

MA/ST 810

Mathematical-Statistical Modeling and Analysis of Complex Systems

Review of Basic Probability

- The fundamentals, random variables, probability distributions
- Probability mass/density functions
- Transformations, random vectors, independence, expectation, variance, covariance, correlation
- Some specific probability distributions
- Conditional probability and conditional expectation
- Joint probability distributions

Review of basic probability

Probability distributions in statistical models:

- Are used to formalize assumptions on model components
- Arise in formalizing assessments of uncertainty based on statistical models (*inference*)

Here: Review

- Basics of probability theory*
- Some important, specific probability distributions

* A comprehensive introduction is given in Casella and Berger (2002), *Statistical Inference, Second Edition*

The fundamentals

(Statistical) experiment: Formal conceptualization

- One toss of a coin
- Choose a person from a population of size N at random
- Observe concentration in a blood sample

Sample space Ω : Set of *all possible outcomes* of an experiment

Examples: Countable or uncountable

- One toss of a coin: $\Omega = \{H, T\}$
- Choose a person from a population of size N : $\Omega = \{\omega_1, \dots, \omega_N\}$
- Observe concentration: $\Omega = (0, \infty)$
- Observe error committed: $\Omega = (-\infty, \infty)$

Event: A *collection of possible outcomes* of an experiment; i.e., any *subset* A of Ω

- Events $A \subset \Omega$ obey usual set-theoretic rules; e.g., union, intersection

The fundamentals

Probability: For each $A \subset \Omega$, assign a *number between 0 and 1*, denoted by $P(A)$

- Technically, not that simple
- \mathcal{B} = collection of subsets of S that includes \emptyset , closed under complementation, closed under countable unions (σ -algebra)

Probability function: For Ω with associated \mathcal{B} , P is a *probability function* with domain \mathcal{B} if

- $P(A) \geq 0$ for $A \in \mathcal{B}$
- $P(\Omega) = 1$
- $A_1, A_2, \dots \in \mathcal{B} \Rightarrow P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

The fundamentals

Properties: For $A, B \in \mathcal{B}$

- $P(\emptyset) = 0$
- $P(A) \leq 1$
- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A) \leq P(B)$ if $A \subset B$
- $P(A) = \sum_{i=1}^{\infty} P(A \cap B_i)$, B_i disjoint partition of Ω

Random variables

Random variable: A *function* from Ω into the real numbers (assigns a real number to each element of the sample space)

- Mapping from original sample space Ω to new sample space \mathcal{X}
- Often denoted by capital letters, e.g. X, Y

Example: Toss a coin two times

- $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\} = \{HH, HT, TH, TT\}$,
 $X(\omega) = \#$ of heads in two tosses taking values in $\mathcal{X} = \{0, 1, 2\}$

Example: Sample a person from a population of size N and observe survival time

- $\Omega = \{\omega_1, \dots, \omega_N\}$,
 $X(\omega) =$ survival time taking values in $\mathcal{X} = (0, \infty)$

Example: Measurement error (uncountable)

- $\Omega = \{ \text{all possible conditions of measurement} \}$,
 $\epsilon(\omega) =$ error committed taking values in $\mathcal{X} = \mathbb{R}$

Random variables

Probability function for X :

- *Countable* Ω and \mathcal{X} – $X = x_i \in \mathcal{X}$ iff ω_j is such that $X(\omega_j) = x_i$

$$P_X(X = x_i) = P(\{\omega_j \in \Omega : X(\omega_j) = x_i\})$$

- *Uncountable* \mathcal{X} – for $A \in \mathcal{X}$ (actually, in a certain σ -algebra of subsets of \mathcal{X})

$$P_X(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\})$$

- Customary to discuss probability with respect to *random variables* and suppress X subscript
- Write X for the random variable (the function) and x for its possible values (realizations, elements of \mathcal{X})
- “*Probability distribution*”

Probability distributions

Cumulative distribution function (cdf): For random variable X

$$F_X(x) = F(x) = P(X \leq x) \quad \text{for all } x$$

(not just $x \in \mathcal{X}$)

