Quantitative Genomics and Genetics

BioCB 4830/6830; PBSB.5201.03

Lecture 6: Variances and Covariances of Random Vectors & Prob Models

Jason Mezey

Feb 8, 2024 (Th) 8:40-9:55

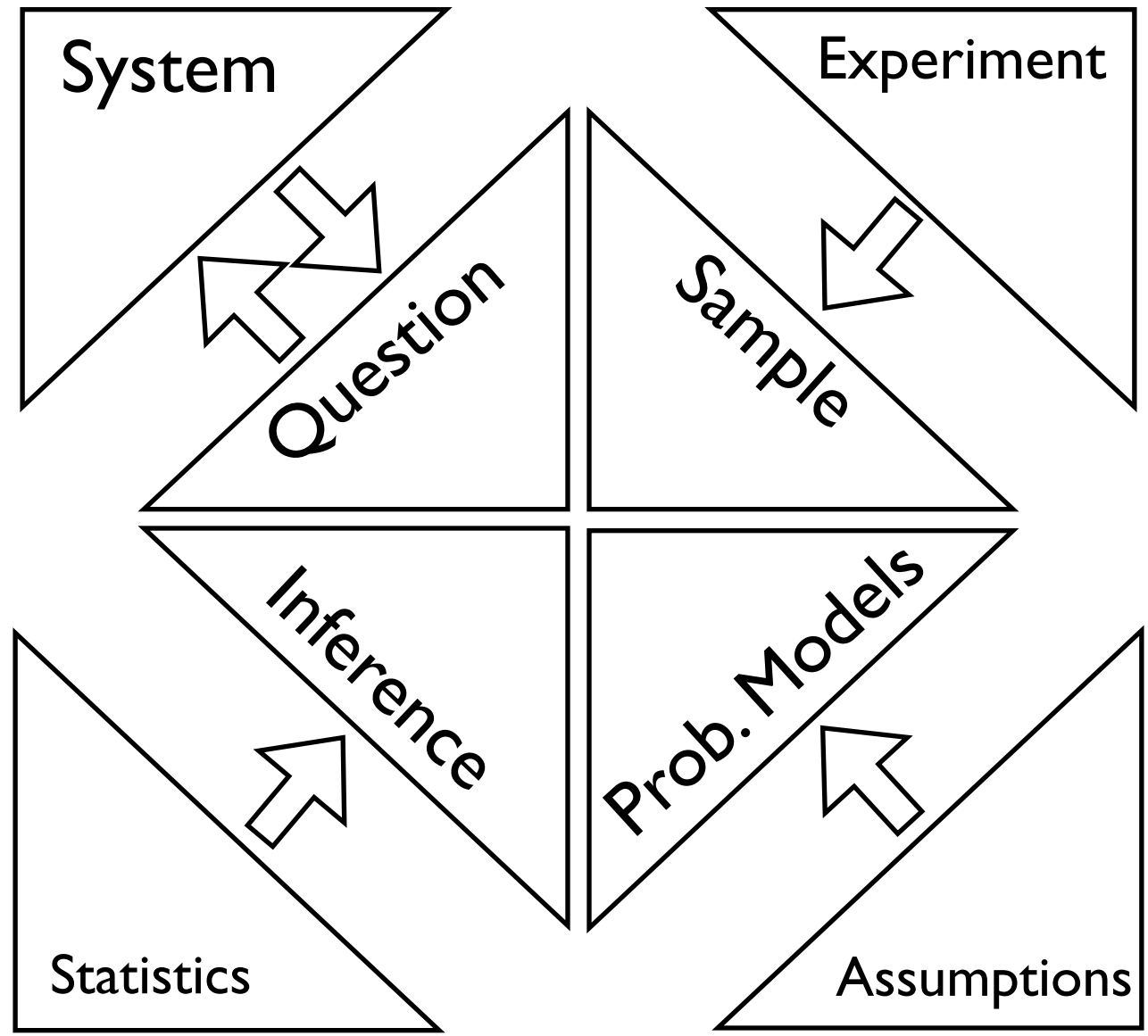
Announcements

- Registering for labs in Ithaca - everyone should be registered for EITHER Thurs or Fri lab, if not - PLEASE CANVAS EMAIL ME ASAP
- For Weill (NYC) students - we have lecture classrooms for next week (and today) Feb 13 and Feb 15: A-250 (1300 York Ave, 2nd floor)
- Reminder: 1st homework is due tomorrow (Fri, Feb 9) by 11:59PM (!!)

