prepared by Professor Jenny Baglivo

1 MATH448001 Notebook 1

This notebook reviews concepts from probability theory and mathematical statistics. The notes include topics from Chapters 1–5 (probability theory), Chapter 6 (distributions derived from the normal distribution), Chapter 8 (estimation theory), Chapter 9 (hypothesis testing), and Chapter 11 (comparing two samples) of the Rice textbook.

1.1 Review of Probability Theory

This section reviews fundamental concepts from probability theory. Concepts from probability theory are essential to understanding statistical inference.

1.1.1 Random Variable, Range, Types of Random Variables

A *random variable* is a function from the sample space of an experiment to the real numbers. The *range* of a random variable is the set of values the random variable assumes.

Random variables are usually denoted by capital letters (X, Y, Z, $\ldots$) and their values by lower case letters (x, y, z, $\ldots$). The letter $\mathcal{R}$ is often used to denote the range.

If the range of a random variable is a finite or countably infinite set, then the random variable is said to be *discrete*; if the range is an interval or a union of intervals, then the random variable is said to be *continuous*; otherwise, the random variable is said to be of *mixed* type.

For example, if X is the height of an individual measured in inches with infinite precision, then X is a continuous random variable whose range is the positive real numbers: $\mathcal{R} = (0, \infty)$.

1.1.2 CDF, PDF, Quantiles

The *cumulative distribution function (CDF)* of the random variable X is defined as follows:

$$F(x) = P(X \leq x) \text{ for all real numbers } x.$$

- If X is a discrete random variable, then $F(x)$ is a step function.
- If X is a continuous random variable, then $F(x)$ is a continuous function.

The *probability density function (PDF)* of the random variable X is defined as follows:

1. *Discrete:* If X is a discrete random variable, then the PDF is a probability:
$$p(x) = P(X = x) \text{ for all real numbers } x.$$
2. *Continuous:* If X is a continuous random variable, then the PDF is a rate:
$$f(x) \;=\; \frac{d}{dx}F(x) \quad \text{whenever the derivative exists.}$$

Let X be a continuous random variable. Then

1. The p^{th} *quantile* (or $100p^{\text{th}}$ *percentile*), x_p, is the point satisfying the equation $P(X \leq x_p) = p$. To find the p^{th} quantile, solve the equation $F(x) = p$ for x.

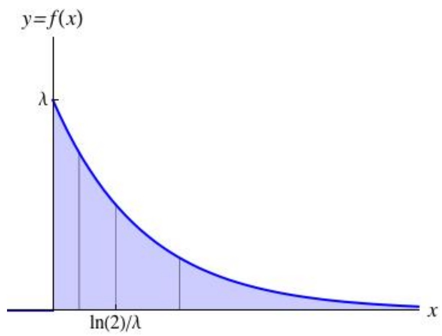
2. The *median* of X is the 50^{th} percentile, $\text{Med}(X) = x_{1/2}$.

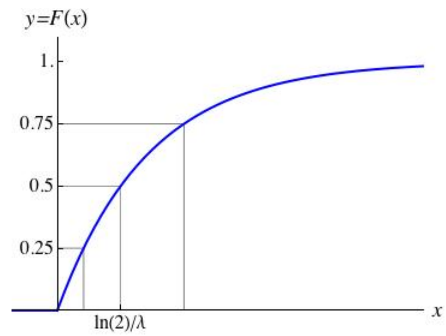
3. The *interquartile range (IQR)* of X is the length of the interval from the 25^{th} percentile to the 75^{th} percentile, $\text{IQR}(X) = x_{3/4} - x_{1/4}$.

For example, let λ be a positive real number and let X be the exponential random variable with parameter λ. Then

$$f(x) = \lambda e^{-\lambda x} \text{ when } x \in (0, \infty); \text{ and } 0 \text{ otherwise.}$$

The PDF is shown in the *left plot* below, and the CDF is shown in the *right plot*.





Vertical lines indicate the locations of the 25^{th}, 50^{th} and 75^{th} percentiles.

1. *To find the CDF:* Since the range of X is the positive reals, $F(x) = 0$ when $x \leq 0$.

   Given $x > 0$,
   $$F(x) = \int_0^x \lambda e^{-\lambda t}\, dt = \Big[-e^{-\lambda t} \Big]_0^x = -e^{-\lambda x} + 1.$$
   Thus, $F(x) = 1 - e^{-\lambda x}$ when $x > 0$, and $F(x) = 0$ otherwise.

2. *To find a general formula for the pth quantile:* Since
   $$1 - e^{-\lambda x} = p \;\Rightarrow\; 1 - p = e^{-\lambda x} \;\Rightarrow\; x = -\frac{\ln(1-p)}{\lambda},$$
   a general formula for the p^{th} quantile of X is $x_p = -\frac{\ln(1-p)}{\lambda}$ when $p \in (0, 1)$.

3. *To find formulas for the median and interquartile range:* Using the general formula,
   $$Med(X) = -\frac{\ln(1/2)}{\lambda} = \frac{\ln(2)}{\lambda} \quad\text{and}\quad IQR(X) = -\frac{\ln(1/4)}{\lambda} + \frac{\ln(3/4)}{\lambda} = \frac{\ln(3)}{\lambda},$$
   where I have used properties of logarithms to simplify each formula.

1.1.3 Expected Values, Mean, Variance, Standard Deviation

Let $g(X)$ be a real-valued function of the random variable X.

1. *Discrete Case:* If X is a discrete random variable with range $\mathcal{R}$ and PDF $p(x)$, then the *expected value* of $g(X)$ is defined as follows:

$$E(g(X)) = \sum_{x \in \mathcal{R}} g(x)\ p(x),$$

   as long as $\sum_{x \in \mathcal{R}} |g(x)|\ p(x)$ converges. If the sum does not converge absolutely, then the expected value is said to be *indeterminate*.

2. *Continuous Case:* If X is a continuous random variable with range $\mathcal{R}$ and PDF $f(x)$, then the *expected value* of $g(X)$ is defined as follows:

$$E(g(X)) = \int_{\mathcal{R}} g(x)\ f(x)\ dx,$$

   as long as $\int_{\mathcal{R}} |g(x)|\ f(x)\ dx$ converges. If the integral does not converge absolutely, then the expected value is said to be *indeterminate*.

The following summary measures are defined using expectations:

Summary Measure:	*Notation and Definition:*
Mean of X, Expected Value of X, or Expectation of X	$\mu = E(X)$
Variance of X	$\sigma^2 = Var(X) = E((X-\mu)^2)$
Standard Deviation of X	$\sigma = SD(X) = \sqrt{Var(X)}$

The mean is a measure of the *center* of a probability distribution, and the variance and standard deviation are measures of the *spread* of the distribution.

Properties of expectations: Properties of sums and integrals imply the following useful properties of expectations.

1. $E(a) = a$, where a is a constant.
2. $E(a + bg(X)) = a + bE(g(X))$, where a and b are constants.
3. $E(c_1 g_1(X) + c_2 g_2(X)) = c_1 E(g_1(X)) + c_2 E(g_2(X))$, where c_1 and c_2 are constants.
4. $Var(X) = E(X^2) - \mu^2$, where $\mu = E(X)$.
5. If $Y = a + bX$ where a and b are constants, then

$$E(Y) = a + bE(X),\ Var(Y) = b^2 Var(X) \text{ and } SD(Y) = |b| SD(X).$$

Property 4 is especially useful for finding variances in new situations. It can be proven using the first 3 properties as follows:

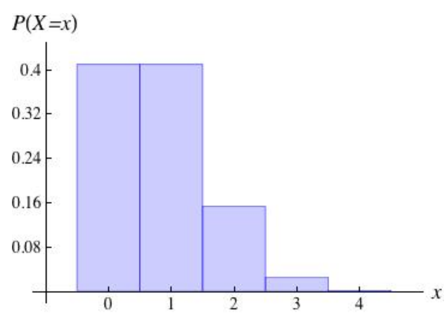
$$\begin{aligned} Var(X) &= E\left((X-\mu)^2\right) && \text{(definition of variance)} \\ &= E\left(X^2 - 2\mu X + \mu^2\right) && \text{(using polynomial expansion)} \\ &= E(X^2) - 2\mu E(X) + \mu^2 && \text{(using the first 3 properties)} \\ &= E(X^2) - 2\mu^2 + \mu^2 && \text{(since } \mu = E(X)\text{)} \\ &= E(X^2) - \mu^2 && \text{(the expression in property 4)} \end{aligned}$$

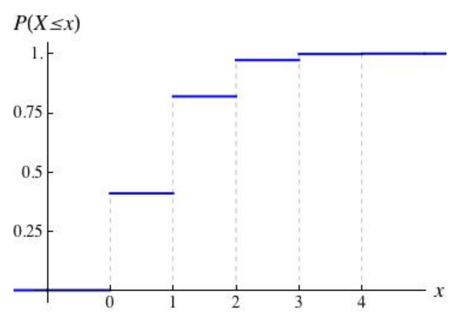
To illustrate these ideas, let X be be the number of successes in 4 independent trials of a Bernoulli experiment with success probability $\frac{1}{5}$.

Then X has a binomial distribution. The PDF of X is

$$p(x) = \binom{4}{x}\left(\frac{1}{5}\right)^x\left(\frac{4}{5}\right)^{4-x} \text{ when } x = 0, 1, 2, 3, 4; \text{ and 0 otherwise,}$$

as shown on the *left*, and the CDF of X is the step function shown on the *right*.





The PDF of X is displayed as a *probability histogram.* The rectangle whose base is centered at x has area $p(x)$; the sum of the areas of the rectangles is 1.

Since

1. $E(X) = 0 \times 0.4096 + 1 \times 0.4096 + 2 \times 0.1536 + 3 \times 0.0256 + 4 \times 0.0016 = 0.80$ and
2. $E(X^2) = 0^2 \times 0.4096 + 1^2 \times 0.4096 + 2^2 \times 0.1536 + 3^2 \times 0.0256 + 4^2 \times 0.0016 = 1.28$,

we know that

$$Var(X) = E(X^2) - (E(X))^2 = 1.28 - (0.80)^2 = 0.64 \text{ and } SD(X) = \sqrt{0.64} = 0.80.$$

Note that common discrete probability distributions are given in Section 1.3.1 (page 37), common continuous probability distributions are given in Section 1.3.2 (page 39), and summaries of these models are given in Section 1.3.3 (page 41).

1.1.4 Joint, Marginal and Conditional Distributions; Independence

Let X be a random variable with range $\mathcal{R}_x$, and Y be a random variable with range $\mathcal{R}_y$.

1. <u>*Discrete Case:*</u> If X and Y are discrete, then

   (a) *Joint PDF:* The *joint PDF* of the random pair (X, Y) is defined as follows:

   $$p(x, y) = P(X = x, Y = y), \quad \text{for all } (x, y) \in \mathbf{R}^2,$$

   where the comma on the right is understood to mean the intersection of events.

   (b) *Marginal PDFs:* Let $p(x, y)$ be the joint PDF of the random pair (X, Y). Then

   - the *marginal PDF* of X is

   $$p_X(x) = P(X = x) = \sum_{y \in \mathcal{R}_y} p(x, y) \quad \text{for all real numbers } x; \text{ and}$$

   - the *marginal PDF* of Y is

   $$p_Y(y) = P(Y = y) = \sum_{x \in \mathcal{R}_x} p(x, y) \quad \text{for all real numbers } y.$$

   (c) *Conditional PDFs:* Let $p(x, y)$ be the joint PDF of the random pair (X, Y). Then

   - if $p_X(x) \neq 0$, then the *conditional PDF* of Y given $X = x$ is

   $$p_{Y|X=x}(y|x) = P(Y = y|X = x) = \frac{p(x, y)}{p_X(x)} \quad \text{for all real numbers } y; \text{ and}$$

   - if $p_Y(y) \neq 0$, then the *conditional PDF* of X given $Y = y$ is

   $$p_{X|Y=y}(x|y) = P(X = x|Y = y) = \frac{p(x, y)}{p_Y(y)} \quad \text{for all real numbers } x.$$

   (d) *Joint PDF as a Product:* The joint PDF of a random pair can be written as the product of a marginal PDF and a conditional PDF. Specifically, if $p(x, y)$ is the joint PDF of the random pair (X, Y), then

   $$p(x, y) = p_X(x) p_{Y|X=x}(y|x) \quad \text{when } p_X(x) \neq 0, \text{ and 0 otherwise; or}$$

   $$p(x, y) = p_Y(y) p_{X|Y=y}(x|y) \quad \text{when } p_Y(y) \neq 0, \text{ and 0 otherwise.}$$

   (Writing the joint PDF as a product is often quite useful in applications.)

   (e) *Independence:* Let $p(x, y)$ be the joint PDF of the random pair (X, Y). Then X and Y are said to be *independent* if

   $$p(x, y) = p_X(x) p_Y(y) \quad \text{for all } (x, y) \in \mathbf{R}^2.$$

   (The discrete random variables are independent when the probability of the intersection is equal to the product of the probabilities for events of the form "$X = x$" and "$Y = y$".)

2. <u>*Continuous Case:*</u> If X and Y are continuous, then

   (a) *Joint CDF:* The *joint CDF* of the random pair (X, Y) is defined as follows:

   $$F(x, y) = P(X \le x, Y \le y), \quad \text{for all } (x, y) \in \mathbf{R}^2,$$

   where the comma on the right is understood to mean the intersection of events.

   (b) *Joint PDF:* If $F(x, y)$ has continuous second order partial derivatives, then the *joint PDF* is defined as follows:

   $$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y) \ \text{ for all possible } (x, y).$$

   (c) *Marginal PDFs:* Let $f(x, y)$ be the joint PDF of the random pair (X, Y). Then

   - the *marginal PDF* of X is

     $$f_X(x) = \int_{\mathcal{R}_y} f(x, y)\, dy \ \text{ for all real numbers } x\text{; and}$$

   - the *marginal PDF* of Y is

     $$f_Y(y) = \int_{\mathcal{R}_x} f(x, y)\, dx \ \text{ for all real numbers } y.$$

   (d) *Conditional PDFs:* Let $f(x, y)$ be the joint PDF of the random pair (X, Y). Then

   - if $f_X(x) \ne 0$, then the *conditional PDF* of Y given $X = x$ is

     $$f_{Y|X=x}(y|x) = \frac{f(x, y)}{f_X(x)} \ \text{ for all real numbers } y\text{; and}$$

   - if $f_Y(y) \ne 0$, then the *conditional PDF* of X given $Y = y$ is

     $$f_{X|Y=y}(x|y) = \frac{f(x, y)}{f_Y(y)} \ \text{ for all real numbers } x.$$

   (e) *Joint PDF as a Product:* The joint PDF of a random pair can be written as the product of a marginal PDF and a conditional PDF. Specifically, if $f(x, y)$ is the joint PDF of the random pair (X, Y), then

   $$f(x, y) = f_X(x) f_{Y|X=x}(y|x) \ \text{ when } f_X(x) \ne 0\text{, and 0 otherwise; or}$$

   $$f(x, y) = f_Y(y) f_{X|Y=y}(x|y) \ \text{ when } f_Y(y) \ne 0\text{, and 0 otherwise.}$$

   (Writing the joint PDF as a product is often quite useful in applications.)

   (f) *Independence:* X and Y are said to be *independent* if

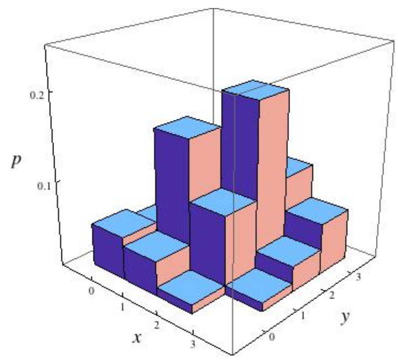
   $$f(x, y) = f_X(x) f_Y(y) \ \text{ for all possible } (x, y).$$

   (The continuous random variables are independent when the joint density can be written as the product of the marginal densities for all pairs for which the functions are defined.)

