_{n-1} + Z_n, \qquad n = 1, 2, \ldots,$$

where ρ is a real number. The time series X_1, X_2, ... is AR(1).

AR(1) Time Series (cont.)

By linearity of expectation

$$E(X_n) = \rho E(X_{n-1}) + E(Z_n) = \rho E(X_{n-1})$$

Since $E(X_0) = 0$, we have $E(X_n) = 0$ for all n.

Since X_{n-1} is a function of X_0 and Z_1, ..., Z_{n-1}, which are independent of Z_n,

$$\begin{aligned} \mathrm{var}(X_n) &= \rho^2 \, \mathrm{var}(X_{n-1}) + \mathrm{var}(Z_n) \\ &= \rho^2 \, \mathrm{var}(X_{n-1}) + \tau^2 \end{aligned}$$

AR(1) Time Series (cont.)

In order for this time series to be weakly stationary, all of the X_n must have the same variance, say $\operatorname{var}(X_n) = \sigma^2$. Then

$$\sigma^2 = \rho^2\sigma^2 + \tau^2$$

which we solve for σ^2 obtaining

$$\sigma^2 = \frac{\tau^2}{1 - \rho^2}$$

Since a variance must be nonnegative, weak stationarity requires $-1 < \rho < 1$.

AR(1) Time Series (cont.)

For $k > 0$

$$\begin{aligned}
\mathrm{cov}(X_{n+k}, X_n) &= \mathrm{cov}(\rho X_{n+k-1} + Z_{n+k}, X_n) \\
&= \rho \, \mathrm{cov}(X_{n+k-1}, X_n) + \mathrm{cov}(Z_{n+k}, X_n) \\
&= \rho \, \mathrm{cov}(X_{n+k-1}, X_n)
\end{aligned}$$

because Z_{n+k} and X_n are independent random variables. Hence

$$\begin{aligned}
\mathrm{cov}(X_{n+1}, X_n) &= \rho \, \mathrm{cov}(X_n, X_n) = \rho \, \mathrm{var}(X_n) = \rho\sigma^2 \\
\mathrm{cov}(X_{n+2}, X_n) &= \rho \, \mathrm{cov}(X_{n+1}, X_n) = \rho^2\sigma^2 \\
\mathrm{cov}(X_{n+k}, X_n) &= \rho^k\sigma^2
\end{aligned}$$

AR(1) Time Series (cont.)

In summary, for a weakly stationary AR(1) time series

$$E(X_n) = 0$$

and

$$\text{cov}(X_{n+k}, X_n) = \rho^k \sigma^2$$

hold for all n and all $k \geq 0$, and the parameter ρ must satisfy $-1 < \rho < 1$.

Monotonicity of Expectation

Suppose $U \leq V$, which means $U(s) \leq V(s)$ for all s in the sample space.

Then we know $V - U \geq 0$, hence $E(V - U) \geq 0$. Since $E(V - U) = E(V) - E(U)$, we conclude $E(U) \leq E(V)$.

In summary,

$$U \leq V \quad \text{implies} \quad E(U) \leq E(V)$$

Markov's Inequality

For any nonnegative random variable Y and positive real number λ

$$\Pr(Y \geq \lambda) \leq \frac{E(Y)}{\lambda}$$

This is called *Markov's inequality*.

The proof is simple

$$\lambda I_{[\lambda,\infty)}(Y) \leq Y$$

always holds, hence

$$E(Y) \geq E\{\lambda I_{[\lambda,\infty)}(Y)\} = \lambda E\{I_{[\lambda,\infty)}(Y)\} = \lambda \Pr(Y \geq \lambda)$$

and rearranging this gives Markov's inequality.