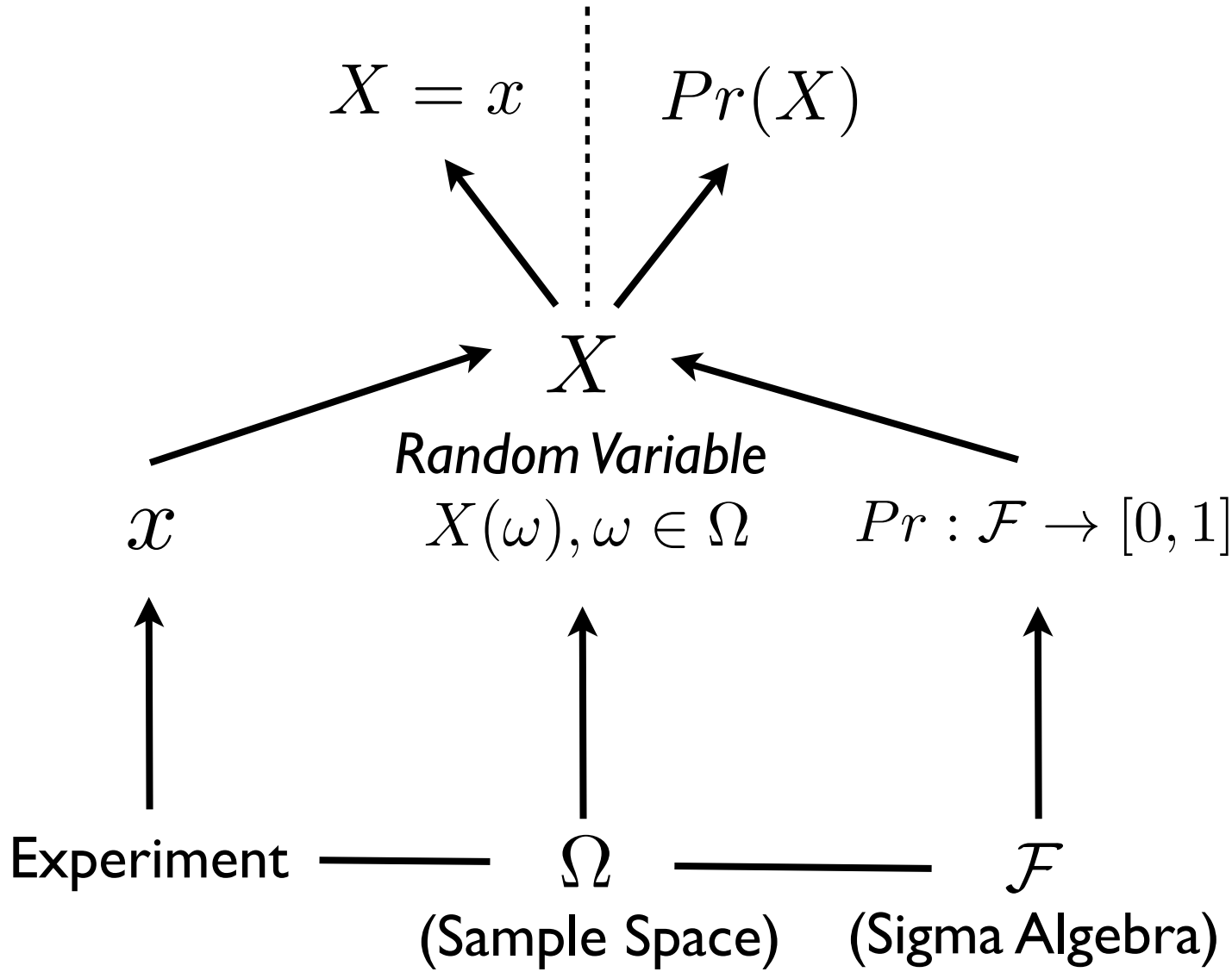
Summary of lecture 6: Introduction expectations / variances of random AND Intro to Inference

- Last class, we introduced expectations
- Today we will continue our discussion of expectations and variances of random variables and vectors!
- We will also begin our discussion of probability models

Conceptual Overview



Review: Random Variables



Review: Random vectors

- We are often in situations where we are interested in defining more than one r.v. on the same sample space
- When we do this, we define a **random vector**
- Note that a vector, in its simplest form, may be considered a set of numbers (e.g. [1.2, 2.0, 3.3] is a vector with three elements)
- Also note that vectors (when a vector space is defined) ARE NOT REALLY NUMBERS although we can define operations for them (e.g. addition, "multiplication"), which we will use later in this course
- Beyond keeping track of multiple r.v.'s, a *random vector* works just like a r.v., i.e. a probability function induces a probability function on the random vector and we may consider discrete or continuous (or mixed!) random vectors
- Note that we can define several r.v.'s on the same sample space (= a random vector), but this will result in one probability distribution function (why!?)

Review: Example of a discrete random vector

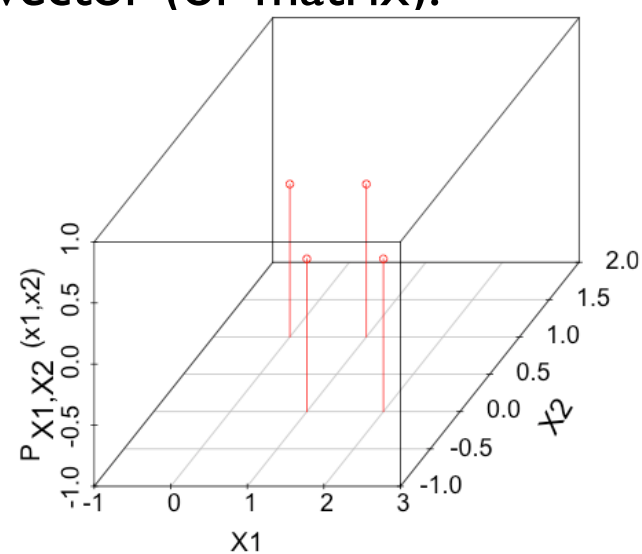
- Consider the two coin flip experiment and assume a probability function for a fair coin: $Pr(\{HH\}) = Pr(\{HT\}) = Pr(\{TH\}) = Pr(\{TT\}) = 0.25$
- Let's define two random variables: "number of Tails" and "first flip is Heads"

$$X_1 = \begin{cases} X_1(HH) = 0 \\ X_1(HT) = X_1(TH) = 1 \\ X_1(TT) = 2 \end{cases} \qquad X_2 = \begin{cases} X_2(TH) = X_2(TT) = 0 \\ X_2(HH) = X_2(HT) = 1 \end{cases}$$

- The probability function induces the following pmf for the random vector $\mathbf{X}=[X_1, X_2]$, where we use bold $\mathbf{X}$ do indicate a vector (or matrix):

$$Pr(\mathbf{X}) = Pr(X_1 = x_1, X_2 = x_2) = P_{\mathbf{X}}(\mathbf{x}) = P_{X_1,X_2}(x_1, x_2)$$

$$Pr(X_1 = 0, X_2 = 0) = 0.0, Pr(X_1 = 0, X_2 = 1) = 0.25$$
$$Pr(X_1 = 1, X_2 = 0) = 0.25, Pr(X_1 = 1, X_2 = 1) = 0.25$$
$$Pr(X_1 = 2, X_2 = 0) = 0.25, Pr(X_1 = 2, X_2 = 1) = 0.0$$