Example (Olkin et al, Macmillan, 1994). In an experiment to study the relationship between spatial perception and the ability to use a graphical method to solve algebra problems, subjects were asked to solve 3 puzzles and 3 algebra problems. Let X be the number of puzzles solved, and let Y be the number of algebra problems solved.

The following table summarizes the results, using proportions of individuals in each cross-classification:

	$y = 0$	$y = 1$	$y = 2$	$y = 3$	*sum:*
$x = 0$	0.05	0.04	0.01		0.10
$x = 1$	0.04	0.16	0.10	0.10	0.40
$x = 2$	0.01	0.09	0.20	0.10	0.40
$x = 3$		0.01	0.03	0.06	0.10
sum:	0.10	0.30	0.34	0.26	1.00



Consider the experiment "Choose a name from the population of subjects and record

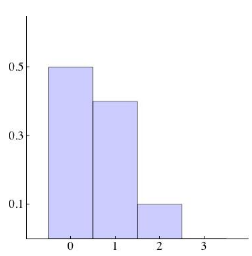
$$(X, Y) \;=\; (\#\text{ puzzles solved, }\#\text{ algebra problems solved}),\text{"}$$

and assume each choice is equally likely. Then

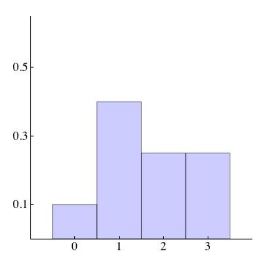
- the body of the table gives the joint (X, Y) distribution,
- the rightmost column of the table gives the marginal X distribution,
- the bottom row of the table gives the marginal Y distribution, and
- the plot on the right above gives the *joint probability histogram* of the random pair. The joint probability histogram is constructed using a box whose base is a square of area 1 centered at (x, y), and whose height is $p(x, y)$, for each pair with nonzero probability.

  The sum of the volumes of the boxes is 1.

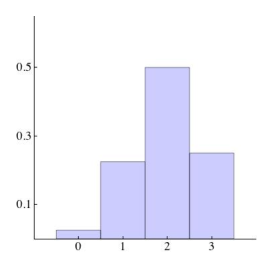
Here are probability histograms for the conditional distributions of Y given $X = x$, for each x:



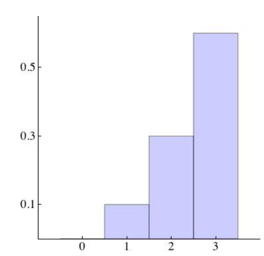
$Y|X = 0$



$Y|X = 1$



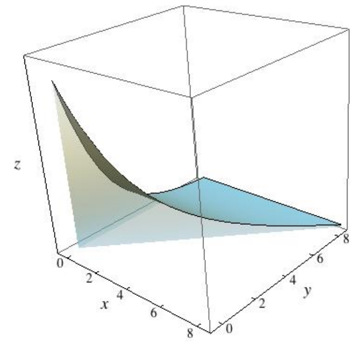
$Y|X = 2$



$Y|X = 3$

As x increases, the conditional distribution of Y given $X = x$ changes, with more weight being given to larger values of Y. (I will let you figure out the probabilities in each case.)

Example. Let (X,Y) be the continuous random pair with joint density function



$$f(x,y) = \frac{1}{4}e^{-y/2} \quad \text{when } 0 < x < y,$$

and 0 otherwise. The joint range of the random pair is

$$\begin{aligned} \mathcal{R} &= \{(x,y) | 0 < x < \infty,\ x < y < \infty\} \\ &= \{(x,y) | 0 < y < \infty,\ 0 < x < y\} \subset \mathbf{R}^2. \end{aligned}$$

The marginal and conditional PDFs are as follows:

- The range of X is the positive real numbers. Given $x > 0$,

$$f_X(x) = \int_x^\infty f(x,y)\ dy = \Big[-\frac{1}{2}e^{-y/2} \Big]_x^{y\to\infty} = 0 + \frac{1}{2}e^{-x/2}.$$

  Thus, the marginal PDF of X is $f_X(x) = \frac{1}{2}e^{-x/2}$ when $x \in (0,\infty)$, and 0 otherwise.

  (X has an exponential distribution.)

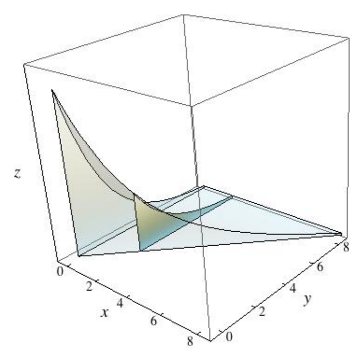
- The range of Y is the positive real numbers. Given $y > 0$,

$$f_Y(y) = \int_0^y f(x,y)\ dx = \Big[\frac{1}{4}e^{-y/2}x\Big]_0^y = \frac{1}{4}ye^{-y} - 0.$$

  Thus, the marginal PDF of Y is $f_Y(y) = \frac{1}{4}ye^{-y/2}$ when $y \in (0,\infty)$, and 0 otherwise.

  (Y has a gamma distribution.)

- Given $x > 0$, the range of Y given $X = x$ is the interval (x,∞), and the conditional PDF is

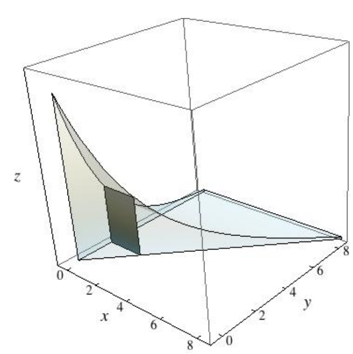


$$f_{Y|X=x}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{2}e^{-(y-x)/2}$$

  when $y \in (x,\infty)$, and equals 0 otherwise.

  (Y given $X = x$ has a shifted exponential distribution.)

- Given $y > 0$, the range of X given $Y = y$ is the interval $(0,y)$, and the conditional PDF is



$$f_{X|Y=y}(x|y) = \frac{f(x,y)}{f_X(x)} = \frac{1}{y}$$

  when $x \in (0,y)$, and equals 0 otherwise.

  (X given $Y = y$ has a uniform distribution.)

1.1.5 Expected Values of Functions of Random Pairs, Covariance, Correlation

Let $g(X,Y)$ be a real-valued function of the random pair (X,Y).

1. <u>*Discrete Case:*</u> If X and Y are discrete random variables with joint range $\mathcal{R} \subseteq \mathbf{R}^2$ and joint PDF $p(x,y)$, then the *expected value* of $g(X,Y)$ is defined as follows:

$$E(g(X,Y)) = \sum_{(x,y)\in\mathcal{R}} g(x,y)\ p(x,y),$$

   as long as $\sum_{(x,y)\in\mathcal{R}} |g(x,y)|\ p(x,y)$ converges. If the sum does not converge absolutely, then the expected value is said to be *indeterminate*.

2. <u>*Continuous Case:*</u> If X and Y are continuous random variables with joint range $\mathcal{R} \subseteq \mathbf{R}^2$ and joint PDF $f(x,y)$, then the *expected value* of $g(X,Y)$ is defined as follows:

$$E(g(X,Y)) = \iint_{\mathcal{R}} g(x,y)\ f(x,y)\ dA,$$

   as long as $\iint_{\mathcal{R}} |g(x,y)|\ f(x,y)\ dA$ converges. If the integral does not converge absolutely, then the expected value is said to be *indeterminate*.

The following summary measures are defined using expectations:

Summary Measure:	*Notation and Definition:*
Covariance of X and Y	$\sigma_{xy} = Cov(X,Y) = E((X-\mu_x)(Y-\mu_y))$ where $\mu_x = E(X)$ and $\mu_y = E(Y)$.
Correlation of X and Y	$\rho = Corr(X,Y) = \dfrac{\sigma_{xy}}{\sigma_x\sigma_y} = \dfrac{Cov(X,Y)}{SD(X)SD(Y)} = \dfrac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ where $\sigma_x = SD(X)$ and $\sigma_y = SD(Y)$.

Properties of expectations: Properties of sums and integrals imply the following useful properties of expectations.

1. If $g_1(X,Y)$ and $g_2(X,Y)$ are real-valued functions, and a, b_1, b_2 are constants, then

$$E(a + b_1 g_1(X,Y) + b_2 g_2(X,Y)) = a + b_1 E(g_1(X,Y)) + b_2 E(g_2(X,Y)).$$

2. If X and Y are independent, and $g(X)$ and $h(Y)$ are real-valued functions, then

$$E(g(X)h(Y)) = E(g(X))E(h(Y)).$$

3. $Cov(X,Y) = Cov(Y,X)$, and $Cov(X,X) = Var(X)$.

4. $Cov(X,Y) = E(XY) - E(X)E(Y)$.

5. $|Corr(X,Y)| \leq 1$. Further, $|Corr(X,Y)| = 1$ if and only if $Y = a + bX$, except possibly on a set of probability zero.

6. If X and Y are independent, then $Cov(X,Y) = 0$ and $Corr(X,Y) = 0$.

7. Let a, b, c, and d be constants. Then

   (a) $Cov(a + bX, c + dY) = bdCov(X,Y)$.

   (b) $Corr(a + bX, c + dY) = \begin{cases} Corr(X,Y) & \text{when } bd > 0 \\ -Corr(X,Y) & \text{when } bd < 0 \end{cases}$

Property 4 is especially useful for finding covariances in new situations. It can be proven using the first property as follows:

$$\begin{aligned}
Cov(X,Y) &= E\left((X - \mu_x)(Y - \mu_y)\right) && \text{(definition of covariance)} \\
&= E\left(XY - X\mu_y - \mu_x Y + \mu_x\mu_y\right) && \text{(using polynomial expansion)} \\
&= E(XY) - \mu_y E(X) - \mu_x E(Y) + \mu_x\mu_y && \text{(using the first property)} \\
&= E(XY) - \mu_y\mu_x - \mu_x\mu_y + \mu_x\mu_y && \text{(since } \mu_x = E(X),\ \mu_y = E(Y)\text{)} \\
&= E(XY) - \mu_x\mu_y && \text{(the expression in property 4)}
\end{aligned}$$

Property 7 tells us that the covariance changes under linear transformations, but correlation is unchanged (up to a change of sign). As a general example, let X be the height in feet and Y be the weight in pounds of a randomly chosen individual from a certain population. If we wanted to change the measurement scales so that height was measured in inches and weight in ounces, then the covariance would change since

$$Cov(12X, 16Y) = (12)(16)Cov(X,Y),$$

but the correlation would remain the same since $Corr(12X, 16Y) = Corr(X,Y)$.

Example. An urn contains 5 red, 2 white and 3 blue chips. Let X be the number of red chips and Y be the number of white chips in a subset of size 3 chosen without replacement from the urn, and assume that each choice of subset is equally likely.

Then the joint PDF is

$$p(x,y) = \frac{\binom{5}{x}\binom{2}{y}\binom{3}{3-x-y}}{\binom{10}{3}}$$

when $x = 0,1,2,3,\quad y = 0,1,2,\quad x + y \leq 3$, and $p(x,y) = 0$ otherwise.

	$y = 0$	$y = 1$	$y = 2$	*sum:*
$x = 0$	$\frac{1}{120}$	$\frac{6}{120}$	$\frac{3}{120}$	$\frac{10}{120}$
$x = 1$	$\frac{15}{120}$	$\frac{30}{120}$	$\frac{5}{120}$	$\frac{50}{120}$
$x = 2$	$\frac{30}{120}$	$\frac{20}{120}$		$\frac{50}{120}$
$x = 3$	$\frac{10}{120}$			$\frac{10}{120}$
sum:	$\frac{56}{120}$	$\frac{56}{120}$	$\frac{8}{120}$	1

Further,

- $E(X) = 0\left(\frac{10}{120}\right) + 1\left(\frac{50}{120}\right) + 2\left(\frac{50}{120}\right) + 3\left(\frac{10}{120}\right) = \frac{3}{2}$,
- $E(X^2) = 0^2\left(\frac{10}{120}\right) + 1^2\left(\frac{50}{120}\right) + 2^2\left(\frac{50}{120}\right) + 3^2\left(\frac{10}{120}\right) = \frac{17}{6}$,
- $E(Y) = 0\left(\frac{56}{120}\right) + 1\left(\frac{56}{120}\right) + 2\left(\frac{3}{120}\right) = \frac{3}{5}$,
- $E(Y^2) = 0^2\left(\frac{56}{120}\right) + 1^2\left(\frac{56}{120}\right) + 2^2\left(\frac{3}{120}\right) = \frac{11}{15}$, and
- $E(XY) = 0\left(\frac{65}{120}\right) + 1\left(\frac{30}{120}\right) + 2\left(\frac{25}{120}\right) = \frac{2}{3}$,
- $Var(X) = \left(\frac{17}{6}\right) - \left(\frac{3}{2}\right)^2 = \frac{7}{12}$, $Var(Y) = \left(\frac{11}{15}\right) - \left(\frac{3}{5}\right)^2 = \frac{28}{75}$,
- $Cov(X,Y) = \left(\frac{2}{3}\right) - \left(\frac{3}{2}\right)\left(\frac{3}{5}\right) = -\frac{7}{30}$ and $Corr(X,Y) = \frac{-7/30}{\sqrt{(7/12)(28/75)}} = -\frac{1}{2}$.

Example. Let (X,Y) be the continuous random pair with joint PDF

$$f(x,y) = \frac{2}{9}(y-x) \ \text{ when } 0 < x < y < 3, \text{ and 0 otherwise.}$$

The joint range of the random pair is

$$\mathcal{R} = \{(x,y) | 0 < x < 3,\ x < y < 3\} = \{(x,y) | 0 < y < 3,\ 0 < x < y\} \subset \mathbf{R}^2.$$

- The range of X is the interval $(0,3)$. Given $x \in (0,3)$,

  $$f_X(x) = \int_x^3 \frac{2}{9}(y-x)\ dy = \left[\frac{1}{9}(y-x)^2\right]_x^3 = \frac{1}{9}(x-3)^2; \ \text{ and } f_X(x) = 0 \text{ otherwise.}$$

  Then $E(X) = \frac{3}{4}$, $E(X^2) = \frac{9}{10}$ and $Var(X) = \frac{27}{80}$.
- The range of Y is the interval $(0,3)$. Given $y \in (0,3)$,

  $$f_Y(y) = \int_0^y \frac{2}{9}(y-x)\ dx = \left[-\frac{1}{9}(y-x)^2\right]_0^y = \frac{1}{9}y^2; \ \text{ and } f_Y(y) = 0 \text{ otherwise.}$$

  Then $E(Y) = \frac{9}{4}$, $E(Y^2) = \frac{27}{5}$ and $Var(Y) = \frac{27}{80}$.
- $E(XY) = \int_0^3 \int_0^y xy\frac{2}{9}(y-x)\ dx\ dy = \int_0^3 \frac{1}{27}y^4\ dy = \left[\frac{1}{135}y^5\right]_0^3 = \frac{9}{5}$,
  $Cov(X,Y) = \left(\frac{9}{5}\right) - \left(\frac{3}{4}\right)\left(\frac{9}{4}\right) = \frac{9}{80}$ and $Corr(X,Y) = \frac{9/80}{\sqrt{(27/80)(27/80)}} = \frac{1}{3}$.

(I will let you sketch the joint range, and check the intermediate steps in this example.)

1.1.6 Linear Functions of Independent Random Variables

Let X and Y be independent random variables with finite means and variances. Further, let $W = a + bX + cY$ be a linear function of X and Y, where a, b and c are constants. Then properties of sums and integrals can be used to demonstrate that the mean and variance of W satisfy the following rules:

$$E(W) = a + bE(X) + cE(Y) \ \text{ and } \ Var(W) = b^2Var(X) + c^2Var(Y).$$

(This result tells us about summary measures of W, but does not tell us about the distribution of W.)

