Probability/Statistics Review

H. T. Banks

Center for Research in Scientific Computation (CRSC)
Center for Quantitative Sciences in Biomedicine (CQSB)
North Carolina State University
Raleigh, NC 27695

Center for Research in Scientific Computation
North Carolina State University

Center for Quantitative Sciences in Biomedicine
North Carolina State University

PROBABILITY REVIEW

- probability and statistics essential mathematical tool in development of inverse problem formulations and
- subsequent analysis as well as for approaches to statistical hypothesis testing
- here a few definitions and basic concepts in the theory of probability and statistics that are essential for the understanding of estimators, confidence intervals and hypothesis testing

Probability

- denote events by capital letters –the probability of event A is $P(A)$.
- set of all possible outcomes, the *sample space* denoted by S.
- Example: experiment of rolling of a die in which there are six possible outcomes–sample space is

$$S = \{1, 2, 3, 4, 5, 6\} \tag{1}$$

and an event A might be defined as

$$A = \{1, 5\}, \tag{2}$$

which consists of the outcomes 1 and 5.

Associated with event A contained in S is its probability $P(A)$. In case the sample space S is discrete (finite or countably infinite), probability satisfies the following postulates:

(i) $P(A) \geq 0$,

(ii) $P(S) = 1$,

(iii) If $A_1, A_2, A_3, \ldots,$ is a finite or an infinite sequence of disjoint subsets of S, then

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots . \tag{3}$$

In our fair experiment of the rolling of a die, each possible outcome has probability $\frac{1}{6}$. The event A as defined by (2) consists of two disjoint subevents, and hence $P(A) = \frac{2}{6} = \frac{1}{3}$. Using the three postulates of probability, a number of immediate consequences can also be derived which have important applications. For example, probabilities cannot exceed 1 ($P(A) \leq 1$ for any event A), the empty set $\emptyset$ has probability 0 ($P(\emptyset) = 0$), and the probability that an event will occur and that it will not occur always add up to 1 ($P(A) + P(\bar{A}) = 1$ where $\bar{A}$ denotes the complement of the event A which consists of all sample points in S that are not in A). If $P(A) = 1$, then we say that "the event A occurs with probability 1 or *almost surely* (a.s.)."

Now perform two experiments and consider outcomes. For example, the two experiments may be two separate tosses of a single die or a single toss of two dice. The sample space in this case consists of 36 pairs (i, j), where $i, j = 1, 2, \ldots, 6$. Note each point in sample space has probability $\frac{1}{36}$. Consider the probability of joint events, such as $\{i = 2, j = \text{odd}\}$. Denote the possible outcomes of one experiment by A_i, $i = 1, 2, \ldots, n$, and by B_j, $j = 1, 2, \ldots, m$ the possible outcomes of the second experiment. The combined experiment has possible joint outcomes (A_i, B_j), where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. The joint probability $P(A_i, B_j)$ satisfies $0 \leq P(A_i, B_j) \leq 1$.If outcomes B_j for $j = 1, 2, \ldots, m$ are mutually exclusive (i.e., $B_i \bigcap B_j = \emptyset, i \neq j$), then

$$\sum_{j=1}^{m} P(A_i, B_j) = P(A_i). \tag{4}$$

Furthermore, if all the outcomes of the two experiments are mutually exclusive, then

$$\sum_{i=1}^{n}\sum_{j=1}^{m} P(A_i, B_j) = 1. \tag{5}$$

The generalization of the above concept to more than two experiments follows in a straightforward manner.