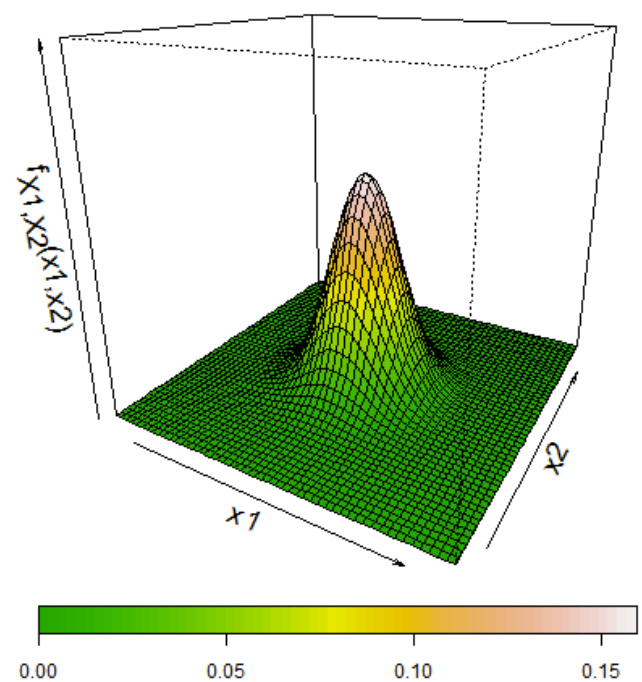
Review: Example of a continuous random vector

- Consider an experiment where we define a two-dimensional *Reals* sample space for "height" and "IQ" for every individual in the US (as a reasonable approximation)
- Let's define a bivariate normal probability function for this sample space and random variables X_1 and X_2 that are identity functions for each of the two dimensions
- In this case, the pdf of $\mathbf{X}=[X_1, X_2]$ is a bivariate normal (we will not write out the formula for this distribution - yet):

$$Pr(\mathbf{X}) = Pr(X_1 = x_1, X_2 = x_2) = f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,X_2}(x_1, x_2)$$

Again, note that we cannot use this probability function to define the probabilities of points (or lines!) but we can use it to define the probabilities that values of the random vector fall within (square) intervals of the two random variables (!) [a,b], [c,d]

$$Pr(a \leqslant X_1 \leqslant b, c \leqslant X_1 \leqslant d) = \int_a^b \int_c^d f_{X_1,X_2}(x_1, x_2) dx_1, dx_2$$



Review: Random vector conditional probability and independence I

- Just as we have defined *conditional probability* (which are probabilities!) for sample spaces, we can define conditional probability for random vectors:

$$Pr(X_1|X_2) = \frac{Pr(X_1 \cap X_2)}{Pr(X_2)}$$

- As a simple example (discrete in this case - but continuous is analogous!), consider the two flip sample space, fair coin probability model, random variables: "number of tails" and "first flip is heads":

	$X_2 = 0$	$X_2 = 1$	
$X_1 = 0$	0.0	0.25	0.25
$X_1 = 1$	0.25	0.25	0.5
$X_1 = 2$	0.25	0.0	0.25
	0.5	0.5	

$$Pr(X_1 = 0|X_2 = 1) = \frac{Pr(X_1 = 0 \cap X_2 = 1)}{Pr(X_2 = 1)} = \frac{0.25}{0.5} = 0.5$$

- We can similarly consider whether r.v.'s of a random vector are independent, e.g.

$$Pr(X_1 = 0 \cap X_2 = 1) = 0.25 \neq Pr(X_1 = 0)Pr(X_2 = 1) = 0.25 * 0.5 = 0.125$$

- NOTE I: we can use either $Pr(X_i|X_j) = Pr(X_i)$ or $Pr(X_i \cap X_j) = Pr(X_i)Pr(X_j)$ to check independence!

- NOTE II: to establish Xi, Xj are independent you must check all possible relationships but the opposite is not true: if one does not show independence you've established they are not independent (!!)

Review: Random vectors conditional probability and independence II

For random variables that are NOT independent…

	$X_2 = 0$	$X_2 = 1$	
$X_1 = 0$	0.0	0.25	0.25
$X_1 = 1$	0.25	0.25	0.5
$X_1 = 2$	0.25	0.0	0.25
	0.5	0.5	

To establish non-independence, just show ONE case that does not conform to the independence definition (e.g.,):

$$Pr(X_2 = 0|X_1 = 0) = \frac{Pr(X_2 = 0 \cap X_1 = 0)}{Pr(X_1 = 0)} = 0 \neq Pr(X_2 = 0) = 0.5$$

OR

$$Pr(X_2 = 0 \cap X_1 = 0) = 0 \neq Pr(X_2 = 0)Pr(X_1 = 0) = 0.5 * 0.25 = 0.125$$

And you're done!

For random variables that ARE independent…

	$X_2 = 0$	$X_2 = 1$	
$X_1 = 0$	0.125	0.125	0.25
$X_1 = 1$	0.25	0.25	0.5
$X_1 = 2$	0.125	0.125	0.25
	0.5	0.5	

To establish independence, you need to show ALL combinations of random variable states conform to the independence definition (!!):

$$Pr(X_2 = 0 \cap X_1 = 0) = Pr(X_2 = 0)Pr(X_1 = 0) = 0.5 * 0.25 = 0.125$$
$$Pr(X_2 = 0 \cap X_1 = 1) = Pr(X_2 = 0)Pr(X_1 = 1) = 0.5 * 0.5 = 0.25$$
$$Pr(X_2 = 0 \cap X_1 = 2) = Pr(X_2 = 0)Pr(X_1 = 2) = 0.5 * 0.25 = 0.125$$
$$Pr(X_2 = 1 \cap X_1 = 0) = Pr(X_2 = 1)Pr(X_1 = 0) = 0.5 * 0.25 = 0.125$$
$$Pr(X_2 = 1 \cap X_1 = 1) = Pr(X_2 = 1)Pr(X_1 = 1) = 0.5 * 0.5 = 0.25$$
$$Pr(X_2 = 1 \cap X_1 = 2) = Pr(X_2 = 1)Pr(X_1 = 2) = 0.5 * 0.25 = 0.125$$