Chebyshev's Inequality

The special case of Markov's inequality where $Y = (X - \mu)^2$ with $\mu = E(X)$ is called *Chebyshev's inequality*.

$$\Pr\{(X - \mu)^2 \geq \lambda\} \leq \frac{E\{(X - \mu)^2\}}{\lambda} = \frac{\text{var}(X)}{\lambda}$$

This is usually rewritten with absolute values rather than squares

$$\Pr(|X - \mu| \geq \delta) \leq \frac{\text{var}(X)}{\delta^2}$$

where $\delta = \sqrt{\lambda}$.

The Law of Large Numbers

Now replace X by

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where X_1, ..., X_n are identically distributed with mean μ, but not necessarily independent or even uncorrelated — they may be a stationary time series, for example.

We know $E(\overline{X}_n) = \mu$ by linearity of expectation. Chebyshev's inequality says

$$\Pr(|\overline{X}_n - \mu| \geq \delta) \leq \frac{\mathrm{var}(\overline{X}_n)}{\delta^2}$$

The Law of Large Numbers: Uncorrelated Case

Now let us specialize to the case where $X_1, \ldots, X_n$ are uncorrelated (which includes independent). Then

$$\text{var}(\overline{X}_n) = \frac{\sigma^2}{n}$$

where σ^2 is the variance of the X_i, which is the same for all because they are assumed identically distributed.

In this case, Chebyshev's inequality says

$$\Pr(|\overline{X}_n - \mu| \geq \delta) \leq \frac{\sigma^2}{n\delta^2}$$

Convergence in Probability to a Constant

A sequence of random variables Z_1, Z_2, ... *converges in probability* to the constant a if for every $\delta > 0$

$$\Pr(|Z_n - a| \geq \delta) \to 0, \qquad \text{as } n \to \infty.$$

Since we use this concept a lot, there is a shorthand for it

$$Z_n \xrightarrow{P} a$$

From our application of Chebyshev's inequality to $\overline{X}_n$ we see that

$$\overline{X}_n \xrightarrow{P} \mu$$

a statement known as the *law of large numbers*.

Convergence in Probability to a Constant (cont.)

The reason why we say convergence in probability *to a constant* is that there is a more general notion of convergence in probability *to a random variable*, which we will not define and will not use in this course.

Philosophy and the Law of Large Numbers

Now we can return to the frequentist philosophy of statistics.

Now we can see that what it tries to do is turn the law of large numbers into a *definition* of expectation.

In order to do that, it must somehow use the concept of *independence* or the concept of *uncorrelated* — both of which are defined in terms of expectation in conventional probability theory — without defining them, or at least without defining them in the conventional way.

No way of accomplishing this has ever been found that is not far more complicated than conventional probability theory.

Little Oh Pee Notation

The Chebyshev's inequality statement

$$\Pr(|\overline{X}_n - \mu| \geq \delta) \leq \frac{\sigma^2}{n\delta^2}$$

which we can rewrite as

$$\Pr(\sqrt{n}|\overline{X}_n - \mu| \geq \delta) \leq \frac{\sigma^2}{\delta^2}$$

is actually a lot stronger than the law of large numbers $\overline{X}_n \xrightarrow{P} \mu$.

To capture that strength of mathematical idea, we introduce the following concepts.

Little Oh Pee Notation (cont.)

For any sequence of random variables Z_1, Z_2, ... and any sequence of positive constants b_1, b_2, ..., we write

$$Z_n = o_p(b_n)$$

read "Z_n is little oh pee of b_n", to indicate

$$\frac{Z_n}{b_n} \xrightarrow{P} 0$$

Using this concept, we can say

$$\overline{X}_n - \mu = o_p(n^{-\alpha})$$

for any $\alpha < 1/2$.