Next, we consider a joint event with probability $P(A, B)$. Assuming that event A has occurred, we wish to determine the probability of the event B. This is called the *conditional probability* of the event B given the occurrence of the event A and is given by

$$P(B|A) = \frac{P(A, B)}{P(A)}, \tag{6}$$

where $P(A) > 0$. A very useful relationship for conditional probabilities, which is known as *Bayes' theorem*, states that if A_i,

where $i = 1, 2, \ldots, n$, are mutually exclusive events such that

$$\bigcup_{i=1}^{n} A_i = S \tag{7}$$

and B is an arbitrary event with $P(B) > 0$, then

$$\begin{aligned} P(A_i|B) &= \frac{P(A_i, B)}{P(B)} \\ &= \frac{P(B|A_i)P(A_i)}{\sum_{l=1}^{n} P(B|A_l)P(A_l)}. \end{aligned}$$

Random Variables

In most applications of probability theory, we are interested only in a particular aspect of the outcome of an experiment. For example, in the experiment of the rolling of a pair of dice, we are generally interested only in the total and not in the outcome for each die. In the language of probability and statistics, the total which we obtain with a pair of dice is called a *random variable*. More formally, the random variable $X(A)$ represents the functional relationship between a random event A and a real number. For example, if we flip a coin the possible outcomes are heads, H, or tails, T. We may define a random variable $X(A)$ by

$$X(A) = \begin{cases} 1, & A = H, \\ -1, & A = T. \end{cases} \tag{8}$$

We note that the random variable may be continuous or discrete. Associated with a random variable X, we consider the event $X \leq x$, where $-\infty < x < \infty$. The probability of this event is defined by

$$F(x) = P(X \leq x), \tag{9}$$

where the function $F(x)$ is called the *probability distribution function* of the random variable X. It is also called the *cumulative distribution function*

or *(cdf)*. The distribution function is right continuous and has the following properties:

(i) $0 \leq F(x) \leq 1$,

(ii) $F(x_1) \leq F(x_2)$ if $x_1 \leq x_2$,

(iii) $F(-\infty) = 0$,

(iv) $F(\infty) = 1$.

The derivative $p(x)$ (when it exists) of the distribution function $F(x)$ given by

$$p(x) = \frac{dF(x)}{dx} \tag{10}$$

is called the *probability density function* or *(pdf)*. The name "density function" comes from the fact that the probability of the event $x_1 \leq X \leq x_2$ is given by

$$\begin{aligned} P(x_1 \leq X \leq x_2) &= P(X \leq x_2) - P(X \leq x_1) \\ &= F(x_2) - F(x_1) \\ &= \int_{x_1}^{x_2} p(x)\, dx. \end{aligned}$$

The probability density function $p(x)$ satisfies the following properties:

(i) $p(x) \geq 0$,

(ii) $\int_{-\infty}^{\infty} p(x)\, dx = F(\infty) - F(-\infty) = 1$.

Moreover, it is common to denote *random variables* by capital letters X, Y, Z, etc., while one denotes particular *realizations* by the corresponding lower case letters x, y, z, etc.

Statistical Averages of Random Variables

Of particular importance in the characterization of the outcomes of experiments and random variables are the concepts of first and second moments of a single random variable and the joint moments (correlation and covariance) between any pair of random variables in a multi-dimensional set of random variables.

We begin the discussion of these statistical averages by considering first a single random variable X and its pdf $p(x)$. The *mean value* μ or *expected value* of the random variable X is defined by

$$\mu = E(X) = \int_{-\infty}^{\infty} xp(x)\,dx, \tag{11}$$

where $E(\cdot)$ is called the expected value operator (or statistical averaging operator). This is the *first moment* of the random variable

X. The nth moment of a probability distribution of a random variable X is defined as

$$E(X^n) = \int_{-\infty}^{\infty} x^n p(x)\, dx. \tag{12}$$

We can also define the *central moments*, which are the moments of the difference between X and μ. The second central moment, which is called the *variance* of X, is defined by

$$\sigma^2 = \text{var}(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx. \tag{13}$$

The square root σ of the variance of X is called the *standard deviation* of X. Variance is a measure of the "randomness" of the random variable X. It is related to the first and second moments

through the relationship

$$\begin{aligned}\sigma^2 &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - \mu^2.\end{aligned}$$

In the important case of multi-dimensional or R^p–valued vector random variables $X = (X_1, X_2, \ldots, X_p)^T$, we can define joint moments of any order. However, the joint moments that are most useful in practical applications are the *joint moments* defined by

$$E(X_i X_j) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_i x_j p(x_i, x_j)\, dx_i dx_j, \tag{14}$$

which are called the *correlation* (**NOT** to be confused with *correlation coefficients*) between the random variables X_i and X_j.