The following theorem, which can be proven using moment generating functions, tells us that we can say more about linear functions of independent normal random variables.

> ***Theorem (Independent Normal RVs).*** Let X and Y be independent normal random variables, and let $W = a + bX + cY$ be a linear function, where a, b and c are constants. Then W is a normal random variable with the following mean and variance:
>
> $$E(W) = a + bE(X) + cE(Y) \text{ and } Var(W) = b^2Var(X) + c^2Var(Y).$$
>
> (A linear function of independent normal random variables is a normal random variable.)

For example, suppose that X and Y are independent normal random variables, each with mean 10 and standard deviation 3, and let $W = X - Y$ be their difference. Then W is a normal random variable with the following mean, variance and standard deviation:

$$E(W) = 10 - 10 = 0, \; Var(W) = 3^2 + (-1)^2 3^2 = 18, \; SD(W) = \sqrt{18} = 3\sqrt{2}.$$

Distribution of the sum of independent random variables. Let X and Y be independent random variables and let $W = X + Y$ be their sum.

Although the distribution of W can be hard to find in general, there are certain situations where the distribution is known. The following table gives several important cases:

Distribution of X:	*Distribution of Y:*	*Distribution of* $W = X + Y$ *When X and Y are Independent:*
Bernoulli p	Bernoulli p	Binomial 2, p
Binomial n_1, p	Binomial n_2, p	Binomial $n_1 + n_2$, p
Geometric p	Geometric p	Negative Binomial 2, p
Negative Binomial r_1, p	Negative Binomial r_2, p	Negative Binomial $r_1 + r_2$, p
Poisson λ_1	Poisson λ_2	Poisson $\lambda_1 + \lambda_2$
Exponential λ	Exponential λ	Gamma $\alpha = 2$, $\beta = \frac{1}{\lambda}$
Normal μ_1, σ_1	Normal μ_2, σ_2	Normal $\mu_1 + \mu_2$, $\sqrt{\sigma_1^2 + \sigma_2^2}$

(*Note:* The 1st and 3rd lines are special cases of the 2nd and 4th lines, respectively.)

To illustrate the second line of the table about binomial distributions, let X be the number of successes in 10 independent trials of a Bernoulli experiment with success probability 0.30, let Y be the number of successes in 15 independent trials of a Bernoulli experiment with success probability 0.30, and suppose that X and Y are independent. Then $W = X + Y$ is the number of successes in 25 independent trials of a Bernoulli experiment with success probability 0.30.

1.1.7 Conditional Expectation, Regression

Let X and Y be random variables.

1. <u>*Discrete Case:*</u> Let X and Y be discrete random variables with joint PDF $p(x, y)$.

   (a) If $p_X(x) \neq 0$, then the *conditional expectation* (or the *conditional mean*) of Y given $X = x$ is defined as follows:

   $$E(Y|X = x) = \sum_{y \in \mathcal{R}_{y|x}} y \; p_{Y|X=x}(y|x),$$

   where $\mathcal{R}_{y|x}$ is the conditional range (the collection of y values with $p_{Y|X=x}(y|x) \neq 0$), provided the series converges absolutely.

   (b) If $p_Y(y) \neq 0$, then the *conditional expectation* (or the *conditional mean*) of X given $Y = y$ is defined as follows:

   $$E(X|Y = y) = \sum_{x \in \mathcal{R}_{x|y}} x \; p_{X|Y=y}(x|y),$$

   where $\mathcal{R}_{x|y}$ is the conditional range (the collection of x values with $p_{X|Y=y}(x|y) \neq 0$), provided the series converges absolutely.

2. <u>*Continuous Case:*</u> Let X and Y be continuous random variables with joint PDF $f(x, y)$.

   (a) If $f_X(x) \neq 0$, then the *conditional expectation* (or the *conditional mean*) of Y given $X = x$ is defined as follows:

   $$E(Y|X = x) = \int_{\mathcal{R}_{y|x}} y \; f_{Y|X=x}(y|x) \; dy,$$

   where $\mathcal{R}_{y|x}$ is the conditional range (the collection of y values with $f_{Y|X=x}(y|x) \neq 0$), provided the integral converges absolutely.

   (b) If $f_Y(y) \neq 0$, then the *conditional expectation* (or the *conditional mean*) of X given $Y = y$ is defined as follows:

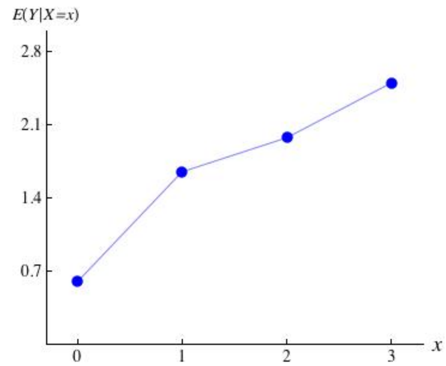
   $$E(X|Y = y) = \int_{\mathcal{R}_{x|y}} x \; f_{X|Y=y}(x|y) \; dx,$$

   where $\mathcal{R}_{x|y}$ is the conditional range (the collection of x values with $f_{X|Y=y}(x|y) \neq 0$), provided the integral converges absolutely.

Regression of Y on X. The formula for the conditional mean $E(Y|X = x)$, as a function of x, is often called the *regression* of Y on X.

Regression of X on Y. The formula for the conditional mean $E(X|Y = y)$, as a function of y, is often called *regression* of X on Y.

Puzzle-algebra example, continued from page 9: The plot on the right shows the average number of algebra problems solved correctly given that a subject solved exactly x puzzles correctly, for $x = 0, 1, 2, 3$.

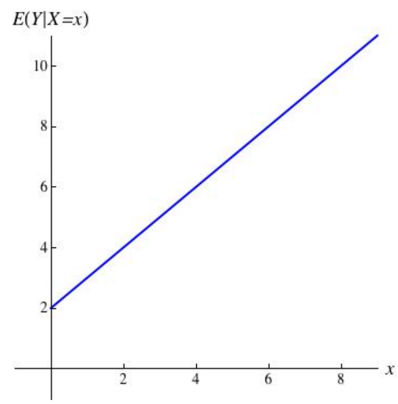


That is, the plot shows pairs $(x, E(Y|X = x))$, where

$$E(Y|X = x) = \sum_{y=0}^{3} y \; p_{Y|X=x}(y|x),$$

for $x = 0, 1, 2, 3$. (I will let you find each value.)

Joint continuous example, continued from page 10: The plot on the right shows the average value of Y for each $x \in (0, \infty)$, for the joint (X, Y) distribution given earlier.



That is, the plot shows pairs $(x, E(Y|X = x))$, where

$$E(Y|X = x) = \int_{x}^{\infty} y \; f_{Y|X=x}(y|x) \; dy = x + 2,$$

for $x \in (0, \infty)$. (When you integrate out y, you get a formula involving x only. I will let you work out the details.)

Positive association: In each example above, X and Y are *positively associated.* As X increases, the expected value of Y given X generally increases as well.

1.1.8 Mutual Independence, Random Samples, Repeated Trials

Let X_1, X_2, ..., X_k be k random variables.

1. *Discrete Case:* If the X_i's are discrete, then
   (a) *Joint PDF:* The *joint PDF* of the random k-tuple $(X_1, X_2, \ldots, X_k)$ is defined as follows
   $$p(x_1, x_2, \ldots, x_k) = P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$$
   for all $(x_1, x_2, \ldots, x_k) \in \mathbf{R}^k$, where commas are understood to mean intersection.
   (b) *Mutual Independence:* The discrete X_i's are said to be *mutually independent* (or *independent* when the context is clear) if
   $$p(x_1, x_2, \ldots, x_k) = p_1(x_1)p_2(x_2) \cdots p_k(x_k)$$
   for all $(x_1, x_2, \ldots, x_k) \in \mathbf{R}^k$, where $p_i(x_i) = P(X_i = x_i)$ for all i.
   (The discrete random variables are independent when the probability of the intersection is equal to the product of the probabilities for all events of the form "$X_i = x_i$".)

2. <u>*Continuous Case:*</u> If the X_i's are continuous, then

   (a) *Joint CDF:* The joint CDF of the random k-tuple $(X_1, X_2, \ldots, X_k)$ is defined as follows:
   $$F(x_1, x_2, \ldots, x_k) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_k \le x_k)$$
   for all k-tuples $(x_1, x_2, \ldots, x_k) \in \mathbf{R}^k$, where commas are understood to mean the intersection of events.

   (b) *Joint PDF:* If the X_i's are continuous and $F(x_1, x_2, \ldots, x_k)$ has continuous k^{th} order partial derivatives, then the *joint PDF* is defined as follows:
   $$f(x_1, x_2, \ldots, x_k) = \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F(x_1, x_2, \ldots, x_k)$$
   for all possible $(x_1, x_2, \ldots, x_k)$.

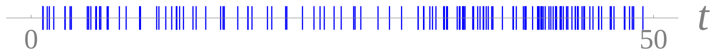
   (c) *Mutual Independence:* If the joint PDF exists, then the continuous X_i's are *mutually independent* (or *independent* when the context is clear) if
   $$f(x_1, x_2, \ldots, x_k) = f_1(x_1) f_2(x_2) \cdots f_k(x_k)$$
   for all possible $(x_1, x_2, \ldots, x_k)$, where $f_i(x_i)$ is the density function of X_i, for all i.
   (The continuous random variables are independent when the joint density can be written as the product of the marginal densities for all k-tuples for which the functions are defined.)

Random samples and repeated trials. Suppose that X_1, X_2, ..., X_k are mutually independent random variables. If the X_i's have a common distribution (that is, if each marginal distribution is the same), then the X_i's are said to be a *random sample* from that distribution.

Consider k repetitions of an experiment, with the outcomes of the trials having no influence on one another, and let X_i be the result of the i^{th} repetition, for $i = 1, 2, \ldots, k$. Then the X_i's are mutually independent, and form a random sample from the common distribution.

To illustrate repeated trials of an experiment, consider the following example.

The graph below shows the times (measured in years from 1947) of the 137 minor-to-light earthquakes (magnitudes 3.5 to 4.4) that occurred in the northeastern United States and eastern Canada between 1947 and 1996. (*Source:* BC Prof. J. Ebel, Weston Observatory.)



Each event is represented by a vertical line located at the time the earthquake occurred.

Geophysicists often use exponential distributions to model the time between successive events. The data pictured above,

$$t_1 = 0.94,\ t_2 = 0.951,\ t_3 = 0.9647,\ \ldots,\ t_{136} = 48.4054,\ t_{137} = 49.1497,$$

yield 136 time differences,

$$d_1 = t_2 - t_1 = 0.011,\ d_2 = t_3 - t_2 = 0.0137,\ \ldots,\ d_{136} = t_{137} - t_{136} = 0.7425,$$

which can be summarized as follows:

Interval:	$[0, 0.2)$	$[0.2, 0.4)$	$[0.4, 0.6)$	$[0.6, 0.8)$	$[0.8, 1.0)$	$[1.0, 1.2)$	$[1.2, 1.4)$
Number in Interval:	63	31	14	9	9	4	6

A time difference in the interval $[0, 0.2)$ occurred 63 times, a time difference in the interval $[0.2, 0.4)$ occurred 31 times, and so forth. The average time difference was ≈ 0.354 years.

The reciprocal of the average, $2.821 \approx \frac{1}{0.354}$, is an estimate of the yearly rate earthquakes occurred during the interval from the first to the last observed earthquake.

Assuming the list giving the differences between successive earthquakes can be thought of as the values of a random sample from an exponential distribution, then it is reasonable to use 2.821 to estimate the parameter of the distribution. (The reciprocal of the sample mean is the maximum likelihood estimate of λ for an exponential distribution. A goodness-of-fit test can be used to check that the exponential distribution is a reasonable model for these data.)

Empirical histograms. Practitioners often use a graphical method, known as an *empirical histogram* (or a *histogram*), to summarize data:

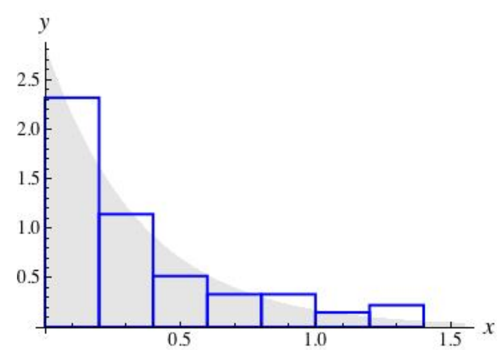
1. An interval containing the observed data is divided into k subintervals of equal length, and the number of observations in each subinterval is computed.
2. For each subinterval, a rectangle whose base is the subinterval itself, and whose *area* is the proportion of observations falling in the subinterval, is drawn.

In this way, the sum of the areas of the rectangles in a histogram is exactly 1.

To illustrate the technique, we use the time differences data with the 7 intervals

$$[0, 0.2),\ [0.2, 0.4),\ \ldots,\ [1.2, 1.4).$$

The plot on the right shows the histogram for these data superimposed on the density function for an exponential distribution with parameter 2.821.



Note that we will use graphical tools such as *comparison plots* (comparing empirical histograms to model PDFs, as above), *probability plots* and *box plots* (discussed in the second set of course notes) throughout the term. The graphical tools will allow us to understand probability distributions and numerical data better.

1.1.9 Sequences of IID Random Variables and the Central Limit Theorem

Let $X_1,\ X_2,\ X_3,\ \ldots$ be a sequence of mutually independent, identically distributed (IID) random variables, each with the same distribution as X.

Two related sequences are of interest:

1. *Sequence of Running Sums:* $S_m = \sum_{i=1}^{m} X_i$, for $m = 1, 2, 3, \ldots$.
2. *Sequence of Running Averages:* $\overline{X}_m = \frac{1}{m}\sum_{i=1}^{m} X_i$, for $m = 1, 2, 3, \ldots$.

The central limit theorem, whose proof is attributed to Laplace and deMoivre, gives us information about the probability distribution of S_m when m is large. Note that the distribution of X must be particularly well-behaved for the result of the theorem to hold.

> ***Theorem (Central Limit Theorem).*** Let $X_1,\ X_2,\ X_3,\ \ldots$ be a sequence of independent, identically distributed random variables, each with the same distribution as X. Assume that the moment generating function of X converges in an open interval containing 0, and let $\mu = E(X)$ and $\sigma = SD(X)$. Then, for every real number x,
>
> $$\lim_{m\to\infty} P\left(\frac{S_m - m\mu}{\sigma\sqrt{m}} \leq x\right) = \Phi(x),$$
>
> where $S_m = \sum_{i=1}^{m} X_i$, and $\Phi()$ is the CDF of the standard normal random variable.

Notes:

1. A useful way to think about the conclusion of the central limit theorem is that

   $$P(S_m \leq x) \approx \Phi\left(\frac{x - m\mu}{\sigma\sqrt{m}}\right), \quad \text{for each } x,$$

   when m is large enough. That is, the distribution of the sum is well approximated by the distribution of a normal random variable with mean $E(S_m) = m\mu$ and standard deviation $SD(S_m) = \sigma\sqrt{m}$ when m is large enough.

2. Since $\overline{X}_m = \frac{1}{m}S_m$, the central limit theorem also gives us information about the probability distribution of $\overline{X}_m$. Specifically,

   $$P(\overline{X}_m \leq x) \approx \Phi\left(\frac{x - \mu}{\sigma/\sqrt{m}}\right), \quad \text{for each } x,$$

   when m is large enough. That is, the distribution of the average $\overline{X}_m$ is well approximated by the distribution of a normal random variable with mean $E(\overline{X}_m) = \mu$ and standard deviation $SD(\overline{X}_m) = \frac{\sigma}{\sqrt{m}}$ when m is large enough.

3. If X has a normal distribution, then the probability statements in items 1 and 2 are exact, not approximate, for all m.