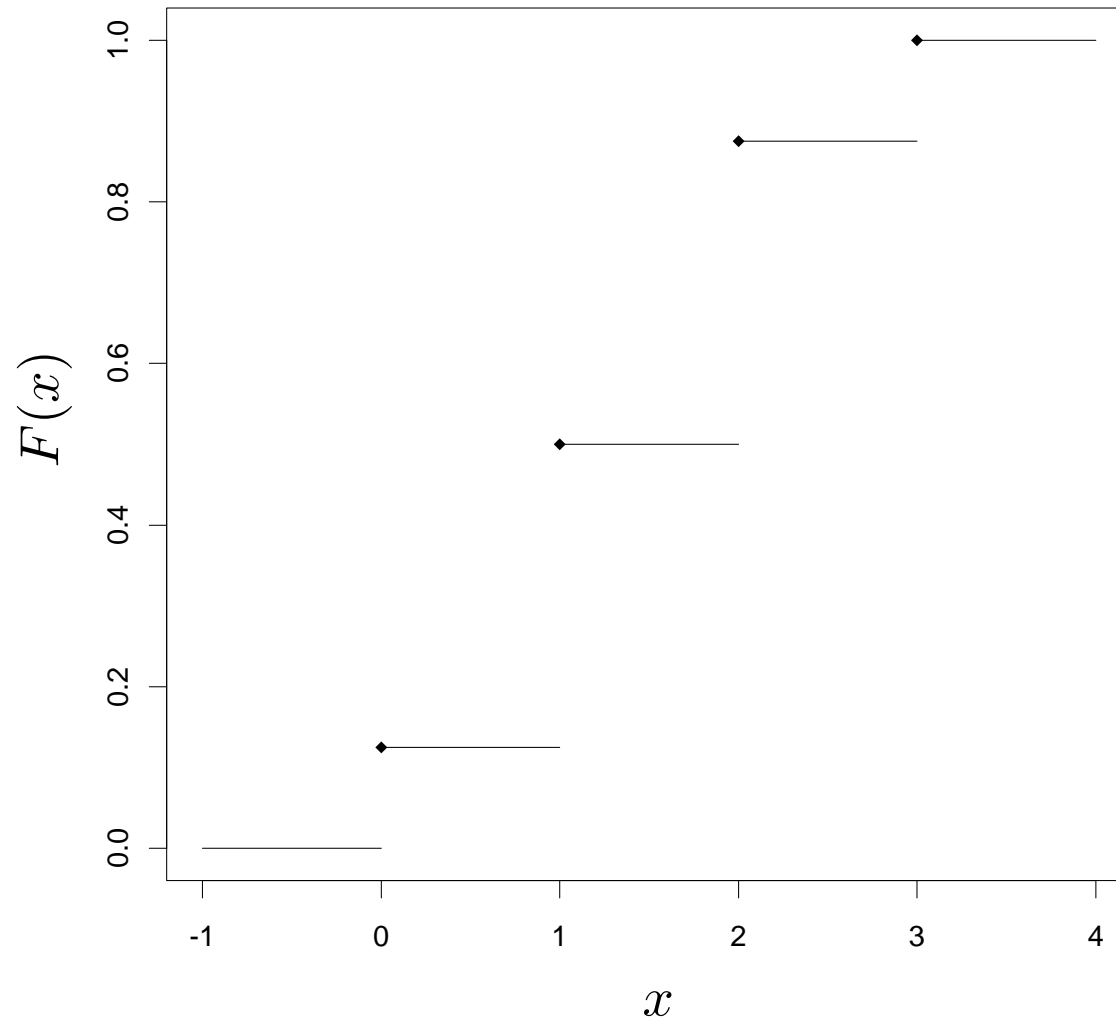
- $F(x)$ is *nondecreasing* and *right continuous*
- $\lim_{x \rightarrow -\infty} F(x) = 0$, $\lim_{x \rightarrow \infty} F(x) = 1$

Example: Toss a coin three times, $X = \#$ heads

x	$P(X = x)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

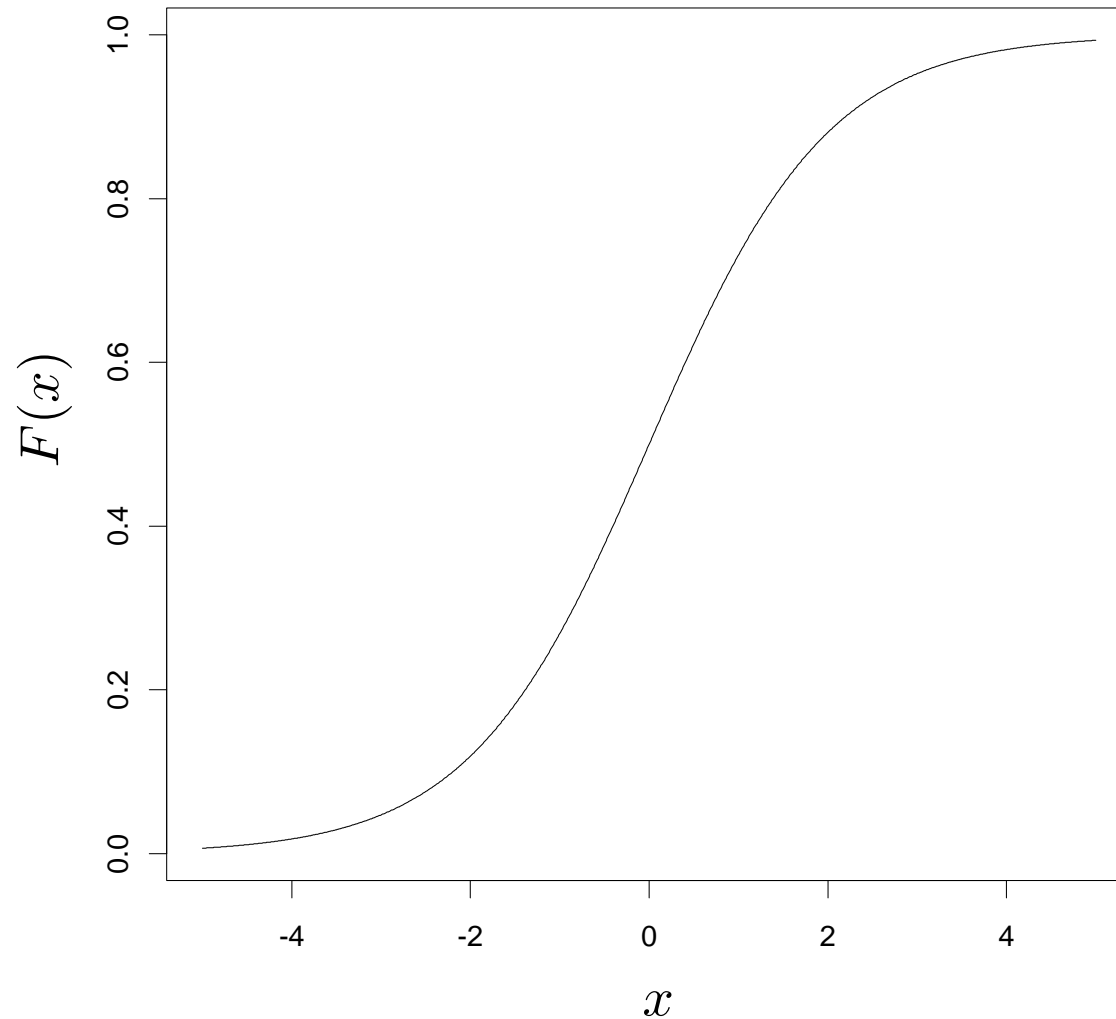
$$F(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \frac{1}{8} & \text{if } 0 \leq x < 1 \\ \frac{4}{8} & \text{if } 1 \leq x < 2 \\ \frac{7}{8} & \text{if } 2 \leq x < 3 \\ 1 & \text{if } 3 \leq x < \infty \end{cases}$$