Here, $p(x_i, x_j)$ are the *marginal densities* defined by

$$p(x_i, x_j) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_1, \ldots, x_p) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_{j-1} dx_{j+1} \ldots dx_p.$$

Also of particular importance is the joint central moment, which is also called the *covariance* of X_i and X_j and is given by

$$\begin{aligned} \mu_{ij} &\equiv E[(X_i - \mu_i)(X_j - \mu_j)] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) p(x_i, x_j)\, dx_i dx_j \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_i x_j p(x_i, x_j)\, dx_i dx_j - \mu_i \mu_j \\ &= E(X_i X_j) - \mu_i \mu_j. \end{aligned} \tag{15}$$

The $(p \times p)$ matrix with elements μ_{ij} is called the *covariance matrix* of the random variable $X = (X_1, \ldots, X_p)$. Two random variables are

said to be *uncorrelated* if $E(X_iX_j) = E(X_i)E(X_j) = \mu_i\mu_j$. In that case, the covariance $\mu_{ij} = 0$. We also note that when X_i and X_j are statistically independent, they are uncorrelated. The reverse is, however, **not true**. That is, if X_i and X_j are **uncorrelated**, they are **not** necessarily **statistically independent.**

Special Probability Distributions

We say a continuous random variable has a certain *distribution* (e.g., *Gaussian distribution*) when it has the corresponding probability density. We review in this section several frequently encountered random variables, their pdf's, and their moments.

Uniform distribution.

The pdf of a uniformly distributed random variable X is given by

$$p(x) = \begin{cases} 1/(b-a), & a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

This is also called a *rectangular distribution* and its graph is depicted in Figure 1. The first two moments of X are

$$\begin{aligned} E(X) &= \mu = \frac{a+b}{2} \\ E(X^2) &= \frac{a^2+b^2+ab}{3} \end{aligned}$$