Review: Expectations and variances

- We are now going to introduce fundamental functions of random variables / vectors: **expectations** and **variances**
- These are **functionals -** map a function to a scalar (number)
- These intuitively (but not rigorously!) these may be thought of as "a function on a function" with the following form:

$$f(\mathbf{X}(\Omega), Pr(\mathbf{X})) : \{\mathbf{X}, Pr(\mathbf{X})\} \rightarrow \mathbb{R}$$

- These are critical concepts for understanding the structure of probability models where the interpretation of the specific probability model under consideration
- They also have deep connections to many important concepts in probability and statistics
- Note that these are distinct from functions (*Transformations*) that are defined directly on *X* and not on Pr(*X*), i.e. *Y* = g(*X*):

$$g(\mathbf{X}) : X \rightarrow Y$$

$$g(\mathbf{X}) \rightarrow Y \Rightarrow Pr(X) \rightarrow Pr(Y)$$

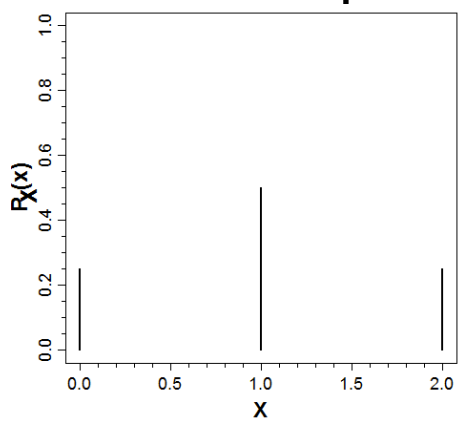
Review: Expectations I

- Following our analogous treatment of concepts for *discrete* and *continuous* random variables, we will do the same for *expectations* (and variances), which we also call *expected values*
- Note that the interpretation of the expected value is the same in each case
- The expected value of a discrete random variable is defined as follows:

$$\mathrm{E}X = \sum_{i=min(X)}^{max(X)} (X = i) Pr(X = i)$$

- For example, consider our two-coin flip experiment / fair coin probability model / random variable "number of tails":

$$\mathrm{E}X = (0)(0.25) + (1)(0.5) + (2)(0.25) = 1$$

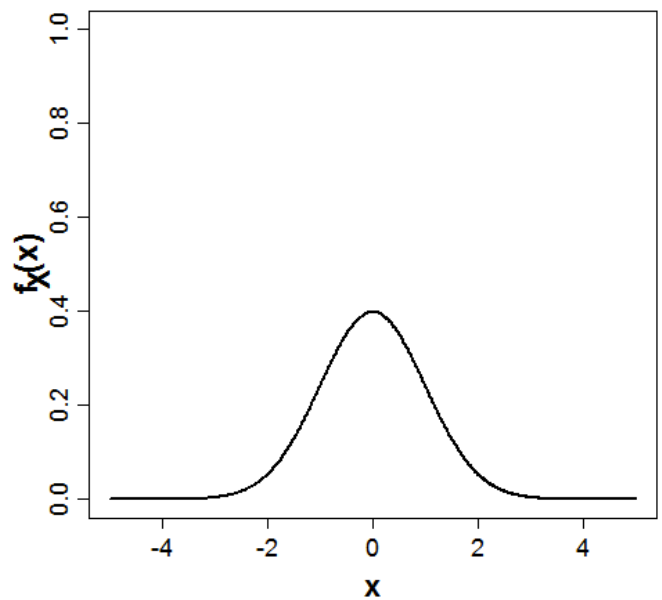


Review: Expectations II

- The expected value of a continuous random variable is defined as follows:

$$\mathrm{E}X = \int_{-\infty}^{+\infty} X f_X(x) dx$$

- For example, consider our height measurement experiment / normal probability model / identity random variable:



Expectations III

- In the discrete case, this is the *same* as adding up all the possibilities that can occur and dividing by the total number, e.g. (0+1+1+2) / 4 = 1 (hence it is often referred to as the *mean* of the random variable
- An expected value may be thought of as the "center of gravity", where a median (defined as the number where half of the probability is on either side) is the "middle" of the distribution (note that for symmetric distributions, these two are the same!)
- The expectation of a random variable X is the value of X that minimizes the sum of the squared distance to each possibility
- For some distributions, the expectation of the random variable may be infinite. In such cases, the expectation does not exist

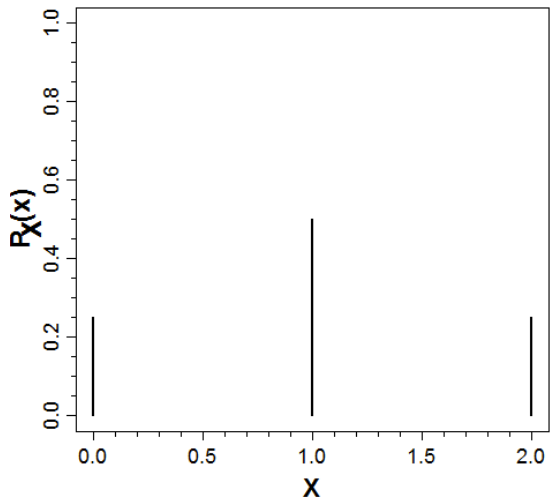
Variances I

- We will define *variances* for *discrete* and *continuous* random variables, where again, the interpretation for both is the same
- The variance of a discrete random variable is defined as follows:

$$\mathrm{Var}(X) = \mathrm{V}(X) = \sum_{i=min(X)}^{max(X)} ((X = i) - \mathrm{E}X)^2 Pr(X = i)$$