4. A table of values of $\Phi(z)$ is given on page A7 (Table 2) in the Rice textbook. An extended table appears in Section 1.3.5 (page 43) of this notebook.

For example, suppose that the occurrences of light-to-moderate earthquakes in the northeastern United States and eastern Canada follow a Poisson process with rate 2.74 events per year. If we observe the process for 50 years, let X_i be the number of earthquakes observed in the i^{th} year and

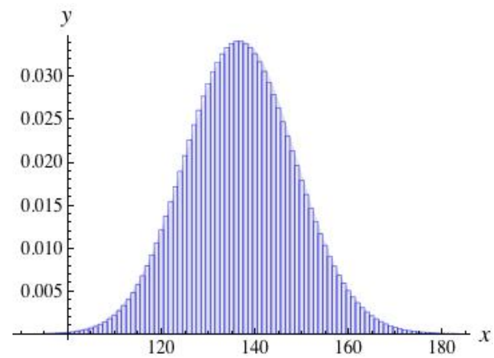
$$S = X_1 + X_2 + \cdots + X_{50}$$

be the total number of earthquakes observed, then S has a Poisson distribution with

$$E(S) = 50E(X) = 50(2.74) = 137 \text{ and } SD(S) = SD(X)\sqrt{50} = \sqrt{2.74}\sqrt{50} = \sqrt{137} \approx 11.7047.$$

The distribution of S is well approximated by the distribution of a normal random variable with mean 137 and standard deviation $\sqrt{137}$. (For Poisson distributions, the approximation is good when the overall mean is greater than 100.)

The plot shows the probability histogram of S superimposed on the density function of the approximating normal distribution. The plots are indistinguishable.



Comparison of distributions. It is interesting to compare distributions as m increases.

For example, suppose that GPA scores of mathematics majors in the United States follow a continuous uniform distribution on $[2.4, 4.0]$.

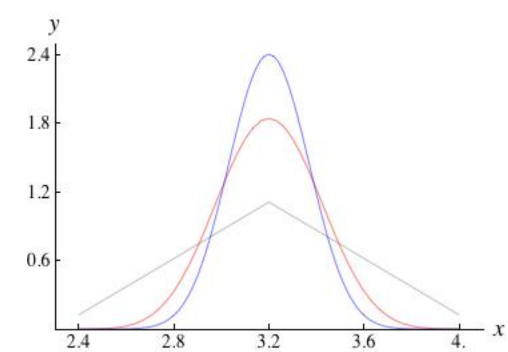
If X is a continuous random variable with this distribution, then

$$E(X) = \frac{2.4 + 4.0}{2} = 3.2 \text{ (the midpoint), and } SD(X) = \sqrt{(4.0 - 2.4)^2/12} \approx 0.462.$$

Let $\overline{X}_m$ be the mean of m IID random variables, each with the same distribution as X, for $m = 2, 6, 10$.

Each distribution is centered at 3.2. The standard deviations are as follows:

m	$SD(X)/\sqrt{m}$
2	≈ 0.327
6	≈ 0.189
10	≈ 0.146



As m increases, the standard deviation of $\overline{X}_m$ decreases, and the distribution becomes more concentrated around 3.2, as the plot above suggests. When $m = 2$, the density curve is the piecewise linear curve shown in the plot (the curve with the corner when $x = 3.2$). The density curves when $m = 6$ and $m = 10$ look like approximate normal density curves; the curve when $m = 10$ is narrower and taller than the curve when $m = 6$.

1.2 Review of Statistical Inference

This section reviews the fundamental concepts of estimation theory and hypothesis testing theory. These concepts will be applied in new ways in this course.

1.2.1 Random Sample, Statistic, Sampling Distribution

The concepts of random sample, statistic, and sampling distribution are essential first concepts:

1. *Random Sample:* A random sample of size n from the X distribution is a list,

   $$X_1,\, X_2,\, \ldots,\, X_n,$$

   of n mutually independent random variables, each with the same distribution as X.

2. *Statistic:* A *statistic* is a function of a random sample (or random samples).

3. *Sampling Distribution:* The probability distribution of a statistic is known as its *sampling distribution.* Sampling distributions are often difficult to find.

1.2.2 Sample Mean, Sample Variance, Sample Standard Deviation

Three useful statistics are the sample mean, the sample variance and the sample standard deviation. We know certain summaries of these random variables (when X has finite mean and variance), but do not always know their exact sampling distributions.

If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with finite mean μ and finite standard deviation σ, then the *sample mean*, $\overline{X}$, and the *sample variance*, S^2, are defined as follows:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

The *sample standard deviation*, S, is the positive square root of the sample variance.

The following theorem can be proven using properties of expectation:

> ***Theorem (Sample Summaries).*** If $\overline{X}$ is the sample mean and S^2 is the sample variance of a random sample of size n from a distribution with mean μ and standard deviation σ, then
>
> 1. $E(\overline{X}) = \mu$ and $Var(\overline{X}) = \sigma^2/n$.
> 2. $E(S^2) = \sigma^2$.

Note that, in general, $E(S) \neq \sigma$. (The value is close, but not exact.)

1.2.3 Sampling from Normal Distributions

When random samples are chosen from normal distributions, then we can say more about the sampling distributions of the sample mean and sample variance. Specifically,

> ***Theorem (Sampling Distributions).*** Let $\overline{X}$ be the sample mean and S^2 be the sample variance of a random sample of size n from a normal distribution with mean μ and standard deviation σ. Then
>
> 1. $\overline{X}$ is a normal random variable with mean μ and standard deviation $\sigma/\sqrt{n}$.
> 2. $V = \frac{(n-1)}{\sigma^2}S^2$ is a chi-square random variable with $(n-1)$ *df*.
> 3. $\overline{X}$ and S^2 are independent random variables.

Notes:

1. The theorem tells us that the sampling distribution of S^2 is a *scaled* chi-square distribution with $(n-1)$ degrees of freedom.
2. The chi-square distribution is discussed in Section 1.3.6 (page 44) of this notebook. The section also contains a table of selected quantiles, χ^2_p.
3. Selected quantiles of the standard normal distribution, z_p, appear as the "∞ row" of the Student t distribution table (see Section 1.3.7, page 47, of this notebook).
4. If X is a normal random variable with mean μ and standard deviation σ, then
   (a) A general formula for the p^{th} quantile of $\overline{X}$ distribution is $x_p = \mu + z_p \frac{\sigma}{\sqrt{n}}$, where z_p is the p^{th} quantile of the standard normal distribution.
   (b) A general formula for the p^{th} quantile of the S^2 distribution is $x_p = \frac{\sigma^2}{n-1}\chi^2_p$, where χ^2_p is the p^{th} quantile of the chi-square distribution with $(n-1)$ degrees of freedom.

Approximate standardization. In many statistical applications, we need to work with an approximate standardization of the sample mean. That is, we need to work with

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \quad \text{instead of} \quad Z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}},$$

where the approximation is produced by replacing σ^2 by S^2 in the exact standardization. The distribution of T is known exactly.

> ***Theorem (Sampling Distribution).*** Let $\overline{X}$ be the sample mean and S^2 be the sample variance of a random sample of size n from a normal distribution with mean μ and standard deviation σ. Then the approximate standardization T defined above has a Student t distribution with $(n-1)$ degrees of freedom.

Note that the Student t distribution is discussed in Section 1.3.7 (page 47) of this notebook.

Independent random samples. In many statistical applications, we need to work with independent random samples from normal distributions. The usual setup is as follows:

1. *X Sample:* Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a normal distribution with mean μ_x and standard deviation σ_x. Let $\overline{X}$ and S_x^2 be the sample mean and sample variance, respectively.

2. *Y Sample:* Let $Y_1, Y_2, \ldots, Y_m$ be a random sample of size m from a normal distribution with mean μ_y and standard deviation σ_y. Let $\overline{Y}$ and S_y^2 be the sample mean and sample variance, respectively.

3. *Independent Samples:* Assume that the samples were chosen independently. Thus, the combined sample,
$$X_1,\ X_2,\ \ldots,\ X_n,\ Y_1,\ Y_2,\ \ldots,\ Y_m$$
is a list of $n+m$ mutually independent random variables, where the first n are IID with the same distribution as X, and the last m are IID with the same distribution as Y.

Difference in means. The statistic $\overline{X} - \overline{Y}$ can be used to study the difference in means parameter $\delta = \mu_x - \mu_y$. In many statistical applications, we need to work with the approximate standardization of the difference in sample means. That is, we need to work with

$$T = \frac{(\overline{X} - \overline{Y}) - \delta}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \quad \text{instead of} \quad Z = \frac{(\overline{X} - \overline{Y}) - \delta}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}},$$

where the approximation is produced by replacing σ_x^2 by S_x^2 and σ_y^2 by S_y^2 in the exact standardization. Surprisingly, the exact sampling distribution of T is *not* known.

If $\sigma_x = \sigma_y = \sigma$, and we use $S_p^2 = ((n-1)S_x^2 + (m-1)S_y^2)/(n+m-2)$ to estimate σ^2, then an exact sampling distribution is known. Specifically, if we work with

$$T = \frac{(\overline{X} - \overline{Y}) - \delta}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \quad \text{as an approximation of} \quad Z = \frac{(\overline{X} - \overline{Y}) - \delta}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)}},$$

where S_p^2 replaces σ^2, then the sampling distribution of T is known exactly.

> ***Theorem (Sampling Distribution).*** Let $\overline{X}$ be the sample mean and S_x^2 be the sample variance of a random sample of size n from a normal distribution with mean μ_x and standard deviation σ. Let $\overline{Y}$ be the sample mean and S_y^2 be the sample variance of a random sample of size m from a normal distribution with mean μ_y and standard deviation σ. Then
>
> $$T = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}, \quad \text{where} \quad S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{(n+m-2)},$$
>
> has a Student t distribution with $(n+m-2)$ degrees of freedom.

Note that the Student t distribution is discussed in Section 1.3.7 (page 47) of this notebook.

Ratio of variances. The statistic S_x^2/S_y^2 can be used to study the ratio of variances parameter $r = \sigma_x^2/\sigma_y^2$. The following theorem tells us about the sampling distribution of the ratio of sample variances when samples are chosen independently from the X and Y distributions.

> ***Theorem (Sampling Distribution).*** Let X be a normal random variable with mean μ_x and standard deviation σ_x, let Y be a normal random variable with mean μ_y and standard deviation σ_y, and let S_x^2 and S_y^2 be the sample variances of independent random samples of sizes n and m, respectively, from the X and Y distributions. Then
>
> $$F = \frac{S_x^2/S_y^2}{\sigma_x^2/\sigma_y^2}$$
>
> has an f ratio distribution with $(n-1)$ and $(m-1)$ degrees of freedom, where the numerator is the ratio of sample variances and the denominator is the ratio of model variances.

Note that the f ratio distribution is discussed in Section 1.3.8 (page 50) of this notebook.

Summary. The following table summarizes the random variables and sampling distributions we use to study various parameters related to normal distributions.

Parameter of interest:	*Random Variable:*	*Sampling Distribution:*
μ, when σ is known	$Z = \frac{(\overline{X}-\mu)}{\sqrt{\sigma^2/n}}$	Standard Normal
μ, when σ is estimated	$T = \frac{(\overline{X}-\mu)}{\sqrt{S^2/n}}$	Student t, $(n-1)$ df
σ^2, when μ is known	$V = \frac{1}{\sigma^2}\sum_i (X_i-\mu)^2$	Chi-square, n df
σ^2, when μ is estimated	$V = \frac{1}{\sigma^2}\sum_i (X_i-\overline{X})^2$	Chi-square, $(n-1)$ df
$\mu_x-\mu_y$, when σ_x and σ_y are known	$Z = \frac{((\overline{X}-\overline{Y})-(\mu_x-\mu_y))}{\sqrt{\sigma_x^2/n+\sigma_y^2/m}}$	Standard Normal
$\mu_x-\mu_y$, when $\sigma_x=\sigma_y$ is estimated	$T = \frac{((\overline{X}-\overline{Y})-(\mu_x-\mu_y))}{\sqrt{S_p^2(1/n+1/m)}}$	Student t, $(n+m-2)$ df
$\mu_x-\mu_y$, when $\sigma_x\neq\sigma_y$ are estimated	$T = \frac{((\overline{X}-\overline{Y})-(\mu_x-\mu_y))}{\sqrt{S_x^2/n+S_y^2/m}}$	Approx Student t
σ_x^2/σ_y^2, when μ_x and μ_y are estimated	$F = \frac{(S_x^2/S_y^2)}{(\sigma_x^2/\sigma_y^2)}$	F ratio, $(n-1)$ and $(m-1)$ df

Since much is known about sampling from normal distributions, many practitioners use techniques for normal distributions without checking assumptions. In this course, we will learn appropriate alternatives when normal theory assumptions are not met.

Note that confidence interval procedures are given in Section 1.3.9 (page 51), and hypothesis test procedures are given in Section 1.3.10 (page 54) of this notebook.

1.2.4 Point and Interval Estimation

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ. The notation

$$\widehat{\theta} \;=\; \widehat{\theta}\,(X_1, X_2, \ldots, X_n)$$

(read "theta-hat") is used to denote an estimator of θ.

1. ***Standard error of $\widehat{\boldsymbol{\theta}}$.*** The standard deviation of the $\widehat{\theta}$ distribution is often called its *standard error*, and we often write $SE(\widehat{\theta}) = SD(\widehat{\theta})$.

   (The definition of the standard error is a handy way to distinguish between the standard deviation of the X distribution, $SD(X) = \sigma$, and the standard deviation of an estimator of a parameter of the X distribution. Some of you may be familiar with the term "standard error of the mean", which refers to the standard deviation of the $\overline{X}$ distribution, $SE(\overline{X}) = \sigma/\sqrt{n}$.)

2. ***Bias of $\widehat{\boldsymbol{\theta}}$.*** The *bias* of the estimator $\widehat{\theta}$ is the difference between its expected value and the true parameter θ:
$$BIAS(\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$
   If $E(\widehat{\theta}) = \theta$ (that is, if the bias is 0), then $\widehat{\theta}$ is said to be an *unbiased estimator* of θ; otherwise, $\widehat{\theta}$ is said to be a *biased estimator* of θ. If $\widehat{\theta}$ is biased, but satisfies
$$\lim_{n\to\infty} E(\widehat{\theta}) = \theta, \quad \text{where } n \text{ is the sample size,}$$
   then $\widehat{\theta}$ is said to be an *asymptotically unbiased estimator* of θ.

   (An accurate estimator will have little or no bias. It is not always possible to find an unbiased estimator of a parameter θ, but the method of moments and the method of maximum likelihood generally produce estimators that are asymptotically unbiased.)

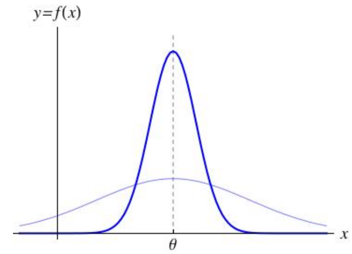
3. ***Mean squared error (MSE) of $\widehat{\boldsymbol{\theta}}$.*** The *mean squared error* (MSE) of an estimator is the expected value of the square of the difference between the estimator and the true parameter θ:
$$MSE(\widehat{\theta}) = E((\widehat{\theta} - \theta)^2).$$
   If $\widehat{\theta}$ is unbiased, then the mean squared error is the same as the variance; that is, $MSE(\widehat{\theta}) = Var(\widehat{\theta})$. Otherwise,
$$MSE(\widehat{\theta}) = Var(\widehat{\theta}) + \left(BIAS(\widehat{\theta})\right)^2.$$
   (A precise estimator will have a small mean squared error. The relationship between mean squared error, variance and bias can be proven using properties of expected values.)