Probability distributions



Probability distributions

$$F(x) = \frac{1}{1 + e^{-x}}$$



Probability mass/density functions

Discrete and continuous random variables: A random variable X is said to be

- *Continuous* if $F(x)$ is a continuous function of x
- *Discrete* if $F(x)$ is a step function of x

Probability mass and density functions: Concerned with “point probabilities” of random variables

- Discrete: *probability mass function*
- Continuous: *probability density function*

Probability mass/density functions

Probability mass function (pmf):

$$f(x) = P(X = x) \quad \text{for all } x$$

- Thus, $P(X \leq x) = F(x) = \sum_{u:u \leq x} f(u)$

Example: Binomial probability mass function

- *Bernoulli trial*: experiment with 2 possible outcomes
- X has a *Bernoulli* probability distribution with $\mathcal{X} = \{0, 1\}$, if

$$X = \begin{cases} 1 & \text{with probability } p & \text{“success”} \\ 0 & \text{with probability } 1 - p & \text{“failure”} \end{cases} \quad 0 \leq p \leq 1$$

- *Binomial distribution*: for n identical Bernoulli trials, let $Y =$ total # successes with sample space $\mathcal{Y} = \{0, 1, \dots, n\}$

$$f(y) = P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}, \quad y \in \mathcal{Y}, \quad = 0 \text{ otherwise}$$

Probability mass/density functions

Probability density function (pdf): Must be a little more careful when X is *continuous*

- $\{X = x\} \subset \{x - \epsilon < X < x\}$ for all $\epsilon > 0 \Rightarrow$

$$0 \leq P(X = x) \leq P(x - \epsilon < X \leq x) = F(x) - F(x - \epsilon)$$

Thus, by continuity of $F(\cdot)$,

$$0 \leq P(X = x) \leq \lim_{\epsilon \downarrow 0} \{F(x) - F(x - \epsilon)\} = 0$$

- By analogy to discrete pmf, for continuous $f(\cdot)$

$$P(X \leq x) = F(x) = \int_{-\infty}^x f(u) du \quad \text{for all } x$$

- $\Rightarrow d/dx F(x) = f(x)$

Probability mass/density functions

Probability mass and density functions satisfy:

- $f(x) \geq 0$ for all x
- $\sum_x f(x) = 1$ (pmf) or $\int_{-\infty}^{\infty} f(x) dx = 1$ (pdf)

Notation in this course: To avoid confusion with our use of f to denote the solution of a system as before

- We will often use $P(x)$ to denote the *cdf* of a random variable X and $p(x)$ to denote the *pmf* or *pdf* as appropriate
- We may add *subscripts* when speaking simultaneously of several random variables; e.g., $p_\epsilon(\epsilon)$ and $p_X(x)$
- We will use “ \sim ” to denote “*distributed as*”

More on random variables

Transformations of random variables: If X is a random variable with cdf $F_X(x)$, then a function $Y = g(X)$ is *also a random variable* with new sample space \mathcal{Y} with elements of form $y = g(x)$

$$P(Y \in A) = P\{g(X) \in A\} = P\{x \in \mathcal{X} : g(x) \in A\} = P\{X \in g^{-1}(A)\}$$

where g^{-1} is inverse mapping from \mathcal{Y} to \mathcal{X} .