- For example, consider our two-coin flip experiment / fair coin probability model / random variable "number of tails":

$$Var(X) = (0-1)^2(0.25) + (1-1)^2(0.5) + (2-1)^2(0.25) = 0.5$$

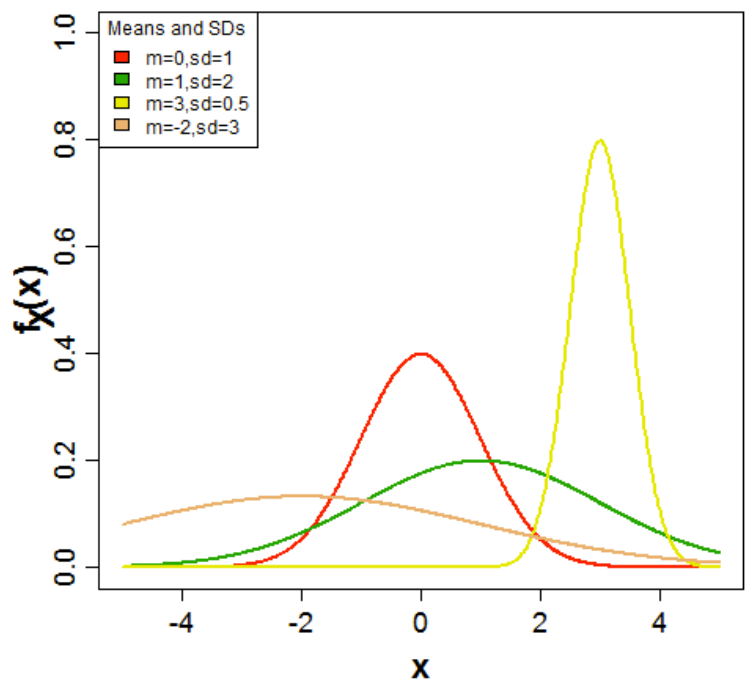


Variances II

- The variance of a continuous random variable is defined as follows:

$$\mathrm{Var}(X) = \mathrm{V}X = \int_{-\infty}^{+\infty} (X - \mathrm{E}X)^2 f_X(x) dx$$

- For example, consider our height measurement experiment / normal probability model / identity random variable:



Variances III

- Intuitively, the variance quantifies the "spread" of a distribution
- The squared component of variance has convenient mathematical properties, e.g. we can view them as sides of triangles
- Other equivalent (and often used) formulations of variance:

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}X)^2\right]$$
$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}X)^2$$

- Instead of viewing variance as including a squared term, we could view the relationship as follows:

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}X)(X - \mathrm{E}X)]$$

Generalization: higher moments

- The expectation of a random variable is the "first" moment and we can generalize this concept to "higher" moments:

$$\mathrm{E}X^k = \sum X^k Pr(X)$$

$$\mathrm{E}X^k = \int X^k f_X(x) dx$$

- The variance is the second "central" moment (i.e. calculating a moment after subtracting off the mean) and we can generalize this concept to higher moments as well:

$$\mathrm{C}(X^k) = \sum (X - \mathrm{E}X)^k Pr(X)$$

$$\mathrm{C}(X^k) = \int (X - \mathrm{E}X)^k f_X(x) dx$$

Random vectors: expectations and variances

- Recall that a generalization of a random variable is a random vector, e.g.

$$\mathbf{X} = [X_1, X_2] \qquad P_{X_1,X_2}(x_1, x_2) \text{ or } f_{X_1,X_2}(x_1, x_2)$$

- The expectation (a function of a random vector and its distribution!) is a vector with the expected value of each element of the random vector, e.g.

$$\mathrm{E}\mathbf{X} = [\mathrm{E}X_1, \mathrm{E}X_2]$$

- Variances also result in variances of each element (and additional terms... see next slide!!)

- Note that we can determine the conditional expected value or variance of a random variable conditional on a value of another variable, e.g.

$$\mathrm{E}(X_1|X_2) = \sum_{i=min(X_1)}^{max(X_1)} (X_1 = i) Pr(X_i = i|X_2) \qquad \mathrm{V}(X_1|X_2) = \sum_{i=min(X_1)}^{max(X_1)} ((X_1 = i) - \mathrm{E}X_1)^2 Pr(X_i = i|X_2)$$

$$\mathrm{E}(X_1|X_2) = \int_{-\infty}^{+\infty} X_1 f_{X_1|X_2}(x_1|x_2) dx_1 \qquad \mathrm{V}(X_1|X_2) = \int_{-\infty}^{+\infty} (X_1 - \mathrm{E}X_1)^2 f_{X_1|X_2}(x_1|x_2) dx_1$$

Random vectors: covariances

- Variances (again a function!) of a random vector are similar producing variances for each element, but they also produce **covariances**, which relate the relationships between random variables *of a random vector*!!