4. ***Consistency of $\widehat{\boldsymbol{\theta}}$.*** The estimator $\widehat{\theta}$ is said to be a *consistent estimator* of θ if, for every positive number ϵ,
$$\lim_{n\to\infty} P(|\widehat{\theta} - \theta| \geq \epsilon) = 0, \quad \text{where } n \text{ is the sample size.}$$
   (A consistent estimator is unlikely to be far from the true parameter when n is large enough. The method of moments and the method of maximum likelihood generally produce estimators that are consistent.)

5. ***Comparing estimators: efficiency.*** Two different estimators are compared by comparing their variances or their mean squared errors:

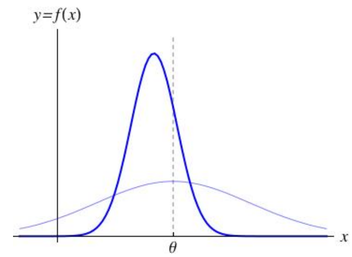
   1. *Unbiased Estimators:* Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be unbiased estimators of θ, each based on a random sample of size n from the X distribution.

      $\widehat{\theta}_1$ is said to be *more efficient* than $\widehat{\theta}_2$ if
      $$Var(\widehat{\theta}_1) < Var(\widehat{\theta}_2).$$



   2. *General Estimators:* More generally, let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two estimators of θ, each based on a random sample of size n from the X distribution.

      $\widehat{\theta}_1$ is said to be *more efficient* than $\widehat{\theta}_2$ if
      $$MSE(\widehat{\theta}_1) < MSE(\widehat{\theta}_2).$$



   (Given two unbiased estimators, we would prefer to use the one with the smaller variance (as illustrated in the *top plot*); otherwise, we would prefer to use the one whose average squared deviation from the true parameter is smaller (as illustrated in the *bottom plot*).)

A natural question to ask is whether a most efficient estimator exists. If we consider unbiased estimators only, we have the following definition.

> ***MVUE:*** The unbiased estimator $\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$ is said to be a *minimum variance unbiased estimator (MVUE)* of θ if
> $$Var(\widehat{\theta}) \le Var(\widehat{\theta}^*) \text{ for all unbiased estimators } \widehat{\theta}^* = \widehat{\theta}^*(X_1, X_2, \ldots, X_n).$$

In many circumstances, the method of maximum likelihood produces estimators that are "asymptotically efficient", that is, if n is large enough, then the bias of the estimator is virtually zero and the variance of the estimator is virtually equal to the minimum possible variance.

$100(1-\alpha)$% confidence intervals for θ: If

$$L = L(X_1, X_2, \ldots, X_n) \text{ and } U = U(X_1, X_2, \ldots, X_n)$$

are statistics (functions of the random sample), then the random interval $[L, U]$ is said to be a $100(1-\alpha)$% *confidence interval* for θ if

$$P(L \le \theta \le U) = 1 - \alpha \quad \text{and} \quad P(\theta < L) = P(\theta > U) = \frac{\alpha}{2}.$$

The quantity α is known as the *error probability.* After the values of the random variables in the random sample have been observed, the computed interval $[\ell, u]$ is said to contain θ "with confidence $1-\alpha$." (The procedure produces an interval containing the true parameter with probability $1-\alpha$. Note that a given computed interval $[\ell, u]$ does *not* contain the true parameter with probability $1-\alpha$.)

To develop confidence interval procedures, we need to know the sampling distributions of important random variables. For example,

1. *Variance Estimation:* Let S^2 be the sample variance of a random sample of size n from a normal distribution with mean μ and standard deviation σ, and assume that both parameters (μ and σ) are unknown. Since $V = \frac{(n-1)}{\sigma^2}S^2$ has a chi-square distribution with $(n-1)$ degrees of freedom, we can write the following probability statement:

$$1-\alpha = P\left(\chi^2_{\alpha/2} \le V \le \chi^2_{1-\alpha/2}\right) = P\left(\chi^2_{\alpha/2} \le \frac{(n-1)}{\sigma^2}S^2 \le \chi^2_{1-\alpha/2}\right),$$

where χ^2_p is the p^{th} quantile of the chi-square distribution with $(n-1)$ df. Since

$$\chi^2_{\alpha/2} \le \frac{(n-1)}{\sigma^2}S^2 \le \chi^2_{1-\alpha/2} \iff \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{\alpha/2}},$$

a $100(1-\alpha)\%$ confidence interval for the variance σ^2 is $[L,U] = \left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2}}\right]$.

To apply the procedure to find a 99% confidence interval: I used the computer to generate a pseudo-random sample from a normal distribution

68.6, 72.4, 79.2, 81.7, 83.7, 84.1, 86.0, 88.4, 89.0, 89.2, 89.4, 90.2, 102.5, 109.4.

For these data, $n = 14$, $\overline{x} = 86.7$ and $s^2 = 10.4826^2$. A 99% confidence interval for σ^2 based on these data is

$$\left[\frac{(13)(10.4826^2)}{29.82}, \frac{(13)(10.4826^2)}{3.57}\right] \approx [47.9041, 400.14].$$

With 99% confidence, we believe that the variance of the X distribution is between about 47.9 and 400.1 (and the standard deviation is between about 6.9 and 20.0).

2. *Difference in Means Estimation:* Let $\overline{X}$ be the sample mean of a random sample of size n from a normal distribution with mean μ_x and standard deviation σ, $\overline{Y}$ be the sample mean of a random sample of size m from a normal distribution with mean μ_y and standard deviation σ, and S_p^2 be the pooled estimate of the common variance σ^2. In addition, assume that the samples were chosen independently and that all three parameters (μ_x, μ_y and σ) are unknown.

Since $T = \frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$ has a Student t distribution with $(n+m-2)$ degrees of freedom, we can write the following probability statement:

$$1-\alpha = P\left(t_{\alpha/2} \le T \le t_{1-\alpha/2}\right) = P\left(t_{\alpha/2} \le \frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}} \le t_{1-\alpha/2}\right),$$

where t_p is the p^{th} quantile of the Student t distribution with $(n+m-2)$ df.

Since $t_{\alpha/2} \le \frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}} \le t_{1-\alpha/2}$

$$\iff (\overline{X}-\overline{Y}) - t_{1-\alpha/2}\sqrt{S_p^2\left(\tfrac{1}{n}+\tfrac{1}{m}\right)} \le \mu_x - \mu_y \le (\overline{X}-\overline{Y}) - t_{\alpha/2}\sqrt{S_p^2\left(\tfrac{1}{n}+\tfrac{1}{m}\right)},$$

or $L \leq \mu_x - \mu_y \leq U$, and since $t_{\alpha/2} = -t_{1-\alpha/2}$, a $100(1-\alpha)\%$ confidence interval for the difference in means parameter $\mu_x - \mu_y$ is usually written in the following form

$$(\overline{X} - \overline{Y}) \pm t_{1-\alpha/2}\sqrt{S_p^2\left(\tfrac{1}{n} + \tfrac{1}{m}\right)}.$$

To apply the procedure to find a 90% confidence interval: I used the computer to generate independent pseudo-random samples from normal distributions with equal variances. Summaries of the samples are as follows:

$$\begin{array}{ll} \textit{X Sample:} & n = 10,\ \overline{x} = 58.7919,\ s_x^2 = 4.57072^2 \\ \textit{Y Sample:} & m = 12,\ \overline{y} = 39.8881,\ s_y^2 = 4.71007^2 \end{array}$$

For these data, the estimated difference in means is 18.9037 and the pooled estimate of the common variance is 21.6028. A 90% confidence interval for $\mu_x - \mu_y$ is

$$18.9037 \pm (1.725)\sqrt{21.6028\left(\frac{1}{10} + \frac{1}{12}\right)} \Rightarrow 18.9037 \pm 3.43293 \Rightarrow [15.4708, 22.3367]$$

With 90% confidence, we believe that the mean of the X distribution is between about 15.5 and 22.3 units larger than the mean of the Y distribution.

Note: Confidence interval procedures for many situations are given in Section 1.3.9 (page 51) of this notebook. Selected quantiles for chi-square and Student t distributions are given in Sections 1.3.6 (page 44) and 1.3.7 (page 47) of this notebook, respectively. In this course, we will develop confidence interval procedures for many new situations, and we will develop good approximate methods for situations where sampling distributions are not known exactly or approximately.

1.2.5 Method of Moments; Method of Maximum Likelihood

The method of moments (MOM) is a general method for estimating parameters developed by Karl Pearson in the 1880's. Maximum likelihood (ML) estimation was developed by R.A. Fisher in the 1920's. Both methods are in current use, although ML estimation is preferred if both can be applied. (ML estimation is preferred because of the large sample theory that Fisher developed for ML estimators; no such theory exists for MOM estimators.)

Method of Moments (MOM) Estimation:

1. Kth Moments: The quantities $\mu_k = E(X^k)$ are known as the k^{th} *moments* of the X distribution, for $k = 1, 2, 3, \ldots$.
2. Kth Sample Moments: If $X_1, X_2, \ldots, X_n$ is a random sample of size n from the X distribution, then the quantities $\widehat{\mu_k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ are known as the k^{th} *sample moments*, for $k = 1, 2, 3, \ldots$. (Note that $\widehat{\mu_k}$ is an unbiased and consistent estimator of μ_k for each k.)

3. *Estimation Method:* Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ, and suppose that $\mu_k = E(X^k)$ is a function of θ for some k.

   Then, a *method of moments estimator* of θ is obtained using the following procedure:

   $$\text{Solve } \mu_k = \widehat{\mu_k} \text{ for the parameter } \theta.$$

In many (but not all) cases, MOM estimators are obtained by using $k = 1$.

For example, suppose that X is the continuous random variable with range $(0, 1)$ and PDF

$$f(x) = (\theta+1)x^\theta \quad \text{when } x \in (0,1), \text{ and } f(x) = 0 \text{ otherwise},$$

where θ is a positive constant. Since

$$\mu_1 = E(X) = \int_0^1 x\ f(x)\ dx = \left[\frac{\theta+1}{\theta+2}x^{\theta+2}\right]_0^1 = \frac{\theta+1}{\theta+2} - 0 = \frac{\theta+1}{\theta+2}$$

is a function of the parameter θ, the MOM estimator can be found by solving

$$\mu_1 = \widehat{\mu_1} \ \Rightarrow\ \frac{\theta+1}{\theta+2} = \overline{X} \ \Rightarrow\ \theta+1 = \overline{X}(\theta+2) \ \Rightarrow\ \theta = \frac{2\overline{X}-1}{1-\overline{X}}.$$

Thus, $\widehat{\theta} = \frac{2\overline{X}-1}{1-\overline{X}}$ is a MOM estimator of θ.

Maximum Likelihood (ML) Estimation:

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ.

1. *Likelihood Function:* The *likelihood function* is the joint PDF of the random sample thought of as a function of the parameter θ, with the random X_i's left unevaluated:

   $$Lik(\theta) = \begin{cases} \prod_{i=1}^n p(X_i) & \text{when } X \text{ is discrete} \\ \prod_{i=1}^n f(X_i) & \text{when } X \text{ is continuous} \end{cases}$$

2. *Log-Likelihood Function:* The *log-likelihood function* is the natural logarithm of the likelihood function:

   $$\ell(\theta) = \log(Lik(\theta))$$

   Note that it is common practice to use log() to denote the natural logarithm, and that this convention is used in the *Mathematica* system as well.

3. *Estimation Method:* The *maximum likelihood estimator* (or *ML estimator*) of θ is the value that maximizes either the likelihood function or the log-likelihood function.

   (In general, ML estimators are obtained using methods from calculus.)

In many (but not all) cases, the ML estimator corresponds to the unique turning point of the likelihood or log-likelihood function.

For example, let X be the continuous random variable used in the last example. Then the likelihood function is

$$Lik(\theta) = \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} (\theta+1) X_i^{\theta} = (\theta+1)^n \prod_{i=1}^{n} X_i^{\theta} = (\theta+1)^n \left(\prod_{i=1}^{n} X_i \right)^{\theta},$$

the log-likelihood function is $\ell(\theta) = n\log(\theta+1) + \theta \log\left(\prod_{i=1}^{n} X_i\right)$, the first derivative of the log-likelihood function is $\ell'(\theta) = \frac{n}{\theta+1} + \log\left(\prod_{i=1}^{n} X_i\right)$, and the second derivative is $\ell''(\theta) = -\frac{n}{(\theta+1)^2}$.

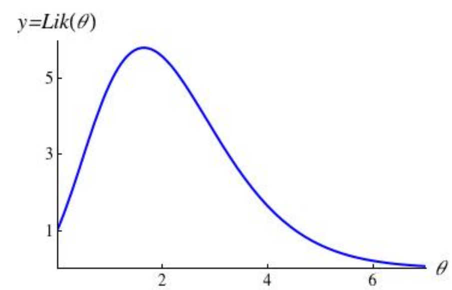
The first derivative equals zero when θ is $-\frac{n}{\log\left(\prod_{i=1}^{n} X_i\right)} - 1$, and the second derivative is negative at this point. Thus, $\widehat{\theta} = -\frac{n}{\log\left(\prod_{i=1}^{n} X_i\right)} - 1$ is the ML estimator of θ for this model.

MOM and ML estimators may be different. In many (but not all) cases, MOM and ML estimators are the same. The estimators for the model above, however, are different.

To illustrate the estimation methods: I used the computer to generate a pseudo-random sample of size 5 from the X distribution:

$$0.79, 0.68, 0.53, 0.76, 0.70.$$

For these data, the MOM estimate is 1.24675, and the ML estimate is 1.64918. The ML estimate corresponds to the first coordinate of the turning point of the observed likelihood function shown on the right.



Large sample theory for ML estimators. R.A. Fisher proved a generalization of the Central Limit Theorem for ML estimators.

Fisher's Theorem. Let $\widehat{\theta}_n$ be the ML estimator of θ based on a random sample of size n from the X distribution, and

$$Z_n = \frac{\widehat{\theta}_n - \theta}{\sqrt{1/(nI(\theta))}}, \quad \text{where } nI(\theta) \text{ is the information.}$$

Under smoothness conditions on the X distribution,

$$\lim_{n\to\infty} P(Z_n \le x) = \Phi(x) \quad \text{for every real number } x, \text{ where } \Phi(\cdot) \text{ is}$$

the cumulative distribution function of the standard normal random variable.

Notes:

1. *Information in a Sample of Size n:* The quantity $nI(\theta)$ is called the *information in a sample of size n*. It can be computed as

$$nI(\theta) = -E(\ell''(\theta)),$$

where $\ell''(\theta)$ is the second derivative of the log-likelihood function, and the expectation is computed using the joint distribution of the X_i's in the random sample for fixed θ.

The reciprocal of the information is the *Cramer-Rao lower bound* for variances of unbiased estimators when the X distribution satisfies certain "smoothness conditions".

2. *Smoothness Conditions:* If the following three conditions hold:

   (a) The PDF of X has continuous second partial derivatives (except, possibly, at a finite number of points),

   (b) The parameter θ is not at the boundary of possible parameter values, and

   (c) The range of X does not depend on θ,

then X satisfies the "smoothness conditions" of the theorem.

The theorem excludes, for example, the Bernoulli distribution with $p = 1$ (the 2^{nd} condition is violated) and the uniform distribution on $[0, b]$ (the 3^{rd} condition is violated).

3. *Large Sample Approximate Confidence Intervals:* Under the conditions of Fisher's theorem, an approximate $100(1-\alpha)\%$ confidence interval for θ has the following form:

$$\widehat{\theta} \ \pm \ z(\alpha/2)\sqrt{\frac{1}{nI(\widehat{\theta})}}$$

where $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

(Fisher's theorem is enormously useful. Under smoothness conditions and when n is large, we have a good approximation to the sampling distribution of the ML estimator. And, since the distribution is approximately normal, we have an easy-to-remember procedure for finding confidence intervals.)

To find the large sample procedure for Poisson models: Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a Poisson distribution with mean λ, and let $Y = \sum_{i=1}^n X_i$ be the sum of the X_i's and $Q = \prod_{i=1}^n X_i$ be the product of the X_i's.