Bounded in Probability

A sequence of random variables Z_1, Z_2, ... is *bounded in probability* if for every $\epsilon > 0$ there exists a $\lambda > 0$ such that

$$\Pr(|Z_n| \geq \lambda) \leq \epsilon, \qquad \text{for all } n.$$

Big Oh Pee Notation

For any sequence of random variables Z_1, Z_2, ... and any sequence of positive constants b_1, b_2, ..., we write

$$Z_n = O_p(b_n)$$

read "Z_n is big oh pee of b_n", to indicate that Z_n/b_n is bounded in probability

Using this concept, we can say

$$\overline{X}_n - \mu = O_p(n^{-1/2})$$

and this concisely and precisely encapsulates what Chebyshev's inequality tells us about the behavior of $\overline{X}_n$ when n is large.

The Square Root Law

A widely used textbook for 1001 calls this appearance of the square root of the sample size the "square root law"

> *statistical precision varies as the square root of the sample size*

Big Oh Pee And Little Oh Pee

$$Z_n = o_p(b_n) \quad \text{implies} \quad Z_n = O_p(b_n)$$

and, if $b_n/a_n \to 0$ as $n \to \infty$, then

$$Z_n = O_p(b_n) \quad \text{implies} \quad Z_n = o_p(a_n)$$

The Law of Large Numbers: AR(1) Time Series

Consider the AR(1) time series. We know

$$\text{cov}(X_{n+k}, X_n) = \rho^k \sigma^2$$

and

$$\begin{aligned}
\text{var}(\overline{X}_n) &= \frac{1}{n}\left(\gamma_0 + 2\sum_{k=1}^{n-1}\frac{n-k}{n}\cdot\gamma_k\right) \\
&= \frac{\sigma^2}{n}\left(1 + 2\sum_{k=1}^{n-1}\frac{n-k}{n}\cdot\rho^k\right) \\
&\leq \frac{\sigma^2}{n}\left(1 + 2\sum_{k=1}^{\infty}|\rho|^k\right)
\end{aligned}$$

The Geometric Series

If $-1 < s < 1$

$$\sum_{k=0}^{\infty} s^k = \frac{1}{1-s}$$

First write

$$S_n = \sum_{k=0}^{n} s^k$$

The Geometric Series (cont.)

Then

$$(1-s)S_n = \sum_{k=0}^{n} s^k - \sum_{k=1}^{n+1} s^k = 1 - s^{n+1}$$

so

$$S_n = \frac{1-s^{n+1}}{1-s}$$

If $|s| < 1$, then $s^{n+1} \to 0$ as $n \to \infty$.

That proves the formula on the previous slide.

The Law of Large Numbers: AR(1) Time Series (cont.)

Continuing the calculation for the AR(1) time series

$$\begin{aligned}
\text{var}(\overline{X}_n) &\leq \frac{\sigma^2}{n}\left(1 + 2\sum_{k=1}^{\infty}|\rho|^k\right) \\
&= \frac{\sigma^2}{n}\left(-1 + 2\sum_{k=0}^{\infty}|\rho|^k\right) \\
&= \frac{\sigma^2}{n}\cdot\frac{1+|\rho|}{1-|\rho|}
\end{aligned}$$

We see again that $\text{var}(\overline{X}_n)$ is bounded by a constant divided by n so again

$$\overline{X}_n - \mu = O_p(n^{-1/2})$$

The Law of Large Numbers: Exchangeable???

For an exchangeable sequence X_1, X_2, ... with $\text{var}(X_i) = v$ and $\text{cov}(X_i, X_j) = c$ we have

$$\text{var}(\overline{X}_n) = \frac{1}{n^2}\left(nv + n(n-1)c\right) \leq \begin{cases} v/n + c, & c > 0 \\ v/n, & c \leq 0 \end{cases}$$

so when $c > 0$ all we can say is that

$$\overline{X}_n - \mu = O_p(1)$$

and we can't say that $\overline{X}_n$ converges in probability to μ or anything else.

Since every exchangeable sequence is a strictly stationary time series, we see that the law of large numbers does not hold for every stationary time series, but it does hold for some, such as the AR(1) time series.

The Law of Large Numbers (cont.)

Thus we see that uncorrelated is a *sufficient* but not *necessary* condition for the law of large numbers to hold.