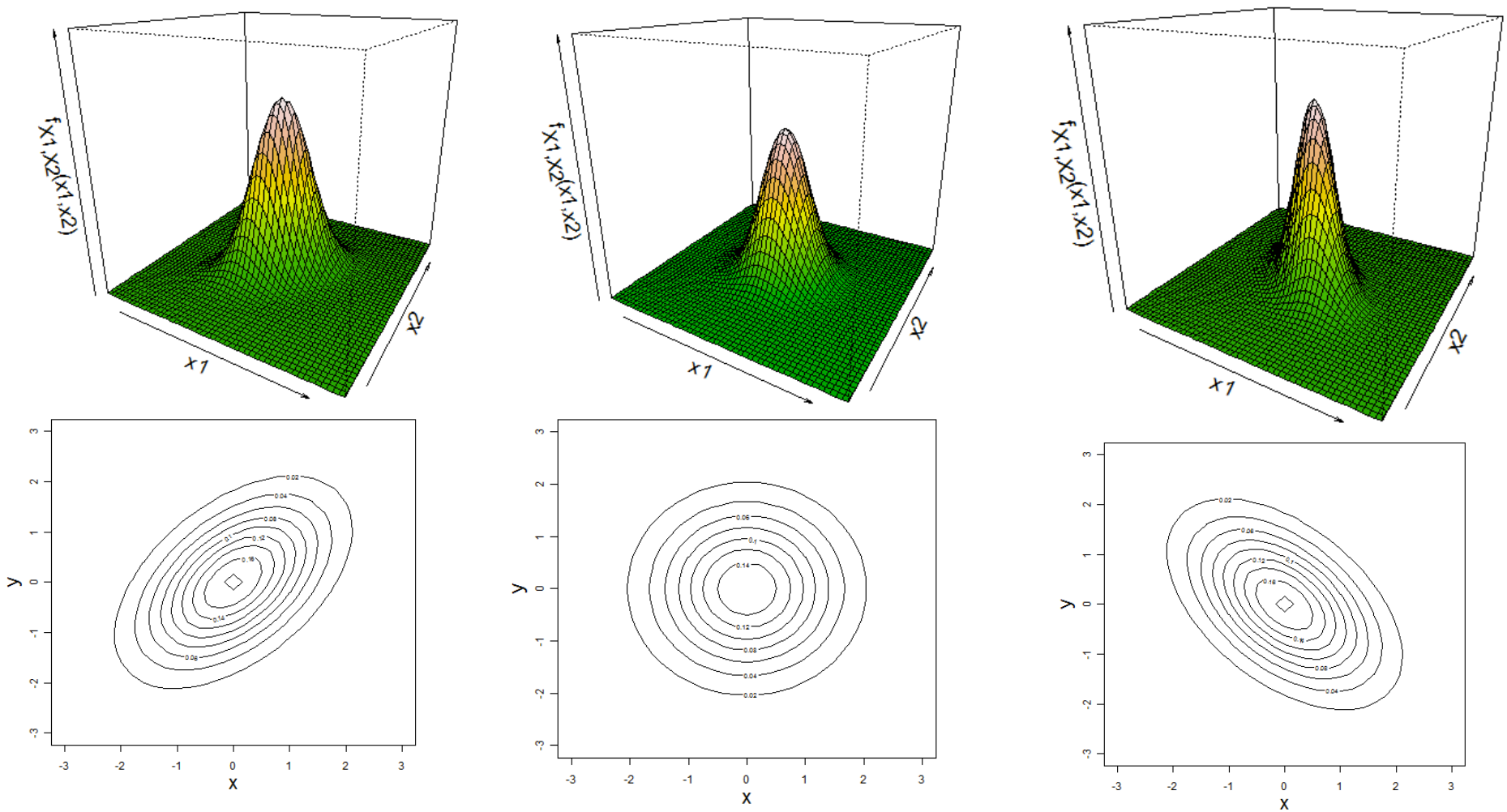
$$Cov(X_1, X_2) = \sum_{i=min(X_1)}^{i=max(X_1)} \sum_{j=min(X_2)}^{j=max(X_2)} ((X_1 = i) - \mathrm{E}X_1)((X_2 = j) - \mathrm{E}X_2) P_{X_1,X_2}(x_1, x_2)$$

$$\mathrm{Cov}(X_1, X_2) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (X_1 - \mathrm{E}X_1)(X_2 - \mathrm{E}X_2) f_{X_1,X_2}(x_1, x_2) dx_1 dx_2$$

- Intuitively, we can interpret a positive covariance as indicating "big values of X_1 tend to occur with big values of X_2 AND small values of X_1 tend to occur with small values of X_2"
- Negative covariance is the opposite (e.g. "big X_1 with small X_2 AND small X_1 with big X_2")
- Zero covariance indicates no relationship between big and small values of X_1 and X_2

An illustrative example

- For example, consider our experiment where we have measured "height" and "IQ" / bivariate normal probability model / identity random variable:



Notes about covariances

- Covariance and independence, while related, are NOT synonymous (!!), although if random variables are independent, then their covariance is zero (but necessarily vice versa!)
- Covariances are symmetric: $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$
- Other equivalent (and often used) formulations of covariances:

$$\mathrm{Cov}(X_1, X_2) = \mathrm{E}\left[(X_1 - \mathrm{E}X_1)(X_2 - \mathrm{E}X_2)\right]$$

$$\mathrm{Cov}(X_1, X_2) = \mathrm{E}(X_1 X_2) - \mathrm{E}X_1 \mathrm{E}X_2$$

- From these formulas, it follows that the covariance of a random variable and itself is the variance:

$$\mathrm{Cov}(X_1, X_1) = \mathrm{E}(X_1 X_1) - \mathrm{E}X_1 \mathrm{E}X_1 = \mathrm{E}(X_1^2) - (\mathrm{E}X_1)^2 = \mathrm{Var}(X_1)$$

Covariance matrices

- Note that we have defined the "output" of applying an expectation function to a random vector but we have not yet defined the analogous output for applying a variance function to a random vector
- The output is a covariance matrix, which is square, symmetric matrix with variances on the diagonal and covariances on the off-diagonals
- For example, for two and three random variables:

$$\mathrm{Var}(\mathbf{X}) = \begin{bmatrix} \mathrm{Var}X_1 & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}X_2 \end{bmatrix}$$

$$\mathrm{Var}(\mathbf{X}) = \begin{bmatrix} \mathrm{Var}X_1 & \mathrm{Cov}(X_1, X_2) & \mathrm{Cov}(X_1, X_3) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}X_2 & \mathrm{Cov}(X_2, X_3) \\ \mathrm{Cov}(X_1, X_3) & \mathrm{Cov}(X_2, X_3) & \mathrm{Var}(X_3) \end{bmatrix}$$

- Note that not all square, symmetric matrices are covariance matrices (!!), technically they must be positive (semi)-definite to be a covariance matrix

Covariances and correlations

- Since the magnitude of covariances depends on the variances of $X1$ and $X2$, we often would like to scale these such that "1" indicates maximum "big with big / small with small" and "-1" indicates maximum "big with small" (and zero still indicates no relationship)
- A correlation captures this relationship:

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)}\sqrt{\mathrm{Var}(X_2)}}$$

- Where we can similarly calculate a correlation matrix, e.g. for three random variables:

$$\mathrm{Corr}(\mathbf{X}) = \begin{bmatrix} 1 & \mathrm{Corr}(X_1, X_2) & \mathrm{Corr}(X_1, X_3) \\ \mathrm{Corr}(X_1, X_2) & 1 & \mathrm{Corr}(X_2, X_3) \\ \mathrm{Corr}(X_1, X_3) & \mathrm{Corr}(X_2, X_3) & 1 \end{bmatrix}$$