and the variance is

$$\text{var}(X) = \sigma^2 = \frac{(a-b)^2}{12}. \tag{17}$$

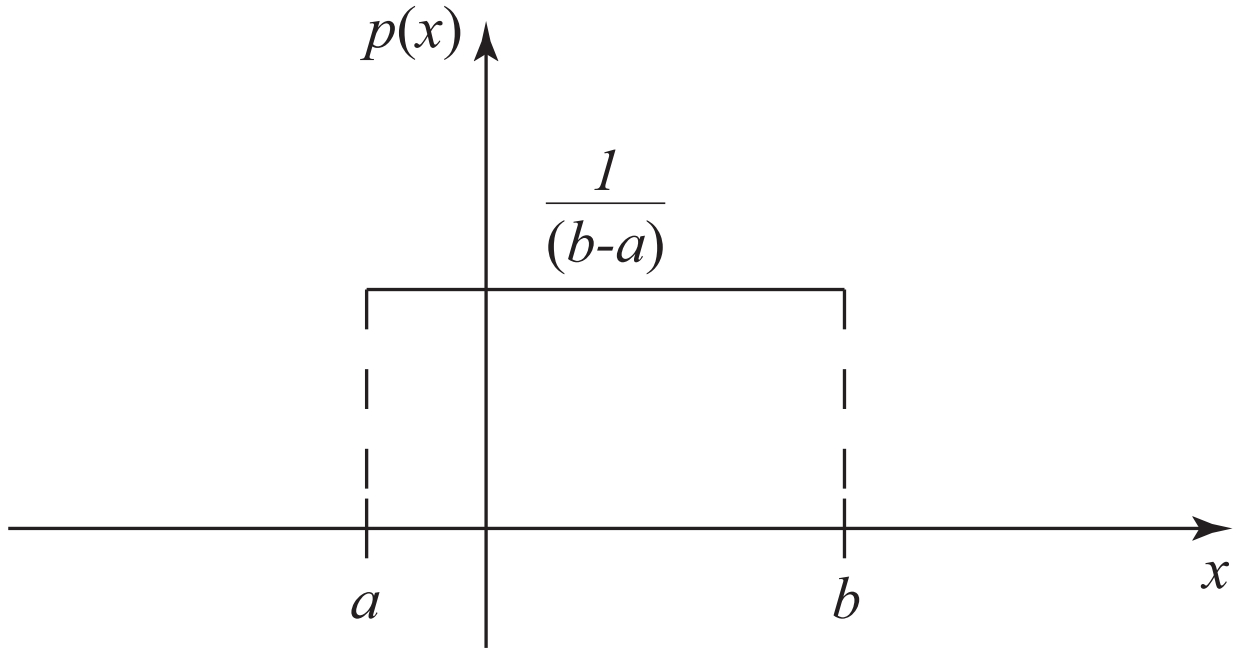

Figure 1: Plot of the pdf $p(x)$ of a uniform distribution.

Gaussian (normal) distribution.

The pdf of a Gaussian or normally distributed random variable is given by

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \tag{18}$$

where μ is the mean and σ^2 is the variance of the random variable. The pdf of a Gaussian distributed random variable is illustrated in Figure 2. The probability distribution function $F(x)$ has the form

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} p(s)\, ds \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-(s-\mu)^2/(2\sigma^2)}\, ds \\ &= \frac{1}{2}\left[1 + \mathrm{erf}(\frac{x-\mu}{\sqrt{2}\sigma})\right], \end{aligned} \tag{19}$$