Identically distributed is also not necessary. Consider X_1, X_2, ... all of which have mean μ and are uncorrelated but $\mathrm{var}(X_n) = \sigma^2$ when n is odd and $\mathrm{var}(X_n) = \tau^2$ when n is even. Then

$$\mathrm{var}(\overline{X}_n) \le \frac{\max(\sigma^2, \tau^2)}{n}$$

and

$$\overline{X}_n - \mu = O_p(n^{-1/2})$$

so the square root law holds for this too.

The Law of Large Numbers (cont.)

In summary, if X_1, X_2, ... all have the same mean μ, then

$$\overline{X}_n - \mu = O_p(n^{-1/2})$$

holds

- always if the X_i are identically distributed and uncorrelated,
- sometimes if the X_i form a weakly stationary time series, and
- sometimes even if the X_i are not identically distributed.

It all depends on the size of $\text{var}(\overline{X}_n)$.

Axioms for Probabilities

So far we have mostly ignored probabilities in this slide deck except that since probability is a special case of expectation

$$\Pr(A) = E(I_A)$$

all our theory about expectation has implications for probabilities. Now we see what these are.

Axioms for Probabilities (cont.)

What does

$$E(X+Y) = E(X) + E(Y)$$

say about probabilities?

To answer that we need to examine when is $I_A + I_B$ an indicator function.

Union and Intersection

For any sets A and B

$$A \cup B = \{ x : x \in A \text{ or } x \in B \}$$
$$A \cap B = \{ x : x \in A \text{ and } x \in B \}$$

$A \cup B$ is called the *union* of A and B.
$A \cap B$ is called the *intersection* of A and B.

In the definition of union "or" means "one or the other or *both*" which is the standard meaning in mathematical logic. So

$$A \cap B \subset A \cup B$$

Union and Intersection (cont.)

If $\mathcal{A}$ is any nonempty family of sets, then we write

$$\bigcup \mathcal{A} = \{\, x : x \in A \text{ for some } A \in \mathcal{A} \,\}$$
$$\bigcap \mathcal{A} = \{\, x : x \in A \text{ for all } A \in \mathcal{A} \,\}$$

If $\mathcal{A} = \{\, A_i : i \in I \,\}$, then

$$\bigcup \mathcal{A} = \bigcup_{i \in I} A_i$$
$$\bigcap \mathcal{A} = \bigcap_{i \in I} A_i$$

Axioms for Probabilities (cont.)

$$I_A(x) + I_B(x) = \begin{cases} 0, & x \notin A \cup B \\ 1, & x \in (A \cup B) \setminus (A \cap B) \\ 2, & x \in A \cap B \end{cases}$$

Hence $I_A + I_B$ is not an indicator function unless $A \cap B = \varnothing$.

So this leads to a definition. Sets A and B are *disjoint* (also called *mutually exclusive*) if $A \cap B = \varnothing$.

When A and B are mutually exclusive, we have

$$I_{A \cup B} = I_A + I_B$$
$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$

Axioms for Probabilities (cont.)

Unfortunately, this simple "addition rule for probabilities" is not very useful, because the required condition — disjoint events — does not arise often.

In general we can write

$$I_A + I_B = I_{A \cup B} + I_{A \cap B}$$
$$\Pr(A) + \Pr(B) = \Pr(A \cup B) + \Pr(A \cap B)$$

Subadditivity of Probability

We can rewrite the last equation as

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

and infer from it

$$\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$$

and then by mathematical induction

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \Pr(A_i)$$

This is called *subadditivity of probability*. Unlike the "addition rule" it holds whether or not events are disjoint. We use this often.