1. The likelihood function is

$$Lik(\lambda) = \prod_{i=1}^n p(X_i) = \prod_{i=1}^n e^{-\lambda}\frac{\lambda^{X_i}}{X_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n X_i}}{\prod_{i=1}^n X_i!} = \frac{e^{-n\lambda}\lambda^Y}{Q},$$

the log-likelihood function is $\ell(\lambda) = -n\lambda + Y\log(\lambda) - \log(Q)$, and the first two derivatives of the log-likelihood function are

$$\ell'(\lambda) = -n + \frac{Y}{\lambda} \ \text{ and } \ \ell''(\lambda) = -\frac{Y}{\lambda^2}.$$

(When computing the derivatives, Y is fixed and λ is variable.)

2. The first derivative equals zero when λ is $\frac{Y}{n}$. (I will let you check this for yourself.) Since the second derivative evaluated at this point is negative, we know that the critical point corresponds to a maximum. Thus, the ML estimator is $\widehat{\lambda} = \frac{Y}{n}$.

3. Since $E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = n\lambda$, the information is

$$nI(\lambda) = -E(\ell''(\lambda)) = -E\left(-\frac{Y}{\lambda^2}\right) = \frac{1}{\lambda^2}E(Y) = \frac{1}{\lambda^2}(n\lambda) = \frac{n}{\lambda},$$

and the reciprocal of the information (the asymptotic variance) is $\frac{1}{nI(\lambda)} = \frac{\lambda}{n}$.

(When computing the expectation, λ is fixed and Y is variable.)

4. Thus, the form of an approximate $100(1-\alpha)\%$ confidence interval for λ is

$$\widehat{\lambda} \pm z(\alpha/2)\sqrt{\frac{\widehat{\lambda}}{n}}, \quad \text{where} \quad \widehat{\lambda} = \frac{Y}{n}$$

and $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

To apply the procedure to find an approximate 95% confidence interval: I used the computer to generate a pseudo-random sample from a Poisson distribution:

Value of x:	0	1	2	3	4	5	6	7
Frequency:	10	7	21	8	5	6	2	1

(0 occurred 10 times, 1 occurred 7 times, and so forth.)

For these data, $n = 60$, $y = 142$ and the ML estimate is $\widehat{\lambda} = 2.36\overline{6}$. An approximate 95% confidence interval for λ based on these data is

$$2.36\overline{6} \pm (1.960)\sqrt{\frac{2.36\overline{6}}{60}} \quad \Rightarrow \quad 2.36\overline{6} \pm 0.389268 \quad \Rightarrow \quad [1.9774, 2.75593].$$

With approximate 95% confidence, we believe that the mean of the Poisson distribution is between about 1.98 and 2.76.

Note that several applications of Fisher's theorem are given in Section 1.3.9 (page 51) of this notebook. (You may find it useful to check that each formula is correct.)

I used the "∞ row" of the reference table for the Student t distribution (see Section 1.3.7, page 47 of this notebook) to find the 97.5% point of the standard normal distribution.

1.2.6 Hypothesis Testing

1. ***Neyman-Pearson framework:*** An *hypothesis* is an assertion about the distribution of a random variable or a random k-tuple (a list of k random variables). In the *Neyman-Pearson framework* of hypothesis testing, there are two competing assertions:

   (a) The null hypothesis, H_O, and

   (b) The alternative hypothesis, H_A.

The null hypothesis is accepted as true unless sufficient evidence is provided to the contrary. If sufficient evidence is provided to the contrary, then the null hypothesis is rejected in favor of the alternative hypothesis.

If there is insufficient evidence to reject a null hypothesis, researchers often say that they have "failed to reject the null hypothesis" rather than they "accept the null hypothesis".

2. ***Parameter Space/Null Space:*** Null and alternative hypotheses are often written as assertions about parameters. Let θ be the parameter of interest.

   We often use the following notation:

$$H_O\colon\ \theta \in \omega_o \qquad \text{versus} \qquad H_A\colon\ \theta \in \Omega \setminus \omega_o,$$

   where the set Ω (the *parameter space*) is the set of all parameters under consideration, and the set ω_o (the *null space*) is the subset of parameters satisfying the null hypothesis.

   For example, if X has a Poisson distribution with parameter λ and we are interested in testing

$$H_O\colon\ \lambda \le 4.5 \qquad \text{versus} \qquad H_A\colon\ \lambda > 4.5,$$

   then the parameter space is $\Omega = (0, \infty)$ and the null space is $\omega_o = (0, 4.5]$.

3. ***Test:*** A *test* is a decision rule allowing the user to choose between competing assertions.

4. ***Components of a Test:*** Let X_1, X_2, $\ldots$, X_n be a random sample from a distribution with parameter θ. To set up a test:

   (a) A *test statistic*, $T = T(X_1, \ldots, X_n)$, is chosen and
   (b) The range of T is subdivided into the *rejection region* (RR) and the complementary *acceptance region* (AR).
   (c) If the observed value of T is in the acceptance region, then the null hypothesis is accepted. Otherwise, the null hypothesis is rejected in favor of the alternative.

   The test statistic, and the acceptance and rejection regions, are chosen so that the probability that T is in the rejection region is small (i.e., near 0) when the null hypothesis is true. Hopefully, the probability that T is in the rejection region is large (i.e., near 1) when the alternative hypothesis is true, but this is *not* guaranteed.

   For example, let X_1, X_2, X_3, X_4, X_5 be a random sample of size 5 from a Poisson distribution with parameter λ, and let $Y = \sum_{i=1}^{5} X_i$ be the sample sum. Consider the test with decision rule

$$\text{Reject } \lambda \le 4.5 \text{ in favor of } \lambda > 4.5 \text{ when } Y \ge 29.$$

   The test statistic, Y, has a Poisson distribution with parameter 5λ. The acceptance and rejection regions are

$$AR = \{0, 1, \ldots, 28\} \ \text{ and } \ RR = \{29, 30, 31, \ldots\}.$$

   (The sum of independent Poisson random variables is a Poisson random variable. I chose the rejection region so that the probability that Y was in the rejection region was at most 0.106 for models satisfying the null hypothesis, and so that the probability increases with increasing λ.)

5. ***Errors of Type I/II:*** When carrying out a test, two types of errors can occur:

   (a) An *error of type I* occurs when a true null hypothesis is rejected.

   (b) An *error of type II* occurs when a false null hypothesis is accepted.

6. ***Size/Power:*** Consider the test with decision rule:

$$\text{Reject } \theta \in \omega_o \text{ in favor of } \theta \in \Omega \setminus \omega_o \text{ when } T \in RR$$

where $T = T(X_1, X_2, \ldots, X_n)$ is a test statistic based on a random sample of size n.

Further, let $P_\theta(T \in RR)$ be the probability that T lies in its rejection region if the true parameter is θ. Then

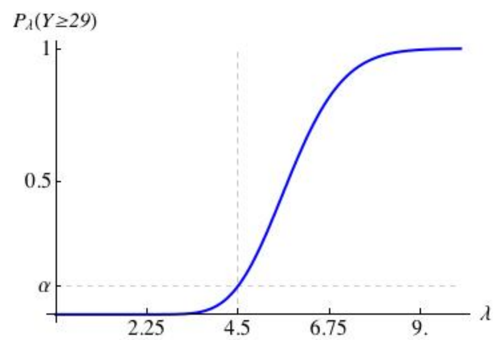
   (a) The *size* (or *significance level*) of the test is the maximum type I error for all models satisfying the null hypothesis, or the least upper bound (the supremum) of type I errors if a maximum does not exist:

$$\alpha = \sup_{\theta \in \omega_o} P_\theta(T \in RR).$$

   (b) The *power* of the test at the parameter θ is $P_\theta(T \in RR)$.

Continuing with the example above, I used the computer to graph the power function,

where $P_\lambda(Y \geq 29) = 1 - \sum_{y=0}^{28} e^{-5\lambda} \frac{(5\lambda)^y}{y!}$.



The size of the test is

$$\alpha = \max_{\lambda \leq 4.5} P_\lambda(Y \geq 29) \approx 0.106.$$

For models satisfying the null hypothesis, the values of the power function are at most α. As λ increases, the power increases.

7. ***Observed Significance Level/p Value:*** The *observed significance level*, or *p value*, is the minimum significance level for which the null hypothesis would be rejected.

Continuing with the example above, suppose that 9, 4, 4, 3, 7 were observed. Then the observed sum is 27, and the observed significance level is

$$P_{4.5}(Y \geq 27) = 1 - \sum_{y=0}^{26} e^{-22.5} \frac{22.5^y}{y!} \approx 0.196.$$

(For this problem, the observed significance level is calculated by placing the observed value of the test statistic at the cutoff for the test, and using $\lambda = 4.5$. When $\lambda = 4.5$, Y is a Poisson random variable with parameter $5\lambda = 22.5$. We would reject the null hypothesis at significance levels of 0.196 or more.)

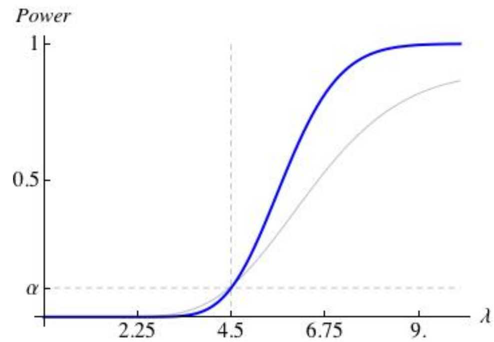
8. ***Comparing tests: power for fixed size α.*** Let T and W be two different test statistics based on a random sample of size n, and suppose that each corresponds to a test of size α. The test based on T is *uniformly more powerful* than the test based on W iff

$$P_\theta(T \in RR_T) \geq P_\theta(W \in RR_W) \text{ for all } \theta \in \Omega \setminus \omega_o,$$

with strict inequality (that is, with ">" replacing "≥") for at least one θ, where RR_T and RR_W are the rejection regions for T and W, respectively.

Continuing with the example above: The test based on Y is uniformly most powerful for samples of size 5 and $\alpha = 0.106$.

Any competitor would have a power curve looking something like the gray curve in the plot to the right: (1) when $\lambda \leq 4.5$, the values lie below α; and (2) when $\lambda > 4.5$, the values for the test based on Y are greater than the values for the competitor test.



A natural question to ask is whether a uniformly most powerful test always exists. Formally,

UMPT: Let $T = T(X_1, X_2, \ldots, X_n)$ be the test statistic for a size α test of

$$H_O : \theta \in \omega_o \text{ versus } H_A : \theta \in \Omega \setminus \omega_o.$$

Then the test based on T is said to be a *uniformly most powerful* test of size α if it is uniformly more powerful than any competitor test of the same size.

In the 1930's, Neyman and Pearson designed the framework for hypothesis testing we use today, and studied the problem of finding uniformly most powerful tests. Under smoothness conditions, their likelihood ratio method for defining tests produces UMP tests when they exist, and produces tests with good statistical power when UMP tests do not exist.

Commonly used tests. Test statistics and rejection regions for commonly used tests are given in Section 1.3.10 (page 54) of this notebook. In each case, the form of the rejection region for a pre-specified significance level is given.

For example, suppose that we are interested in testing the null hypothesis $\mu = 50$ versus the alternative hypothesis $\mu \neq 50$ at the 5% significance level using a random sample of size 15 from a normal distribution, and assume that both parameters are unknown. Then the test statistic is $T = \frac{\overline{X}-50}{\sqrt{S^2/15}}$, and the rejection region is

$$|T| \geq t_{14}(.025) = 2.145.$$ (2.145 is the 97.5% point of the Student t distribution with 14 df.)

1.2.7 Likelihood Ratio Tests

Let X be a random variable, let θ be a parameter of the X distribution, and let $Lik(\theta)$ be the likelihood function for a random sample of size n from the X distribution.

Consider testing

$$H_O\colon\ \theta \in \omega_o \qquad \text{versus} \qquad H_A\colon\ \theta \in \Omega \setminus \omega_o,$$

where Ω is the set of all parameters under consideration, and the $\omega_o \subset \Omega$ is the subset of parameters satisfying the null hypothesis.

1. *Likelihood Ratio Statistic:* The *likelihood ratio statistic*, Λ, is the ratio of the maximum value of the likelihood function for models satisfying the null hypothesis to the maximum value of the likelihood function for all models under consideration:

$$\Lambda = \frac{\max_{\theta \in \omega_o} Lik(\theta)}{\max_{\theta \in \Omega} Lik(\theta)}$$

   Note that

   (a) $\Lambda \leq 1$ since the maximum over a subset of Ω must be less than or equal to the maximum over the full parameter space.
   (b) The denominator is maximized at the maximum likelihood (ML) estimator, $\widehat{\theta}$.

2. *Likelihood Ratio Test:* A *likelihood ratio test* based on this statistic is a test whose decision rule has the following form

$$\text{Reject } \theta \in \omega_o \text{ in favor of } \theta \in \Omega \setminus \omega_o \text{ when } \Lambda \leq c.$$

   Note that

   (a) If $\theta \in \omega_o$, then the numerator and denominator of Λ are likely to be approximately equal, and the value of Λ is likely to be close to 1.
   (b) If $\theta \in \Omega \setminus \omega_o$, then the numerator of Λ is likely to be much smaller than the denominator, and the value of Λ is likely to be close to 0.

Implementing likelihood ratio tests. A likelihood ratio test is not implemented as shown above. Instead, an equivalent test (with a simpler statistic and rejection region) is used. To demonstrate an equivalence with a test based on T, for example, we would need to show that

$$\Lambda \leq c \iff T \in RR_T.$$

(Equivalent events will have equal probabilities for all parameters θ.)

Derivations are often tedious. Suffice it to say that most of the commonly used tests are equivalent forms of likelihood ratio tests, and some are uniformly most powerful. (Distinctions are noted in Section 1.3.10, page 54 of this notebook).

Large sample theory. In the 1930's, Wilks proved a useful large sample approximation to the sampling distribution of a particular equivalent form: $T = -2\log(\Lambda)$.

> ***Wilks' Theorem.*** Let Λ be the likelihood ratio statistic for a test of
>
> $$H_O\colon\ \theta \in \omega_o \qquad \text{versus} \qquad H_A\colon\ \theta \in \Omega \setminus \omega_o$$
>
> based on a random sample of size n from the X distribution. Under smoothness conditions on the X distribution, and when n is large, the test statistic
>
> $-2\log(\Lambda)$ has an approximate chi-square distribution with $r - r_o$ *df*,
>
> where r is the number of free parameters in Ω, r_o is the number of free parameters in ω_o, and log() is the natural logarithm function.

The statistic $-2\log(\Lambda)$ has a nice form when working with discrete data, and we will use $-2\log(\Lambda)$ at the end of the term when studying contingency table problems.

1.3 Summary of Probability Distributions and Reference Tables

This section contains an overview of probability distributions, and reference tables of probabilities, quantiles and statistical procedures. The reference tables are useful for solving problems by hand without the use of a computer.

1.3.1 Discrete Probability Distributions

1. Discrete uniform distribution. Let n be a positive integer. The random variable X is said to be a *discrete uniform random variable*, or to have a *discrete uniform distribution*, with parameter n when its PDF is as follows:

$$p(x) \;=\; P(X = x) \;=\; \frac{1}{n} \quad \text{when } x \in \{1, 2, \ldots, n\}, \text{ and 0 otherwise.}$$

Discrete uniform distributions have n equally likely outcomes $(1, 2, \ldots, n)$.

2. Hypergeometric distribution. Let n, M, and N be integers with $0 < M < N$ and $0 < n < N$. The random variable X is said to be a *hypergeometric random variable*, or to have a *hypergeometric distribution*, with parameters n, M, and N, when its PDF is as follows

$$p(x) \;=\; P(X = x) \;=\; \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}\ ,$$

for integers, x, between $\max(0, n + M - N)$ and $\min(n, M)$, and equals 0 otherwise.