where erf denotes the error function and is given by

$$\mathrm{erf}(x) = \frac{2}{\pi} \int_0^x e^{-s^2}\, ds. \tag{20}$$

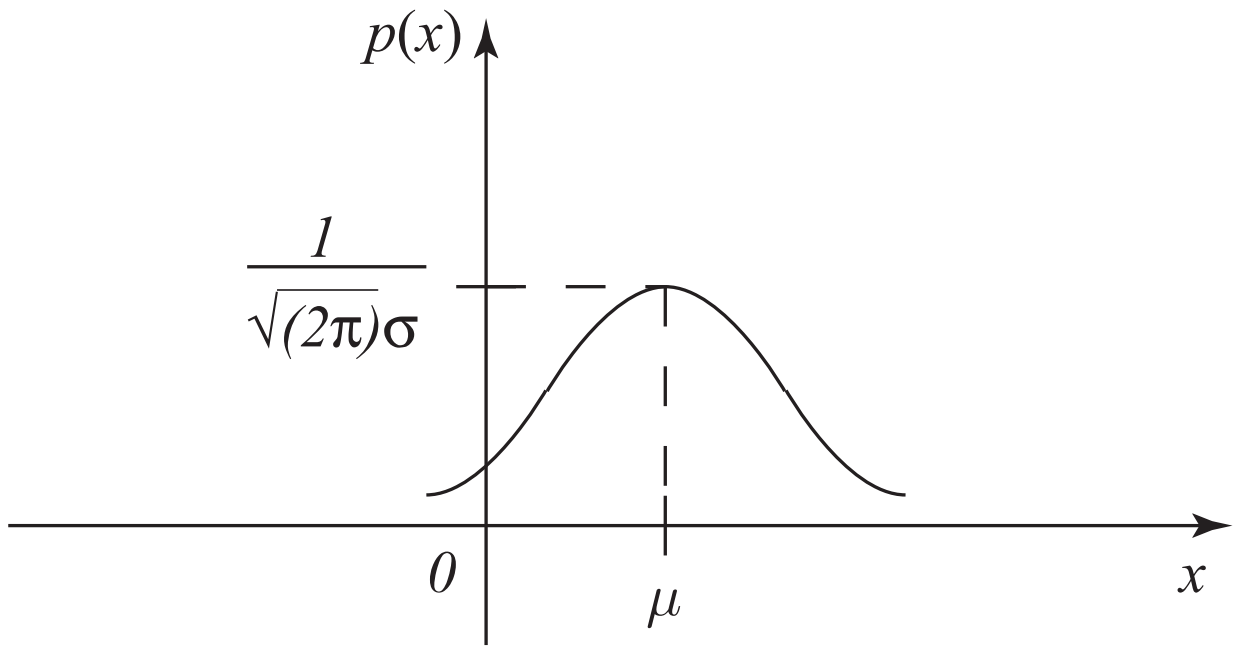

Figure 2: The pdf graph of a Gaussian distributed random variable.

The kth central moment of the random variable X is given by the expression

$$E[(X-\mu)^k] \equiv m_k = \begin{cases} 1 \cdot 3 \cdots (k-1)\sigma^k, & k = \text{even}, \\ 0, & k = \text{odd} \end{cases} \tag{21}$$

and the kth moments are given in terms of the central moments by

$$E(X^k) = \sum_{i=0}^{k} \begin{pmatrix} k \\ i \end{pmatrix} \mu^i m_{k-i}. \tag{22}$$

Finally, if X is a random variable distributed normally with mean μ and variance σ^2, this is commonly denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$.

Log-normal distribution.

A widely employed model for biological (and other) phenomena where the random variable is only allowed positive values is the so-called *log normal* distribution. If $\log X \sim \mathcal{N}(\mu, \sigma^2)$, then X has a lognormal distribution with density given by

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x}\exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}, \quad 0 < x < \infty, \tag{23}$$

with mean and variance

$$E(X) = e^{\mu+\sigma^2/2}, \tag{24}$$

$$\mathrm{var}(X) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}. \tag{25}$$

We observe that $\mathrm{var}(X)$ is proportional to $\{E(X)\}^2$ so that the constant *coefficient of variation (CV)* defined by $\sqrt{\mathrm{var}(X)}/E(X)$, which represents the "*noise-to-signal*" ratio, does not depend on

$E(X)$. The density for this random variable is skewed (asymmetric) with a "*long right tail*" but becomes more and more symmetric as $\sigma \to 0$.

Multivariate normal distribution.

One of the most often encountered multivariate random variables is (as we shall see below in discussing asymptotic theory for confidence intervals) also incredibly important in statistical *modeling* and *inference* and is known as the *multivariate normal* or *multinormal* random variable. A *random vector* $X = (X_1, \ldots, X_p)^T$ has a multivariate (p-variate) normal distribution (denoted by $X \sim \mathcal{N}_p(\mu, \Sigma)$) if $\alpha^T X$ is normal for all $\alpha \in R^p$; its density is given by

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

for $x = (x_1, \ldots, x_p)^T \in R^p$ where the mean is

$$\mu = E(X) = (\mu_1, \ldots, \mu_p)^T = \{E(X_1), \ldots, E(X_p)\}^T$$

and the *covariance matrix* is

$$\Sigma = E\{(x-\mu)(x-\mu)^T\}.$$

The $(p \times p)$ covariance matrix Σ is such that

$$\Sigma_{jj} = \mathrm{var}(X_j), \quad \Sigma_{jk} = \Sigma_{kj} = \mathrm{cov}(X_j, X_k).$$

Finally, we note that the *marginal* probability densities are *univariate* normal.

Chi-square distribution.

The chi-square distribution is important in statistical analysis of variance (ANOVA) and other statistical procedures [7, 10] based on normally distributed random variables. In particular, a chi-square distributed random variable is related to the normally distributed random variable through a transformation. That is, if X is a normally distributed random variable, then $Y = X^2$ has a chi-square distribution. There are two types of chi-square distributions. A central chi-square distribution is obtained when X has zero mean; otherwise, we call it a non-central chi-square distribution.