The Inclusion-Exclusion Rule

We can apply mathematical induction to

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

obtaining

$$\begin{aligned}\Pr(A \cup B \cup C) &= \Pr(A) + \Pr(B \cup C) - \Pr(A \cap (B \cup C)) \\ &= \Pr(A) + \Pr(B \cup C) - \Pr((A \cap B) \cup (A \cap C)) \\ &= \Pr(A) + \Pr(B) + \Pr(C) - \Pr(B \cap C) \\ &\quad - \Pr(A \cap B) - \Pr(A \cap C) + \Pr(A \cap B \cap C)\end{aligned}$$

and increasingly messier formulas for more events.

Axioms for Probabilities (cont.)

The axiom $X \geq 0$ implies $E(X) \geq 0$ says the following about probabilities: since any indicator function is nonnegative, we have $\Pr(A) \geq 0$ for any event A.

Of course, we knew that already, but if we want axioms for probability, that is one of them.

The axiom $E(aX) = aE(X)$ doesn't say anything about probabilities, because multiplying an indicator function by a constant doesn't give an indicator function.

The axiom $E(1) = 1$, says $\Pr(S) = 1$, where S is the sample space, because the indicator function of the whole sample space is equal to one everywhere.

Summary of Axioms for Probabilities

$$\Pr(A \cup B) = \Pr(A) + \Pr(B), \qquad \text{if } A \text{ and } B \text{ are disjoint events}$$

$$\Pr(A) \geq 0, \qquad \text{for any event } A$$

$$\Pr(S) = 1, \qquad \text{where } S \text{ is the sample space}$$

Complementary Events and the Complement Rule

Definition:

$$A^c = S \setminus A$$

is called the *complement* of the event A. Note that A and A^c are disjoint events so

$$\Pr(A) + \Pr(A^c) = \Pr(A \cup A^c) = \Pr(S) = 1$$

hence

$$\Pr(A^c) = 1 - \Pr(A)$$

which is called the *complement rule*.

Complementary Events and the Complement Rule (cont.)

In particular, the complement rule implies $\Pr(A) \leq 1$ so

$$0 \leq \Pr(A) \leq 1, \qquad \text{for any event } A$$

follows from the axioms (we already knew it follows from the definition of probabilities of events in terms of PMF).

Independence and Probabilities

Random variables $X_1, \ldots, X_n$ are independent if

$$E\left\{\prod_{i=1}^{n} h_i(X_i)\right\} = \prod_{i=1}^{n} E\{h_i(X_i)\}$$

for any functions $h_1, \ldots, h_n$ (this just repeats what was said on slide 76). What does this say about probabilities?

Independence and Probabilities (cont.)

The case $h_i = I_{A_i}$ of

$$E\left\{\prod_{i=1}^{n} h_i(X_i)\right\} = \prod_{i=1}^{n} E\{h_i(X_i)\}$$

can be rewritten

$$\Pr(X_1 \in A_1 \text{ and } \cdots \text{ and } X_n \in A_n) = \prod_{i=1}^{n} \Pr(X_i \in A_i)$$

This holds for any events A_1, ..., A_n but only when X_1, ..., X_n are independent random variables.

Independence and Probabilities (cont.)

Definition: the events A_1, ..., A_n are *independent* if the random variables I_{A_1}, ..., I_{A_n} are independent.

The case $X_i = I_{A_i}$ and each h_i is the identity function of

$$E\left\{\prod_{i=1}^{n} h_i(X_i)\right\} = \prod_{i=1}^{n} E\{h_i(X_i)\}$$

can be rewritten

$$\Pr\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \Pr(A_i)$$

This holds for independent events A_1, ..., A_n.

Monotonicity of Probability

As a generalization of the complement rule, if A and B are events and $A \subset B$, then A and $B \setminus A$ are mutually exclusive events and $B = A \cup (B \setminus A)$, hence

$$\Pr(B) = \Pr(A) + \Pr(B \setminus A)$$

(the complement rule is the case $B = S$). From this we conclude

$$\Pr(A) \leq \Pr(B), \qquad \text{whenever } A \subset B$$

which is called *monotonicity of probability*.