- The distribution of Y depends on that of X
- In particular, $F_Y(y)$, $f_Y(y)$ are *related* to $F_X(x)$, $f_X(x)$

Random vectors

Several random variables at once: p -dimensional *random vector*

$(X_1, \dots, X_p)^T$ is a function from Ω into \mathbb{R}^p . Consider $p = 2$

- All components *discrete* – *joint pmf*

$$f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$

Satisfies $\sum_{x_1, x_2} f(x_1, x_2) = 1$

- All components *continuous* – *joint pdf* $f(x_1, x_2)$ from \mathbb{R}^2 into \mathbb{R} satisfies

$$P\{X_1, X_2) \in A\} = \int \int_A f(x_1, x_2) dx_1 dx_2, \quad \int \int f(x_1, x_2) dx_1 dx_2 = 1$$

- *Marginal* pmf and pdf: E.g., X_1

$$f_{X_1}(x_1) = \sum_{x_2} f(x_1, x_2) \text{ or } f_{X_1}(x_1) = \int f(x_1, x_2) dx_2$$

Independence and expectation

Independent random variables: X_1 and X_2 are *independent* if

$$f(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2), \quad \text{write } X_1 \perp\!\!\!\perp X_2$$

Expectation of a random variable: The “*average*” value of a random variable

- “*weighted*” according to the probability distribution
- Measure of “*center*”

Expected value or mean: For random variable X , the expected value of $g(X)$ is

$$E\{g(X)\} = \begin{cases} \int_{-\infty}^{\infty} g(x)f(x) dx & X \text{ continuous} \\ \sum_x g(x)f(x) = \sum_x g(x)P(X = x) & X \text{ discrete} \end{cases}$$

Variance and higher moments

Higher moments: For random variable X and integer k

- The k th moment of X is $E(X^k)$
- The k th *central moment* is $E\left[\{X - E(X)\}^k\right]$

Variance: Second central moment

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

- Measure of degree of “*spread*” of distribution about its mean
- Standard deviation = $\sqrt{\text{var}(X)}$ on same scale of X
- Quantifies *variation*

Random vectors: Element-by-element using marginal pmf/pdf

Covariance and correlation

Covariance and correlation: Measures of “*degree of association*” – For any two random variables

- **Covariance** between X_1 and X_2 is defined as

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

- Will be > 0 if $X_1 > E(X_1)$ and $X_2 > E(X_2)$ or $X_1 < E(X_1)$ and $X_2 < E(X_2)$ tend to happen together
- Will be < 0 if $X_1 > E(X_1)$ and $X_2 < E(X_2)$ or $X_1 < E(X_1)$ and $X_2 > E(X_2)$ tend to happen together
- Will = 0 if X_1 and X_2 are $\perp\!\!\!\perp$
- **Correlation** is covariance put on a unitless basis

$$\rho_{X_1 X_2} = \text{corr}(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sqrt{\text{var}(X_1)\text{var}(X_2)}}$$

- $-1 \leq \rho_{X_1 X_2} \leq 1$; $\rho_{X_1, X_2} = -1$ or 1 iff $X_1 = a + bX_2$

Some specific probability distributions

Discrete probability distributions:

- $X \sim \text{Binomial}(n, p)$

$$f(x) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \dots, n$$

$$E(X) = np, \quad \text{var}(X) = np(1 - p)$$

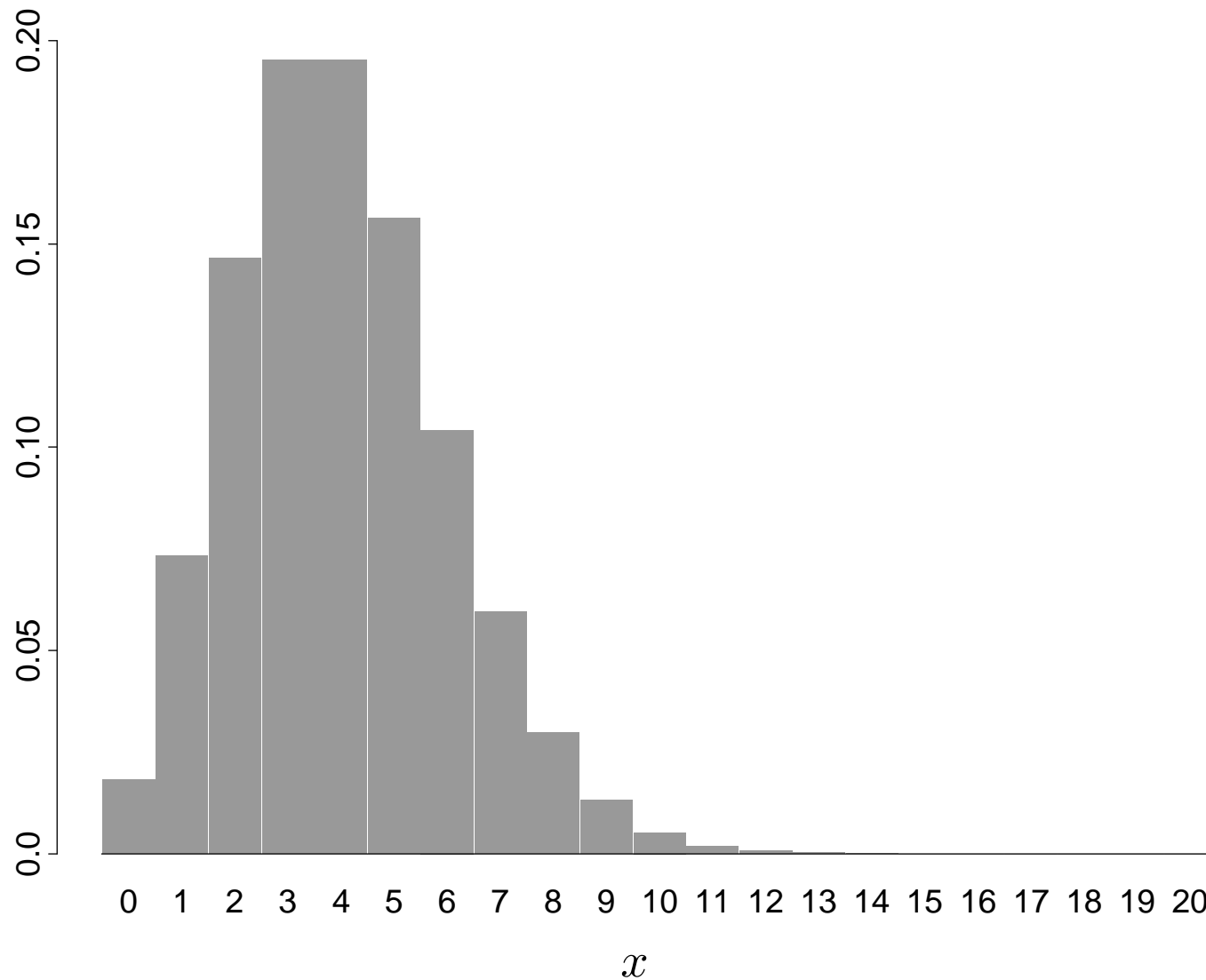
- $X \sim \text{Poisson}(\lambda)$ – a model for *counts*