Hypergeometric distributions are used to model *urn experiments*, where N is the number of objects in the urn, M is the number of "special objects," n is the size of the subset chosen from the urn and X is the number of special objects in the chosen subset. If each choice of subset is equally likely, then X has a hypergeometric distribution.

In addition, hypergeometric distributions are used in *survey analysis*, where a simple random sample of n individuals are chosen from a population of total size N which contains a subpopulation of size M of particular interest to researchers.

3. Bernoulli distribution. A *Bernoulli experiment* is a random experiment with two outcomes. The outcome of chief interest is called "success" and the other outcome "failure." Let p equal the probability of success.

Let $X = 1$ if a "success" occurs, and let $X = 0$ otherwise. Then X is said to be a *Bernoulli random variable*, or to have a *Bernoulli distribution*, with parameter p. The PDF of X is:

$$p(1) = p, \quad p(0) = 1 - p, \quad \text{and} \quad p(x) = 0 \quad \text{otherwise.}$$

4. Binomial distribution. Let X be the number of successes in n independent trials of a Bernoulli experiment with success probability p. Then X is said to be a *binomial random variable*, or to have a *binomial distribution*, with parameters n and p. The PDF of X is:

$$p(x) \;=\; \binom{n}{x}\, p^x\, (1-p)^{n-x} \quad \text{when } x \in \{0, 1, 2, \ldots, n\}, \text{ and 0 otherwise.}$$

X can be thought of as the sum of n independent Bernoulli random variables. Thus, by the central limit theorem, the distribution of X is approximately normal when n is large.

Binomial distributions have many applications. One common application is to *survey analysis*, where a simple random sample of n individuals is drawn from a population of total size N which contains a subpopulation of size M of particular interest to a researcher. If the population size N is very large, and the proportion $p = M/N$ is not too extreme, then binomial probabilities can be used to approximate hypergeometric probabilities.

5. Geometric distribution on $\{0, 1, 2, \ldots\}$. Let X be the number of failures before the first success in a sequence of independent Bernoulli experiments with success probability p. Then X is said to be a *geometric random variable*, or to have a *geometric distribution*, with parameter p. The PDF of X is as follows:

$$p(x) \;=\; (1-p)^x p \quad \text{when } x \in \{0, 1, 2, \ldots\}, \text{ and 0 otherwise.}$$

Note that there is an alternative definition of the geometric random variable. Namely, that "X is the trial number of the first success in a sequence of independent Bernoulli experiments with success probability p." The definition above is the one that is more commonly used in applications, and it is the one that is part of the *Mathematica* system.

6. Negative binomial distribution on $\{0, 1, 2, \ldots\}$. Let X be the number of failures before the r^{th} success in a sequence of independent Bernoulli experiments with success probability p. Then X is said to be a *negative binomial random variable*, or to have a *negative binomial distribution*, with parameters r and p. The PDF of X is as follows:

$$p(x) \;=\; \binom{r-1+x}{x}(1-p)^x p^r \quad \text{when } x \in \{0, 1, 2, \ldots\}, \text{ and 0 otherwise.}$$

X can be thought of as the sum of r independent geometric random variables. Thus, by the central limit theorem, the distribution of X is approximately normal when r is large.

Note that there is an alternative definition of the negative binomial random variable. Namely, that "X is the trial number of the r^{th} success in a sequence of independent Bernoulli experiments with success probability p." The definition above is the one that is more commonly used in applications, and it is the one that is part of the *Mathematica* system.

7. Poisson distribution. Let λ be a positive real number. The random variable X is said to be a *Poisson random variable*, or to have a *Poisson distribution*, with parameter λ when its PDF is as follows:

$$p(x) \;=\; e^{-\lambda}\,\frac{\lambda^x}{x!} \quad \text{when } x \in \{0, 1, 2, \ldots\}, \text{ and 0 otherwise.}$$

If events occurring over time follow an approximate Poisson process, with an average of λ events per unit time, then X is the number of events observed in one unit of time.

The distribution of X is approximately normal when λ is large.

1.3.2 Continuous Probability Distributions

1. Continuous uniform distribution. Let a and b be real numbers with $a < b$. The random variable X is said to be a *continuous uniform random variable*, or to have a *continuous uniform distribution*, on the interval (a, b) when its PDF is as follows:

$$f(x) \;=\; \frac{1}{b-a} \quad \text{when } x \in (a, b), \text{ and 0 otherwise.}$$

The constant density for the continuous uniform random variable takes the place of the equally likely outcomes for the discrete uniform random variable.

2. Exponential distribution. Let λ be a positive real number. The random variable X is said to be an *exponential random variable* or to have an *exponential distribution* with parameter λ when its PDF is as follows:

$$f(x) \;=\; \lambda e^{-\lambda x} \quad \text{when } x \in (0, \infty), \text{ and 0 otherwise.}$$

An important application of the exponential distribution is to Poisson processes. Specifically, the time to the first event, or the time between successive events, of a Poisson process with rate λ has an exponential distribution with parameter λ.

3. Gamma distribution. Let α and β be positive real numbers. The continuous random variable X is said to be a *gamma random variable*, or to have a *gamma distribution*, with parameters α and β when its PDF is as follows:

$$f(x) \;=\; \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\, x^{\alpha-1}e^{-x/\beta} \quad \text{when } x \in (0, \infty), \text{ and 0 otherwise.}$$

Notes:

1. *Gamma Function:* The normalizing constant in the definition of the gamma distribution is known as the *Euler gamma function*; its definition is as follows:

$$\Gamma(\alpha) \; = \; \int_0^\infty x^{\alpha-1}e^{-x}dx \quad \text{when } \alpha > 0.$$

   Properties of the Euler gamma function include:

   (a) $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ for all positive real numbers α.

   (b) $\Gamma(n) = (n-1)!$ for all positive integers n.

   (Thus, the gamma function generalizes the factorial function to all positive numbers.)

2. *Relationship to Poisson Processes:* An important application of the gamma distribution is to Poisson processes. Specifically, the time to the r^{th} event of a Poisson process with rate λ has a gamma distribution with parameters $\alpha = r$ and $\beta = 1/\lambda$.

3. *Parameterizations:* α is a shape parameter, and β is a scale parameter of the gamma distribution. When α is large, the distribution of X is approximately normal.

   An alternative parameterization uses parameters α and $\lambda = 1/\beta$. The choice of parameters given here is the one used in the *Mathematica* system.

4. Cauchy distribution. Let a be a real number and b be a positive real number. The continuous random variable X is said to be a *Cauchy random variable*, or to have a *Cauchy distribution*, with center a and spread b when its PDF and CDF are as follows:

1. *Cauchy PDF:* $f(x) = \dfrac{b}{\pi(b^2+(x-a)^2)}$ for all real numbers x.
2. *Cauchy CDF:* $F(x) = \dfrac{1}{2} + \dfrac{1}{\pi}\tan^{-1}\left(\dfrac{x-a}{b}\right)$ for all real numbers x.

The Cauchy distribution is symmetric around its center a, but the expectation and variance of the Cauchy random variable are indeterminate.

Use the CDF of the Cauchy distribution for by-hand computations of probabilities.

5. Normal distribution. Let μ be a real number and σ be a positive real number. The continuous random variable X is said to be a *normal random variable*, or to have a *normal distribution*, with mean μ and standard deviation σ when its PDF and CDF are as follows:

1. *Normal PDF:* $f(x) = \dfrac{1}{\sqrt{2\pi}\,\sigma}e^{-(x-\mu)^2/(2\sigma^2)}$ for all real numbers x.
2. *Normal CDF:* $F(x) = \Phi\left(\dfrac{x-\mu}{\sigma}\right)$ for all real numbers x,

where $\Phi()$ is the CDF of the standard normal random variable.

Use the table of standard normal probabilities (Section 1.3.5, page 43) for by-hand computations of probabilities.

1.3.3 Model Summaries of Discrete and Continuous Models

Distribution, with PDF for x in the range	Model Summaries
Discrete Uniform on $\{1, 2, \ldots, n\}$ $p(x) = 1/n$, $x \in \{1, 2, \ldots, n\}$	$E(X) = \frac{n+1}{2}$ $Var(X) = \frac{n^2-1}{12}$
Hypergeometric with parameters n, M, N $p(x) = \binom{M}{x}\binom{N-M}{n-x}/\binom{N}{n}$, $x \in \{\max(0, n+M-N), \ldots, \min(n, M)\}$	$E(X) = n\frac{M}{N}$ $Var(X) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\left(\frac{N-n}{N-1}\right)$
Bernoulli with success probability p $p(1) = p$, $p(0) = 1 - p$	$E(X) = p$ $Var(X) = p(1-p)$
Binomial with parameters n, p $p(x) = \binom{n}{x}p^x(1-p)^{n-x}$, $x \in \{0, 1, \ldots, n\}$	$E(X) = np$ $Var(X) = np(1-p)$
Geometric on $\{0, 1, 2, \ldots\}$, with parameter p $p(x) = (1-p)^x p$, $x \in \{0, 1, 2, \ldots\}$	$E(X) = \frac{1-p}{p}$ $Var(X) = \frac{1-p}{p^2}$
Negative Binomial on $\{0, 1, 2, \ldots\}$, with parameters r, p $p(x) = \binom{r-1+x}{x}(1-p)^x p^r$, $x \in \{0, 1, 2, \ldots\}$	$E(X) = \frac{r(1-p)}{p}$ $Var(X) = \frac{r(1-p)}{p^2}$
Poisson with parameter λ $p(x) = e^{-\lambda}\frac{\lambda^x}{x!}$, $x \in \{0, 1, 2, \ldots\}$	$E(X) = \lambda$ $Var(X) = \lambda$
Continuous Uniform on the interval (a, b) $f(x) = \frac{1}{b-a}$, $x \in (a, b)$	$E(X) = \frac{a+b}{2}$, $Var(X) = \frac{(b-a)^2}{12}$ $Med(X) = \frac{a+b}{2}$, $IQR(X) = \frac{b-a}{2}$
Exponential with parameter λ $f(x) = \lambda e^{-\lambda x}$, $x \in (0, \infty)$	$E(X) = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2}$ $Med(X) = \frac{\ln(2)}{\lambda}$, $IQR(X) = \frac{\ln(3)}{\lambda}$
Gamma with shape α and scale β $f(x) = \frac{1}{\Gamma(\alpha)\ \beta^\alpha}\ x^{\alpha-1}e^{-x/\beta}$, $x \in (0, \infty)$	$E(X) = \alpha\beta$ $Var(X) = \alpha\beta^2$
Cauchy with center a and spread b $f(x) = \frac{b}{\pi(b^2+(x-a)^2)}$, $x \in (-\infty, \infty)$	$E(X)$ and $Var(X)$ are indeterminate $Med(X) = a$, $IQR(X) = 2b$
Normal with mean μ and standard deviation σ $f(x) = \frac{1}{\sqrt{2\pi}\ \sigma}e^{-(x-\mu)^2/(2\sigma^2)}$, $x \in (-\infty, \infty)$	$E(X) = \mu$, $Var(X) = \sigma^2$ $Med(X) = \mu$, $IQR(X) \approx 1.35\sigma$

1.3.4 Multivariate Discrete Probability Distributions

1. Multivariate hypergeometric distribution. Let n and M_i, for $i = 1, \ldots, k$, be positive integers with $n < M_1 + M_2 + \cdots + M_k$. The random k-tuple $(X_1, X_2, \ldots, X_k)$ is said to have a *multivariate hypergeometric distribution* with parameters n and $(M_1, M_2, \ldots, M_k)$ when its joint PDF has the following form:

$$p(x_1, x_2, \ldots, x_k) = \frac{\binom{M_1}{x_1}\binom{M_2}{x_2}\cdots\binom{M_k}{x_k}}{\binom{N}{n}},$$

where $N = M_1 + M_2 + \cdots + M_k$, each x_i is an integer satisfying $0 \le x_i \le \min(n, M_i)$, and the sum of the x_i's is exactly n; and the joint PDF $p(x_1, x_2, \ldots, x_k) = 0$ otherwise.

Multivariate hypergeometric distributions are used to model *urn experiments*, where N is the number of objects in the urn, M_i is the number of objects of the i^{th} type in the urn, for $i = 1, 2, \ldots, k$, n is the size of the subset chosen from the urn, and X_i is the number of objects of the i^{th} type in the chosen subset, for $i = 1, 2, \ldots, k$.

In addition, multivariate hypergeometric distributions are used in *survey analysis*, where a simple random sample of n individuals are chosen from a population of total size N which contains subpopulations of sizes M_i, for $i = 1, 2, \ldots, k$, of particular interest to researchers.

2. Multinomial distribution. Let n be a positive integer, and p_i, for $i = 1, 2, \ldots, k$, be positive proportions whose sum is exactly 1. The random k-tuple $(X_1, X_2, \ldots, X_k)$ is said to have a *multinomial distribution* with parameters n and $(p_1, p_2, \ldots, p_k)$ when its joint PDF has the following form:

$$p(x_1, x_2, \ldots, x_k) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

when each x_i is an integer satisfying $0 \le x_i \le n$, and the sum of the x_i's is exactly n; and the joint PDF $p(x_1, x_2, \ldots, x_k) = 0$ otherwise.

Multinomial distributions are used to model *independent* trials of an experiment with exactly k outcomes. Specifically, suppose that an experiment has k outcomes with probabilities

$$p_1,\ p_2,\ \ldots,\ p_k,$$

where the sum of the probabilities is 1. Let X_i be the number of occurrences of the i^{th} outcome in n independent trials of the experiment, for $i = 1, 2, \ldots, k$. Then $(X_1, X_2, \ldots, X_k)$ has a multinomial distribution with parameters n and $(p_1, p_2, \ldots, p_k)$.

Multinomial distributions have many applications. One common application is to *survey analysis*, where a simple random sample of n individuals is drawn from a population of total size N which contains subpopulations of sizes M_i, for $i = 1, 2, \ldots, k$, of particular interest to a researcher. If the population size N is very large, and each $p_i = M_i/N$ is not too extreme, then multinomial probabilities can be used to approximate multivariate hypergeometric probabilities. (Thus, sampling with replacement can be used an approximation to sampling without replacement in certain large population situations.)

1.3.5 Standard Normal Distribution: Table of Quantiles

Let Z be the standard normal random variable ($\mu = 0$, $\sigma = 1$) and let $\Phi(z) = P(Z \leq z)$ be the cumulative distribution function of Z.

- The following tables gives $\Phi(z)$ for $z \geq 0$ (where z = Row Value + Column Value).
- If $z < 0$, then $\Phi(z) = 1 - \Phi(-z)$.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

1.3.6 Chi-Square Distribution; Table of Quantiles

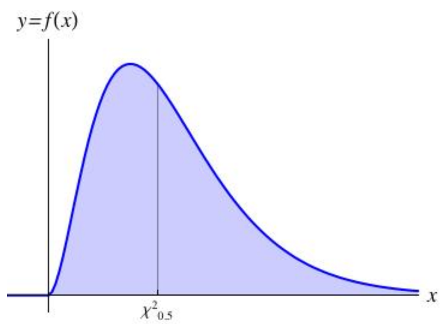
Let Z_1, Z_2, …, Z_m be independent standard normal random variables. Then

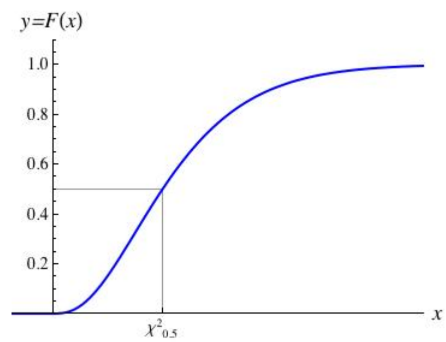
$$V \;=\; Z_1^2 + Z_2^2 + \cdots + Z_m^2$$

is said to be a *chi-square random variable*, or to have a *chi-square distribution*, with m degrees of freedom (*df*). The PDF of V is as follows:

$$f(x) \;=\; \frac{1}{2^{m/2}\Gamma(m/2)}\, x^{(m/2)-1}\, e^{-x/2} \quad \text{when } x > 0 \text{ and } 0 \text{ otherwise.}$$

Typical forms of the PDF and CDF of V are shown below:





The location of the median, $\chi^2_{0.5}$, has been labeled in each plot.