First, let us consider the central chi-square distribution. In this case, the pdf of Y has the form

$$p(y) = \frac{1}{\sqrt{2\pi y}\sigma} e^{-y/(2\sigma^2)}, \tag{26}$$

where $y > 0$. The corresponding probability distribution function $F(y)$ is given by

$$F(y) = \frac{1}{\sqrt{2\pi}\sigma} \int_0^y \frac{1}{\sqrt{s}} e^{-s/(2\sigma^2)}\, ds. \tag{27}$$

More generally, suppose that the random variable Y is defined as

$$Y = \sum_{i=1}^{k} X_i^2, \tag{28}$$

where X_i, $i = 1, 2, \ldots, k$, are statistically independent and identically distributed normal random variables with zero mean and variance σ^2. The pdf is then given by

$$p(y) = \frac{1}{\sigma^k 2^{k/2} \Gamma(k/2)} y^{k/2-1} e^{-y/(2\sigma^2)}, \qquad y > 0, \tag{29}$$

where $\Gamma(p)$ is the gamma function defined as $\Gamma(1/2) = \sqrt{\pi}$,

$\Gamma(3/2) = \sqrt{\pi}/2$, and for $\nu > 0$,

$$\Gamma(\nu) = \int_0^\infty x^{\nu-1} e^{-x}\, dx. \tag{30}$$

By integration by parts, it can be shown that

$$\Gamma(\nu) = (\nu - 1)\Gamma(\nu - 1). \tag{31}$$

For positive and integer ν we obtain

$$\Gamma(\nu) = (\nu - 1)!. \tag{32}$$

This pdf, which is a generalization of (26), is called a *chi-square (or gamma) pdf with k degrees of freedom* (denoted $Y \sim \chi^2(k)$ or $Y \sim \chi^2_k$). Its graphs for several values of k are depicted in Figure 3.

The first two moments of Y are

$$\begin{aligned} E(Y) &= k\sigma^2 \\ E(Y^2) &= 2k\sigma^4 + k^2\sigma^4 \end{aligned}$$