$$f(x) = P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

$$E(X) = \lambda, \quad \text{var}(X) = \lambda$$

Some specific probability distributions

Poisson pmf with $\lambda = 4$:



Some specific probability distributions

Continuous probability distributions:

- *Normal* or *Gaussian* distribution:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

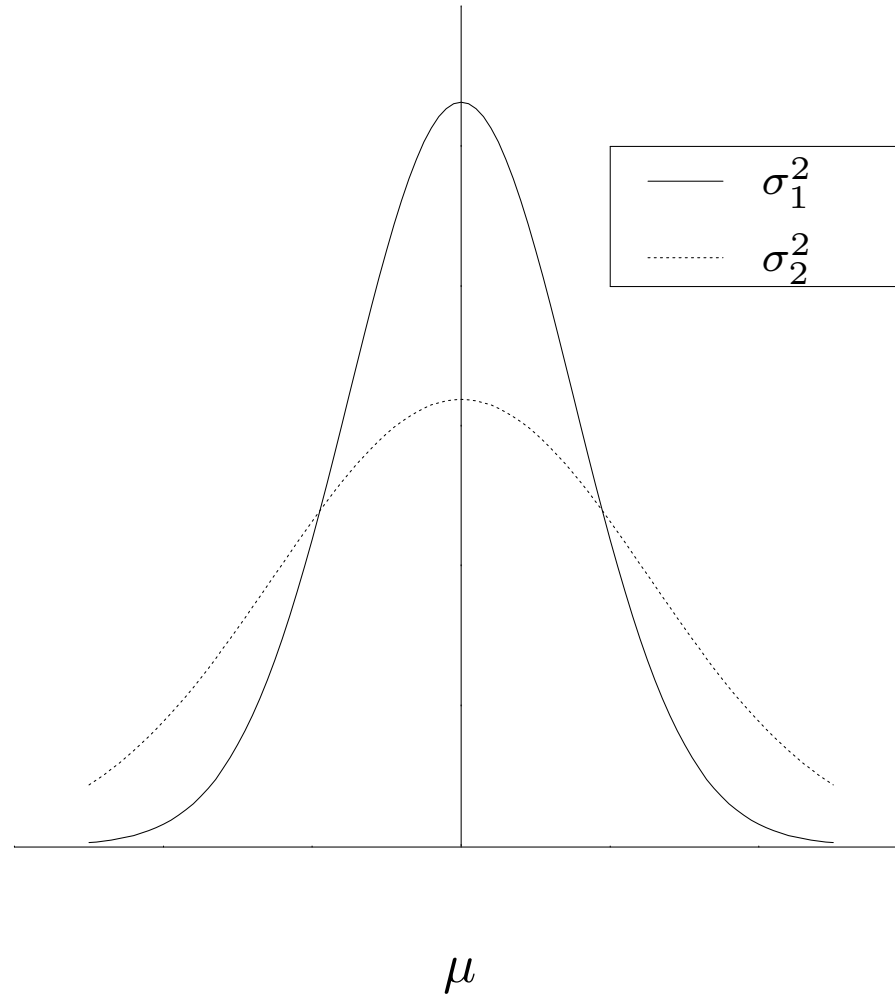
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty$$

$$E(X) = \mu, \text{ var}(X) = \sigma^2, \sigma > 0$$

- *Symmetric* about its mean
- $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$ *standard normal*
- A (the most) popular model for phenomena such as measurement errors, observations on biological, physical phenomena
- Plays a central role in approximate methods of *statistical inference* for complex models

Some specific probability distributions

Two normal pdfs with same mean μ , different variances $\sigma_1^2 < \sigma_2^2$:



Some specific probability distributions

Continuous probability distributions:

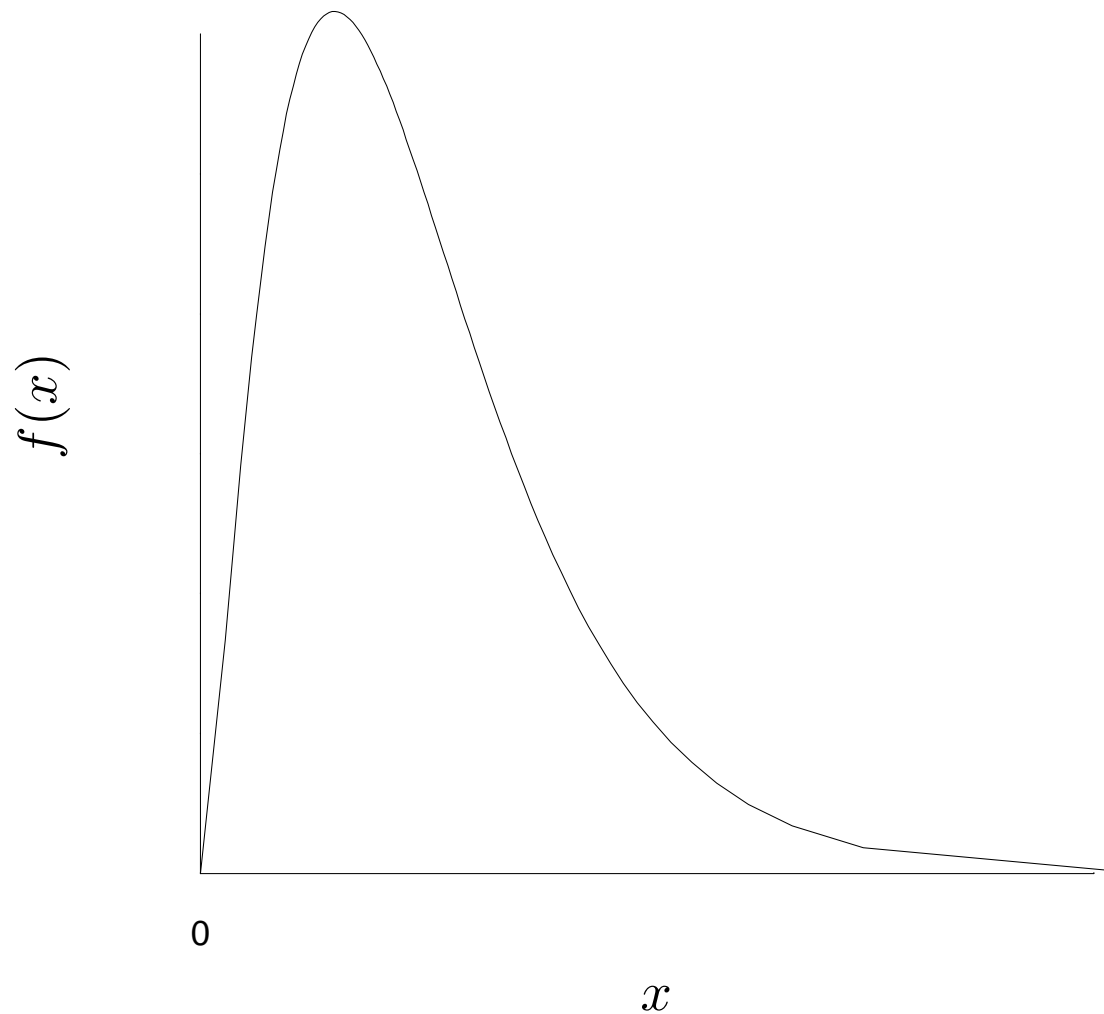
- *Lognormal* distribution: If $\log X \sim \mathcal{N}(\mu, \sigma^2)$, then X has a lognormal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{x} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}, \quad 0 < x < \infty$$

$$E(X) = e^{\mu + \sigma^2/2}, \quad \text{var}(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2} \propto \{E(X)\}^2$$

- Constant *coefficient of variation (CV)* $= \sqrt{\text{var}(X)}/E(X)$ (“*noise-to-signal*”) – does not depend on $E(X)$
- A common model for biological phenomena
- Skewed (asymmetric) with “*long right tail*”
- Looks more and more symmetric as $\sigma \rightarrow 0$

Some specific probability distributions



Some specific probability distributions

Continuous probability distributions:

- *Gamma* distribution

$$f(x) = \frac{1}{\Gamma(a)b^a} x^{a-1} \exp(-x/b), \quad 0 < x < \infty, \quad a, b > 0$$

$$E(X) = ab, \quad \text{var}(X) = ab^2$$

- Constant CV = $a^{-1/2}$
- Similar in shape to lognormal
- Looks more and more symmetric as $a \rightarrow \infty$
- Special case 1: *Exponential distribution* $a = 1$
- Special case 2: *Chi squared (χ^2) distribution with k degrees of freedom*: For integer $k > 0$, set $a = k/2$, $b = k \Rightarrow$ important in *statistical inference*

Some specific probability distributions

Continuous probability distributions: These two are also important in *statistical inference*