Algebra of expectations and variances

- If we consider a function (e.g., a transformation) on X (a function on the random variable but not on the probabilities directly!), recall that this can result in a different probability distribution for Y and therefore different expectations, variances, etc. for Y as well
- We will consider two types of functions on random variables and the result on expectation and variances: sums $Y = X_1 + X_2 + \ldots$ and $Y = a + bX_1$ where a and b are constants
- For example, for sums, $Y = X_1 + X_2$ we have the following relationships:

$$\mathrm{E}(Y) = \mathrm{E}(X_1 + X_2) = \mathrm{E}X_1 + \mathrm{E}X_2$$

$$\mathrm{Var}(Y) = \mathrm{Var}(X_1 + X_2) = \mathrm{Var}X_1 + \mathrm{Var}X_2 + 2\mathrm{Cov}(X_1, X_2)$$

- As another example, for $Y = X_1 + X_2 + X_3$ we have:

$$\mathrm{E}(Y) = \mathrm{E}(X_1 + X_2 + X_3) = \mathrm{E}X_1 + \mathrm{E}X_2 + \mathrm{E}X_3$$

$$\mathrm{Var}(Y) = \mathrm{Var}(X_1+X_2+X_3) = \mathrm{Var}X_1+\mathrm{Var}X_2+\mathrm{Var}X_3+2\mathrm{Cov}(X_1, X_2)+2\mathrm{Cov}(X_1, X_3)+2\mathrm{Cov}(X_2, X_3)$$

Algebra of expectations and variances

- For the function $Y = a + bX_1$ we obtain the following relationships:

$$\mathrm{E}Y = a + b\mathrm{E}X$$

$$\mathrm{Var}(Y) = b^2\mathrm{Var}(X)$$

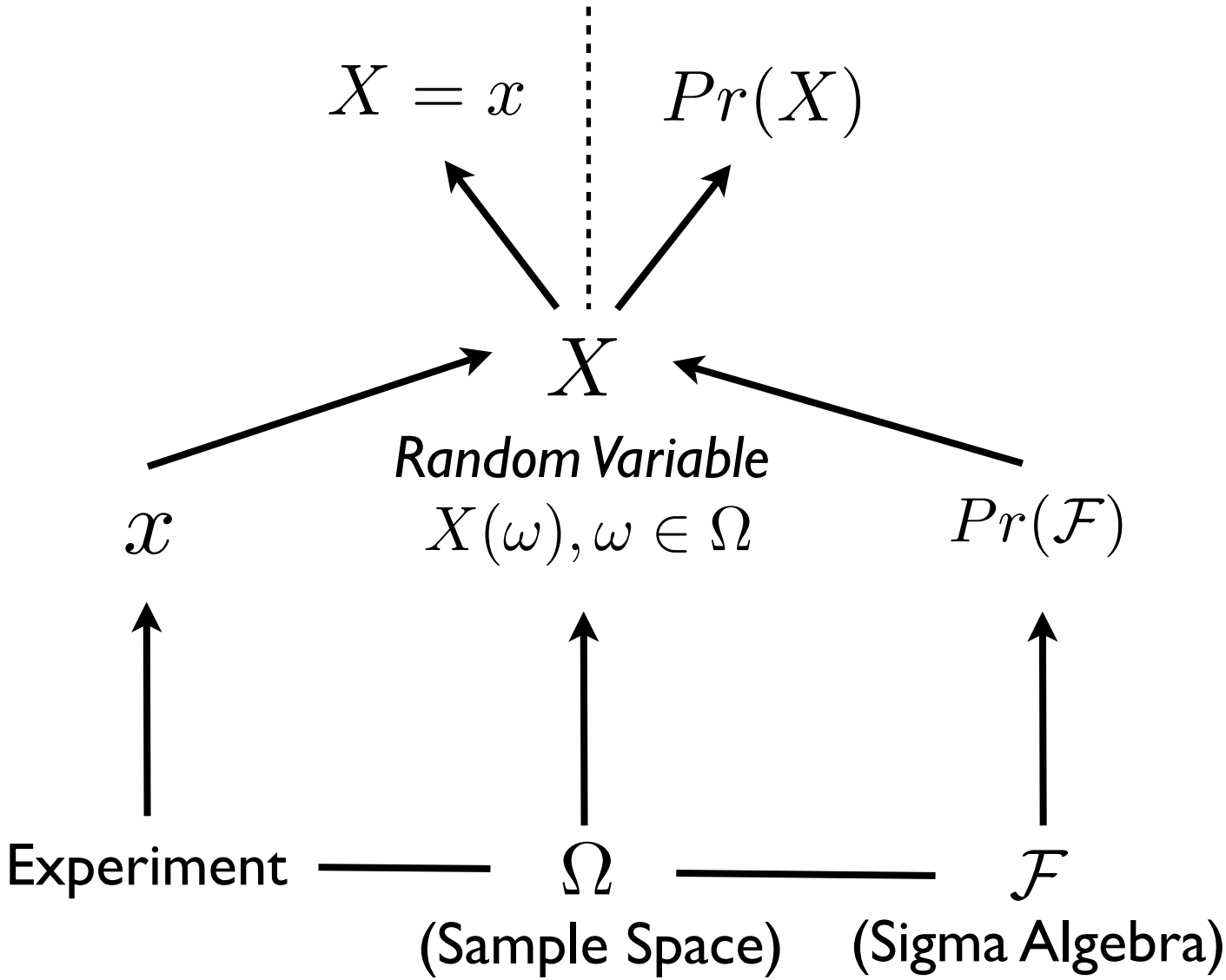
- Finally, note that if we were to take the covariance (or correlation) of two random variables Y_1 and Y_2 with the relationship:

$$Y_1 = a_1 + b_1X_1,\ Y_2 = a_2 + b_2X_2$$

$$\mathrm{Cov}(Y_1, Y_2) = b_1b_2\mathrm{Cov}(X_1, X_2)$$

$$\mathrm{Corr}(Y_1, Y_2) = \mathrm{Corr}(X_1, X_2)$$

So far


$X = x$
$Pr(X)$
X
Random Variable
$X(\omega), \omega \in \Omega$
x
$Pr(\mathcal{F})$
Experiment
Ω
(Sample Space)
$\mathcal{F}$
(Sigma Algebra)