and its variance is

$$\mathrm{var}(Y) = 2k\sigma^4. \tag{33}$$

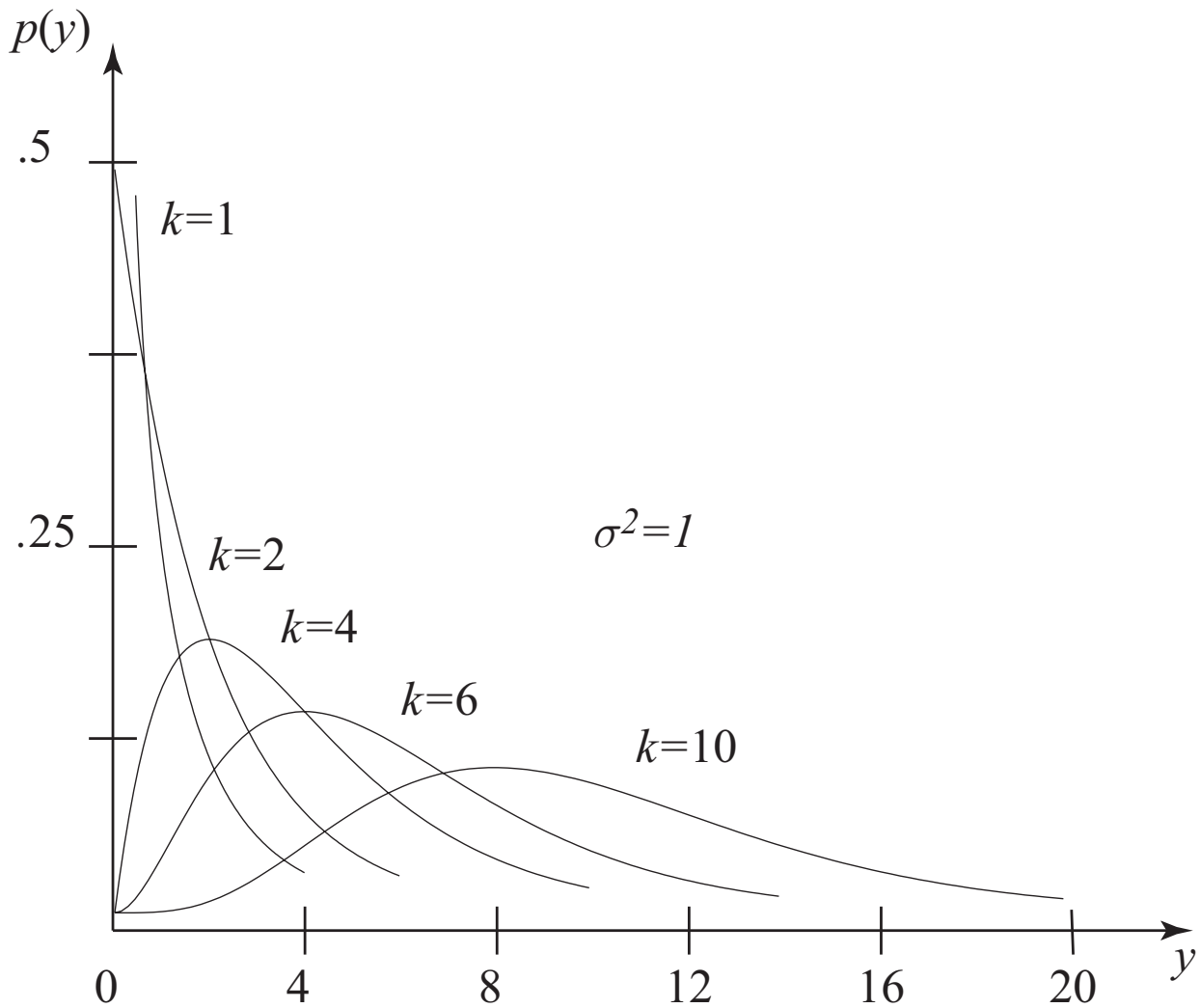

Figure 3: The pdf graph of a chi-square distribution for various degrees of freedom k.

We now turn to the non-central chi-square distribution. Here, let X_i, $i = 1, 2, \ldots, k$, be Gaussian distributed random variables with means μ_i and identical variances equal to σ^2. The random variable $Y = \sum_i^k X_i^2$ has the pdf

$$p(y) = \frac{1}{2\sigma^2} \left(\frac{y}{s}\right)^{(k-2)/4} e^{-(s^2+y)/(2\sigma^2)} I_{k/2-1}\left(\frac{\sqrt{y}s}{\sigma^2}\right), \qquad y > 0, \tag{34}$$

where the parameter s^2, which is called the noncentrality parameter of the distribution, is given by

$$s^2 = \sum_{i=1}^{k} \mu_i^2.$$

The function $I_\alpha(x)$ is the αth-order modified Bessel function of the

first kind and is given by

$$I_\alpha(x) = \sum_{j=0}^{\infty} \frac{(x/2)^{\alpha+2j}}{j!\,\Gamma(\alpha+j+1)}, \qquad x \geq 0. \tag{35}$$

The pdf function given by the expression (34) is called the *non-central chi-square pdf with k degrees of freedom*.

Finally, the first two moments of the non-central chi-square distribution random variable are

$$\begin{aligned} E(Y) &= k\sigma^2 + s^2 \\ E(Y^2) &= 2k\sigma^4 + 4\sigma^2 s^2 + (k\sigma^2 + s^2)^2 \end{aligned}$$

and its variance is

$$\mathrm{var}(Y) = 2k\sigma^4 + 4\sigma^2 s^2.$$

Rayleigh distribution.