- *Student's t distribution with k degrees of freedom:* If $U \sim \mathcal{N}(0, 1)$, $V \sim \chi_k^2$ are $\perp\!\!\!\perp$, then $X = U/\sqrt{V/k} \sim t_k$ with pdf

$$f(x) = \frac{\Gamma\{(k+1)/2\}}{\Gamma(k/2)} \frac{1}{\sqrt{k\pi}} \frac{1}{(1+x^2/k)^{(k+1)/2}}, \quad -\infty < x < \infty$$

$$E(X) = 0 \text{ if } k > 1, \text{ var}(X) = k/(k-2) \text{ if } k > 2$$

- Symmetric like normal, with “*heavier tails*,” becomes normal as $k \rightarrow \infty$
- *F distribution with k_1, k_2 degrees of freedom* If $U \sim \chi_{k_1}^2$, $V \sim \chi_{k_2}^2$ are $\perp\!\!\!\perp$, then $X = (U/k_1)/(V/k_2) \sim \mathcal{F}_{k_1, k_2}$ with pdf

$$f(x) = \frac{\Gamma\{(k_1+k_2)/2\}}{\Gamma(k_1/2)\Gamma(k_2/2)} \left(\frac{k_1}{k_2}\right)^{k_1/2} \frac{x^{k_1/2-1}}{\{1+(k_1/k_2)x\}^{(k_1+k_2)/2}}, \quad 0 < x < \infty$$

Some specific probability distributions

Multivariate normal distribution: *Random vector* $X = (X_1, \dots, X_p)^T$ has a multivariate (p -variate) normal distribution if $\alpha^T X \sim \text{normal}$ $\forall \alpha \in \mathbb{R}^p$

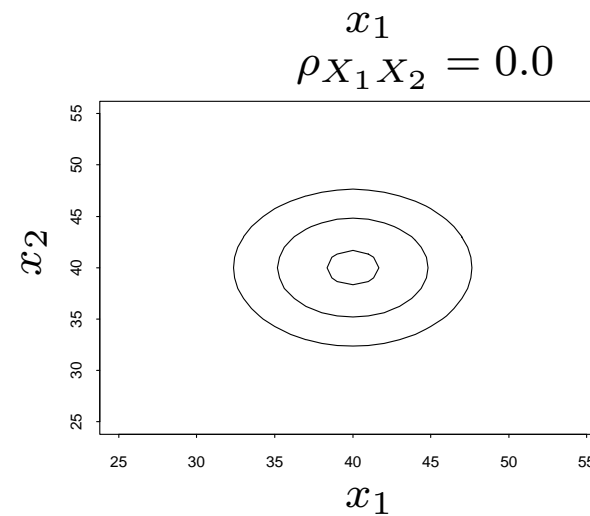
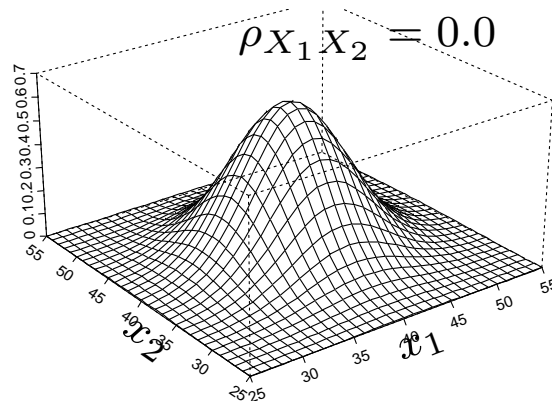
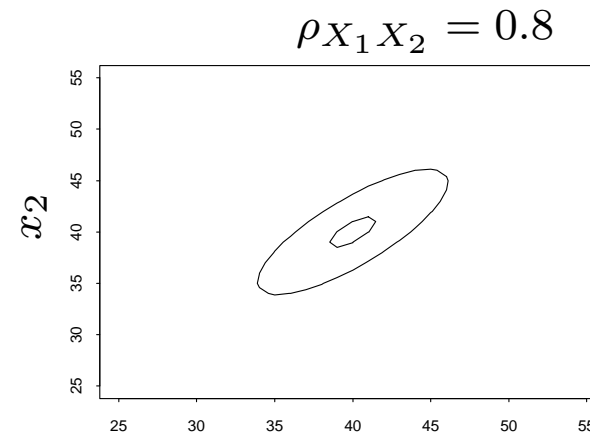
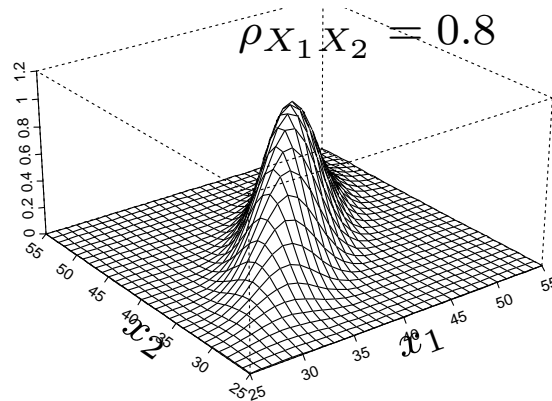
$$f(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-(x - \mu)^T \Sigma^{-1} (x - \mu) / 2\},$$

for $x = (x_1, \dots, x_p)^T \in \mathbb{R}^p$

- $E(X) = \mu = (\mu_1, \dots, \mu_p)^T = \{E(X_1), \dots, E(X_p)\}^T$
- Σ ($p \times p$) is such that $\Sigma_{jj} = \text{var}(X_j)$, $\Sigma_{jk} = \Sigma_{kj} = \text{cov}(X_j, X_k)$
- $\Sigma = E\{(x - \mu)(x - \mu)^T\}$ is the *covariance matrix*
- The *marginal* pdfs are *univariate* normal
- Incredibly important in statistical *modeling* and *inference*