Probability models I

- We have defined Pr(*X*), a probability model (=probability function!) on a random variable, which technically we produce by defining Pr function on the sigma algebra and the X (random variable function) on the sample space
- So far, we have generally considered such probability models / functions without defining them explicitly (except for a illustrative few examples)
- To define an explicit model for a given system / experiment we are going to assume that there is a "true" probability model, that is a consequence of the experiment that produces sample outcomes
- We place "true" in quotes since the defining a single true probability model for a given case could only really be accomplished if we knew every single detail about the system and experiment (would a probability model be useful in this case?)
- In practice, we therefore assume that the true probability distribution is within a restricted family of probability distributions, where we are satisfied if the true probability distribution in the family describes the results of our experiment pretty well / seems reasonable given our assumptions

Probability models II

- In short, we therefore start a statistical investigation *assuming* that there is a single true probability model that correctly describes the possible experiment outcomes given the uncertainty in our system
- In general, the starting point of a statistical investigation is to make *assumptions* about the form of this probability model
- More specifically, a convenient assumption is to assume our true probability model is specific model in a family of distributions that can be described with a compact equation
- This is often done by defining equations indexed by *parameters*

Probability models III

- **Parameter** - a constant(s) θ which indexes a probability model belonging to a family of models Θ such that $\theta \in \Theta$
- Each value of the parameter (or combination of values if there is more than on parameter) defines a different probability model: Pr(*X*)
- We assume one such parameter value(s) is the true model
- The advantage of this approach is this has reduced the problem of using results of experiments to answer a broad question to make an educated guess at the value of the parameter(s)
- Remember that the foundation of such an approach is still an assumption about the properties of the the experiment, and the system of interest (!!!)

Discrete parameterized examples

- Consider the probability model for the one coin flip experiment / number of tails.
- This is the Bernoulli distribution with parameter θ = *p* (what does *p* represent!?) where $\Theta = [0, 1]$
- We can write this *X* ~ Bern(*p*) and this family of probability models has the following form:

$$Pr(X = x|p) = P_X(x|p) = p^x(1-p)^{1-x}$$

- For the experiment of *n* coin flips / number of tails, *one possible* family Binomial distribution *X* ~ Bin(*n*, *p*):

$$Pr(X = x|n, p) = P_X(x|n, p) = \binom{n}{x} p^x(1-p)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

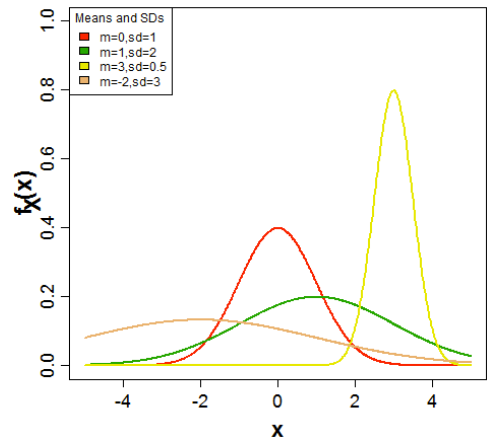
$$n! = n * (n-1) * (n-2) * ... * 1$$

- There are many other discrete examples: hypergeometric, Poisson, etc.

Continuous parameterized examples

- Consider the measure heights experiment (reals as approximation to the sample space) / identity random variable
- For this example we can use the family of normal distributions that are parameterized by $\theta = [\mu, \sigma^2]$ (what do these parameters represent!?) with the following possible values: $\Theta_\mu = (-\infty, \infty)$, $\Theta_{\sigma^2} = [0, \infty)$
- We often write this as $X \sim N(\mu, \sigma^2)$ and the equation has the following form:

$$Pr(X = x|\mu, \sigma^2) = f_X(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

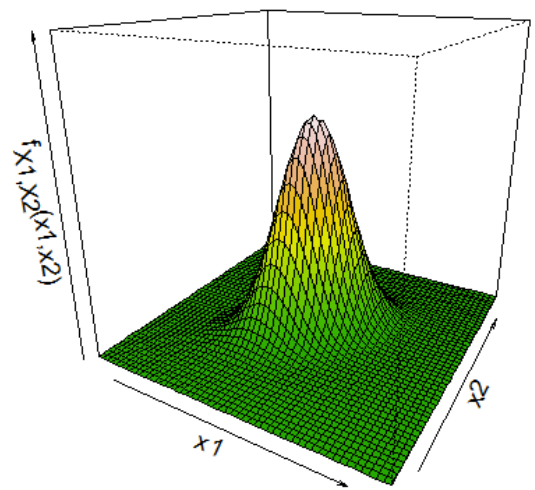


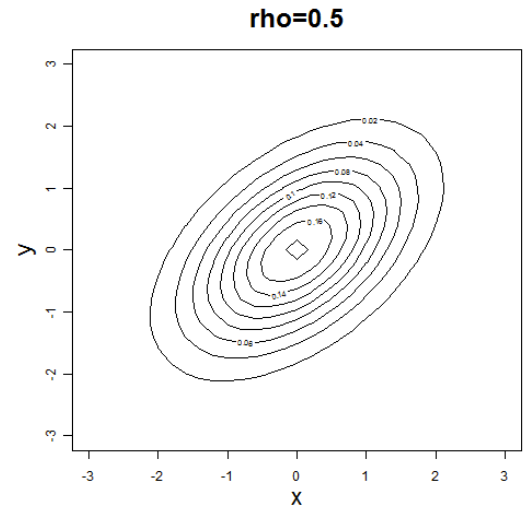
- There are many other continuous examples: uniform, exponential, etc.

Example for random vectors

- Since random vectors are the generalization of r.v.'s, we similarly can define parameterized probability models for random vectors
- As an example, if we consider an experiment where we measure "height" and "IQ" and we take the 2-D reals as the approximate sample space (vector identity function), we could assume the bivariate normal family of probability models:

$$f_{\mathbf{X}}(\mathbf{x}|\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho}} exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_1)^2}{2\sigma_2^2}\right)\right]$$





That's it for today

- Next lecture, we will begin our discussion of inference!