Another frequently encountered random variable which is closely related to the central chi-square distribution is the Rayleigh distribution . To begin the discussion, let us consider a central chi-square distribution with two degrees of freedom, $Y = X_1^2 + X_2^2$, where X_i are zero mean statistically independent Gaussian random variables with identical variances σ^2. The pdf of Y is given by

$$p(y) = \frac{1}{2\sigma^2} e^{-y/(2\sigma^2)}. \tag{36}$$

Define a new variable Z as

$$Z = \sqrt{X_1^2 + X_2^2} = \sqrt{Y}.$$

Then, after a change of variables in equation (36), we obtain the pdf of Z as

$$p(z) = \frac{z}{\sigma^2} e^{-z^2/(2\sigma^2)}, \qquad z \geq 0$$

which is known as the pdf of a Rayleigh distributed random variable. The corresponding probability distribution function is given by

$$\begin{aligned} F(z) &= \int_0^z \frac{s}{\sigma^2} e^{-s^2/(2\sigma^2)}\, ds \\ &= 1 - e^{-z^2/(2\sigma^2)}, \qquad z \geq 0. \end{aligned}$$

The moments of Z are

$$E(Z^k) = (2\sigma^2)^{(k/2)} \Gamma(1 + \frac{k}{2})$$

and the variance is given by

$$\mathrm{var}(Z) = (2 - \frac{\pi}{2})\sigma^2.$$

Student's t distribution.

If $U \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(k)$ are independent, then $X = U/\sqrt{V/k}$ has a t distribution with k degrees of freedom (denoted by $X \sim t^k$) and density function

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

The mean and variance are given by

$$E(X) = 0 \quad \text{if} \quad k > 1 \text{ (otherwise undefined)}$$

and

$$\text{var}(X) = k/(k-2) \quad \text{if} \quad k > 2 \text{ (otherwise undefined)}.$$

The corresponding density is symmetric like that of the normal, with "*heavier tails*", and becomes similar to a normal as $k \to \infty$. As we

shall see below, the Student's t distribution is fundamental to the computation of confidence intervals for estimated parameters using experimental data in inverse problems (where typically $k \gg 2$).

References

[1] H. T. Banks and H. T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, Boca Raton, FL, 2009.

[2] H.T. Banks, M. Davidian, J.R. Samuels, Jr., and K.L. Sutton, An Inverse Problem Statistical Methodology Summary, *CRSC-TR08-01*, January, 2008; Chapter 11 in *Statistical Estimation Approaches in Epidemiology*, (edited by Gerardo Chowell, Mac Hyman, Nick Hengartner, Luis M.A Bettencourt and Carlos Castillo-Chavez), Springer, Berlin Heidelberg New York, 2009, pp. 249–302.

[3] H.T. Banks, S. Dediu and S.E. Ernstberger, Sensitivity functions and their uses in inverse problems, *J. Inverse and Ill-posed Problems*, **15**, 2007, pp. 683-708.

[4] H.T. Banks, S.L. Ernstberger and S.L. Grove, Standard errors and confidence intervals in inverse problems: Sensitivity and associated pitfalls, CRSC-TR06-10, March, 2006; *J. Inv. Ill-posed Problems*, **15**, 2006, pp. 1-18.

[5] H.T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhäuser, Boston, 1989.

[6] R.J. Carroll and D. Ruppert, *Transformation and Weighting in Regression,* Chapman & Hall, New York, 1988.

[7] G. Casella and R. L. Berger, *Statistical Inference,* Duxbury, California, 2002.

[8] M. Davidian and D. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman & Hall, London, 1998.

[9] A. R. Gallant, *Nonlinear Statistical Models*, Wiley, New York, 1987.

[10] F. Graybill, *Theory and Application of the Linear Model*, Duxbury, North Scituate, MA, 1976.

[11] G.A.F. Seber and C.J. Wild, *Nonlinear Regression*, J. Wiley & Sons, Hoboken, NJ, 2003.