Some specific probability distributions

Two bivariate ($p = 2$) normal pdfs:



Conditional probability and expectation

Conditional probability: probabilistic statement of “*relatedness*”

- E.g., weight $Y > 200$ *more likely* for $X = 6$ than $X = 5$ feet tall
- X, Y *discrete*: conditional pmf *given* $X = x$ is function of y

$$p(y|x) = P(Y = y|X = x) = \frac{p(x, y)}{p_X(x)}, \quad p_X(x) > 0$$

and satisfies $\sum_y f(y|x) = 1$ (a pmf for fixed x)

- X, Y *continuous*: conditional pdf *given* $X = x$ is function of y

$$p(y|x) = \frac{p(x, y)}{p_X(x)}, \quad p_X(x) > 0$$

and satisfies $\int_{-\infty}^{\infty} p(y|x) dy = 1$ (a pdf for fixed x)

- Thus, the *conditional distribution* of Y given $X = x$ is *possibly different* for each x
- $Y|X$ denotes the *family* of probability distributions so defined

Conditional probability and expectation

Conditional expectation: For $g(Y)$ a function of Y , define the *conditional expectation of Y given $X = x$*

$$E\{g(Y)|X = x\} = E\{g(Y)|x\} = \sum_y g(y)p(y|x) \quad \textit{discrete}$$

$$E\{g(Y)|X = x\} = E\{g(Y)|x\} = \int_{-\infty}^{\infty} g(y)p(y|x) dy \quad \textit{continuous}$$

- *Conditional expectation* is a function of x taking a value in \mathbb{R} , *possibly different* for each x
- Thus, $E\{g(Y)|X\}$ is a *random variable* whose value depends on the value of X (and takes on values $E\{g(Y)|x\}$ as X takes on values x)
- *Conditional variance* defined similarly

$$\text{var}(Y|x) = E[\{Y - E(Y|x)\}^2 | x] = E(Y^2|x) - \{E(Y|x)\}^2$$

Conditional probability and expectation

Relation to independence: If X and Y are *independent* random variables/vectors, then

$$p(y|x) = \frac{p(x, y)}{p_X(x)} = \frac{p_X(x)p_Y(y)}{p_X(x)} = p_Y(y)$$

and

$$E\{g(Y)|X = x\} = E\{g(Y)\} \quad \text{for any } x$$

so $E\{g(Y)|X\}$ is a *constant random variable* and equal to $E\{g(Y)\}$

Fun facts:

- $E\{E(Y|X)\} = \int_{-\infty}^{\infty} E(Y|x)p(x) dx = \int_{-\infty}^{\infty} yp(y) dy = E(Y)$ [using the definition of $E(Y|X)$]
- $\text{var}(Y) = \text{var}\{E(Y|X)\} + E\{\text{var}(Y|X)\}$

Facts for joint probability distributions

Fun facts: For joint probability distributions (including conditional)

- For *random variables* X_1 and X_2 and constants a and b ,
 - $E(aX_1 + bX_2) = aE(X_1) + bE(X_2)$
 - $\text{var}(aX_1 + bX_2) = a^2\text{var}(X_1) + b^2\text{var}(X_2) + 2ab\text{cov}(X_1, X_2)$ with
$$\text{var}(aX_1 + bX_2) = a^2\text{var}(X_1) + b^2\text{var}(X_2) \quad \text{if } X_1 \perp\!\!\!\perp X_2$$
- For a $(n \times 1)$ *random vector* $X = (X_1, \dots, X_n)$ the *covariance matrix* $E\left[\{X - E(X)\}\{X - E(X)\}^T\right]$ has
 - diagonal elements $\text{var}(X_j)$, $j = 1, \dots, n$
 - off-diagonal elements $\text{cov}(X_j, X_{j'})$
- If X_1 and X_2 are two *independent*, $(n \times 1)$ *random vectors*, each with a *multivariate normal probability distribution*, and A and B are conformable constant matrices, then the probability distribution of $AX_1 + BX_2$ is *also multivariate normal*

Facts for joint probability distributions

Fun facts: For joint probability distributions (including conditional)

- In fact, for two $(n \times 1)$ *random vectors* X_1 and X_2 with covariance matrices Σ_1 and Σ_2 and conformable constant matrices A and B
 - $E(AX_1 + BX_2) = AE(X_1) + BE(X_2)$
 - The *covariance matrix* of $AX_1 + BX_2$ is

$$A\Sigma_1A^T + B\Sigma_2B^T + AE\left[\{X_1 - E(X_1)\}\{X_2 - E(X_2)\}\right]B^T \\ + BE\left[\{X_2 - E(X_2)\}\{X_1 - E(X_1)\}\right]A^T,$$

which equals

$$A\Sigma_1A^T + B\Sigma_2B^T \quad \text{if } X_1 \perp\!\!\!\perp X_2$$

(all elements of X_1 are *independent* of all elements of X_2)