Notes:

1. *Gamma Subfamily:* The chi-square family of distributions is a subfamily of the gamma family, where $\alpha = \frac{m}{2}$ and $\beta = 2$.
2. *Mean and Variance:* For each m, $E(V) = m$ and $Var(V) = 2m$.
3. *Shape:* The parameter m governs the shape of the distribution. As m increases, the shape becomes more symmetric. For large m, the distribution is approximately normal.
4. *Independent Sums:* If V_1 and V_2 are independent chi-square random variables with m_1 and m_2 degrees of freedom, respectively, then the sum V_1+V_2 has a chi-square distribution with $m_1 + m_2$ degrees of freedom.
5. *Normal Random Samples:* If X_1, X_2, …, X_n is a random sample of size n from a normal distribution with mean μ and standard deviation σ, then
$$V = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$
is a chi-square random variable with n degrees of freedom.
6. *Quantiles:* The notation χ^2_p is used to denote the p^{th} quantile ($100p^{\text{th}}$ percentile) of the chi-square distribution. A table with quantiles corresponding to
$$p = 0.005,\ 0.010,\ 0.025,\ 0.050,\ 0.100,\ 0.900,\ 0.950,\ 0.975,\ 0.990,\ 0.995$$
for various degrees of freedom is given on page A8 (Table 3) in the Rice textbook. An extended table is given below.

Reference table of selected quantiles for chi-square distributions. Let V be a chi-square random variable with df degrees of freedom.

(1) The following table gives selected quantiles of the V distribution when $df = 1$:

p	0.005	0.010	0.025	0.050	0.100
χ^2_p	0.000039	0.00016	0.00098	0.0039	0.016

p	0.900	0.950	0.975	0.990	0.995
χ^2_p	2.71	3.84	5.02	6.63	7.88

(2) The following tables give selected quantiles of the V distribution when $df > 1$:

df	$\chi^2_{0.005}$	$\chi^2_{0.01}$	$\chi^2_{0.025}$	$\chi^2_{0.05}$	$\chi^2_{0.10}$	$\chi^2_{0.90}$	$\chi^2_{0.95}$	$\chi^2_{0.975}$	$\chi^2_{0.99}$	$\chi^2_{0.995}$
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67

df	$\chi^2_{0.005}$	$\chi^2_{0.01}$	$\chi^2_{0.025}$	$\chi^2_{0.05}$	$\chi^2_{0.10}$	$\chi^2_{0.90}$	$\chi^2_{0.95}$	$\chi^2_{0.975}$	$\chi^2_{0.99}$	$\chi^2_{0.995}$
31	14.46	15.66	17.54	19.28	21.43	41.42	44.99	48.23	52.19	55.00
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49	56.33
33	15.82	17.07	19.05	20.87	23.11	43.75	47.40	50.73	54.78	57.65
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06	58.96
35	17.19	18.51	20.57	22.47	24.80	46.06	49.80	53.20	57.34	60.27
36	17.89	19.23	21.34	23.27	25.64	47.21	51.00	54.44	58.62	61.58
37	18.59	19.96	22.11	24.07	26.49	48.36	52.19	55.67	59.89	62.88
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16	64.18
39	20.00	21.43	23.65	25.70	28.20	50.66	54.57	58.12	62.43	65.48
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
41	21.42	22.91	25.21	27.33	29.91	52.95	56.94	60.56	64.95	68.05
42	22.14	23.65	26.00	28.14	30.77	54.09	58.12	61.78	66.21	69.34
43	22.86	24.40	26.79	28.96	31.63	55.23	59.30	62.99	67.46	70.62
44	23.58	25.15	27.57	29.79	32.49	56.37	60.48	64.20	68.71	71.89
45	24.31	25.90	28.37	30.61	33.35	57.51	61.66	65.41	69.96	73.17
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.20	74.44
47	25.77	27.42	29.96	32.27	35.08	59.77	64.00	67.82	72.44	75.70
48	26.51	28.18	30.75	33.10	35.95	60.91	65.17	69.02	73.68	76.97
49	27.25	28.94	31.55	33.93	36.82	62.04	66.34	70.22	74.92	78.23
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
55	31.73	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29	85.75
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
65	39.38	41.44	44.60	47.45	50.88	79.97	84.82	89.18	94.42	98.11
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43	104.21
75	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.84	106.39	110.29
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32
85	55.17	57.63	61.39	64.75	68.78	102.08	107.52	112.39	118.24	122.32
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12	128.30
95	63.25	65.90	69.92	73.52	77.82	113.04	118.75	123.86	129.97	134.25
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	140.17
105	71.43	74.25	78.54	82.35	86.91	123.95	129.92	135.25	141.62	146.07
110	75.55	78.46	82.87	86.79	91.47	129.39	135.48	140.92	147.41	151.95
115	79.69	82.68	87.21	91.24	96.04	134.81	141.03	146.57	153.19	157.81
120	83.85	86.92	91.57	95.70	100.62	140.23	146.57	152.21	158.95	163.65
125	88.03	91.18	95.95	100.18	105.21	145.64	152.09	157.84	164.69	169.47
130	92.22	95.45	100.33	104.66	109.81	151.05	157.61	163.45	170.42	175.28
135	96.43	99.74	104.73	109.16	114.42	156.44	163.12	169.06	176.14	181.07
140	100.65	104.03	109.14	113.66	119.03	161.83	168.61	174.65	181.84	186.85
145	104.89	108.35	113.56	118.17	123.65	167.21	174.10	180.23	187.53	192.61
150	109.14	112.67	117.98	122.69	128.28	172.58	179.58	185.80	193.21	198.36

1.3.7 Student t Distribution; Table of Quantiles

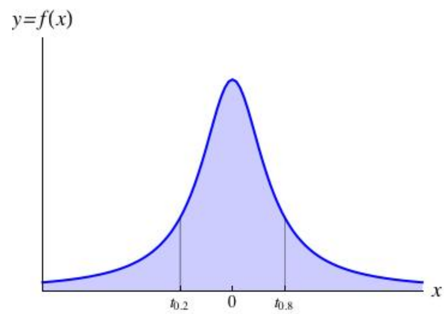
Assume that Z is a standard normal random variable, V is a chi-square random variable with m degrees of freedom, and Z and V are independent. Then

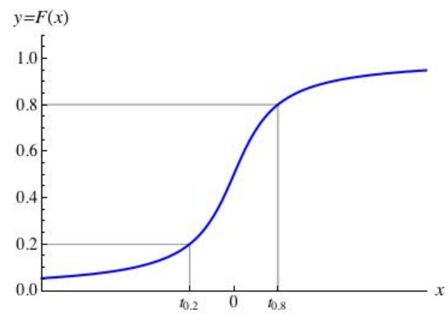
$$T = \frac{Z}{\sqrt{V/m}}$$

is said to be a *Student t random variable*, or to have a *Student t distribution*, with m degrees of freedom (*df*). The PDF of T is as follows:

$$f(x) = \frac{\Gamma((m+1)/2)}{\sqrt{m\pi}\Gamma(m/2)} \left(\frac{m}{m+x^2}\right)^{(m+1)/2} \quad \text{for all real numbers } x.$$

Typical forms of the PDF and CDF of T are shown below:





The symmetric locations of the 20^{th} and 80^{th} percentiles are marked in each plot.

Notes:

1. *Shape and Summary Measures:* The distribution of T is symmetric around 0.

   The median of the distribution is 0, and the mean is 0 as long as $m > 1$.

   When $m > 2$, $Var(T) = \frac{m}{m-2}$.

   For large m, T has an approximate standard normal distribution.

2. *Quantiles:* The notation t_p is used to denote the pth quantile (100pth percentile) of the Student t distribution. A table with quantiles corresponding to

   $$p = 0.60,\ 0.70,\ 0.80,\ 0.90,\ 0.95,\ 0.975,\ 0.99,\ 0.995$$

   for various degrees of freedom is given on page A9 (Table 4) of the Rice textbook. An extended table is given below.

   Since the Student t distribution is symmetric around zero, $t_{1-p} = -t_p$.

Reference table for selected quantiles of Student t distributions. Let T be a Student t random variable with *df* degrees of freedom.

- The following tables give selected quantiles of the T distribution when $p > 0.50$.
- For $p < 0.50$, use $t_p = -t_{1-p}$.
- The $df = \infty$ row of the table corresponds to quantiles of the standard normal distribution.

df	$t_{0.60}$	$t_{0.70}$	$t_{0.80}$	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750
31	0.256	0.530	0.853	1.309	1.696	2.040	2.453	2.744
32	0.255	0.530	0.853	1.309	1.694	2.037	2.449	2.738
33	0.255	0.530	0.853	1.308	1.692	2.035	2.445	2.733
34	0.255	0.529	0.852	1.307	1.691	2.032	2.441	2.728
35	0.255	0.529	0.852	1.306	1.690	2.030	2.438	2.724
36	0.255	0.529	0.852	1.306	1.688	2.028	2.434	2.719
37	0.255	0.529	0.851	1.305	1.687	2.026	2.431	2.715
38	0.255	0.529	0.851	1.304	1.686	2.024	2.429	2.712
39	0.255	0.529	0.851	1.304	1.685	2.023	2.426	2.708
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704

df	$t_{0.60}$	$t_{0.70}$	$t_{0.80}$	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
41	0.255	0.529	0.850	1.303	1.683	2.020	2.421	2.701
42	0.255	0.528	0.850	1.302	1.682	2.018	2.418	2.698
43	0.255	0.528	0.850	1.302	1.681	2.017	2.416	2.695
44	0.255	0.528	0.850	1.301	1.680	2.015	2.414	2.692
45	0.255	0.528	0.850	1.301	1.679	2.014	2.412	2.690
46	0.255	0.528	0.850	1.300	1.679	2.013	2.410	2.687
47	0.255	0.528	0.849	1.300	1.678	2.012	2.408	2.685
48	0.255	0.528	0.849	1.299	1.677	2.011	2.407	2.682
49	0.255	0.528	0.849	1.299	1.677	2.010	2.405	2.680
50	0.255	0.528	0.849	1.299	1.676	2.009	2.403	2.678
51	0.255	0.528	0.849	1.298	1.675	2.008	2.402	2.676
52	0.255	0.528	0.849	1.298	1.675	2.007	2.400	2.674
53	0.255	0.528	0.848	1.298	1.674	2.006	2.399	2.672
54	0.255	0.528	0.848	1.297	1.674	2.005	2.397	2.670
55	0.255	0.527	0.848	1.297	1.673	2.004	2.396	2.668
56	0.255	0.527	0.848	1.297	1.673	2.003	2.395	2.667
57	0.255	0.527	0.848	1.297	1.672	2.002	2.394	2.665
58	0.255	0.527	0.848	1.296	1.672	2.002	2.392	2.663
59	0.254	0.527	0.848	1.296	1.671	2.001	2.391	2.662
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660
65	0.254	0.527	0.847	1.295	1.669	1.997	2.385	2.654
70	0.254	0.527	0.847	1.294	1.667	1.994	2.381	2.648
75	0.254	0.527	0.846	1.293	1.665	1.992	2.377	2.643
80	0.254	0.526	0.846	1.292	1.664	1.990	2.374	2.639
85	0.254	0.526	0.846	1.292	1.663	1.988	2.371	2.635
90	0.254	0.526	0.846	1.291	1.662	1.987	2.368	2.632
95	0.254	0.526	0.845	1.291	1.661	1.985	2.366	2.629
100	0.254	0.526	0.845	1.290	1.660	1.984	2.364	2.626
105	0.254	0.526	0.845	1.290	1.659	1.983	2.362	2.623
110	0.254	0.526	0.845	1.289	1.659	1.982	2.361	2.621
115	0.254	0.526	0.845	1.289	1.658	1.981	2.359	2.619
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617
125	0.254	0.526	0.845	1.288	1.657	1.979	2.357	2.616
130	0.254	0.526	0.844	1.288	1.657	1.978	2.355	2.614
135	0.254	0.526	0.844	1.288	1.656	1.978	2.354	2.613
140	0.254	0.526	0.844	1.288	1.656	1.977	2.353	2.611
145	0.254	0.526	0.844	1.287	1.655	1.976	2.352	2.610
150	0.254	0.526	0.844	1.287	1.655	1.976	2.351	2.609
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576

1.3.8 F Ratio Distribution

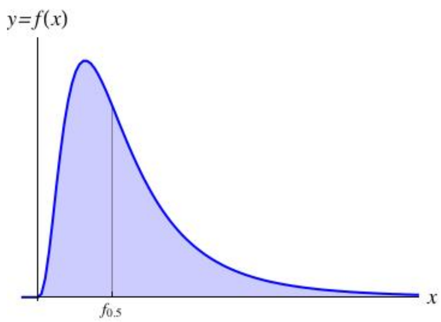
Let U and V be independent chi-square random variables with n_1 and n_2 degrees of freedom, respectively. Then

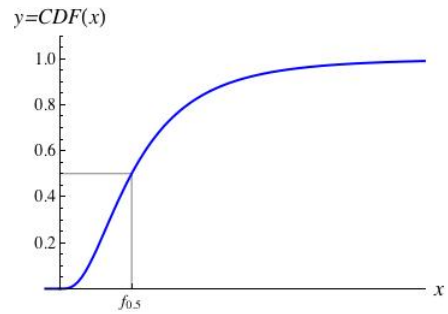
$$F = \frac{U/n_1}{V/n_2}$$

is said to be an *f ratio random variable*, or to have an *f ratio distribution*, with n_1 and n_2 degrees of freedom (*df*). The PDF of F is as follows:

$$f(x) = \frac{\Gamma((n_1+n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} \left(\frac{n_1}{n_2}\right)^{n_1/2} \left(\frac{n_2}{n_2+n_1 x}\right)^{(n_1+n_2)/2} \quad \text{when } x > 0, \text{ and } 0 \text{ otherwise.}$$

Typical forms for the PDF and CDF of F are shown below.





The location of the median, $f_{0.5}$, has been labeled in each plot.

Notes:

1. *"Fisher" Ratio Distribution:* The f in "f ratio distribution" is for R.A. Fisher, who pioneered its use in analyzing the results of comparative studies (that is, in analyzing the results of studies comparing two or more samples).

2. *Shape:* Both parameters govern shape and scale. If $n_2 > 2$, then $E(F) = \frac{n_2}{n_2-2}$; otherwise the mean is indeterminate. Note that $E(F) \to 1$ as $n_2 \to \infty$.

   If $n_2 > 4$, then $Var(F) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-4)(n_2-2)^2}$; otherwise, the variance is indeterminate.

3. *Reciprocal:* If F has an f ratio distribution with n_1 and n_2 degrees of freedom, then the reciprocal of F has an f ratio distribution with n_2 and n_1 degrees of freedom.

4. *Quantiles:* The notation f_p is used to denote the pth quantile (100pth percentile) of the f ratio distribution. The Rice textbook includes tables for $p = 0.90$ (page A10), $p = 0.95$ (page A11), $p = 0.975$ (page A12), and $p = 0.99$ (page A13).

   The $p = 0.10$,0.05, 0.025, 0.01 quantiles can be computed using reciprocals. Specifically,

$$f_p \text{ on } n_1,\, n_2 \ df \quad = \quad \frac{1}{f_{1-p} \text{ on } n_2,\, n_1 \ df}$$

To illustrate the use of the tables in the textbook, let $n_1 = 8$ and $n_2 = 10$. Then

1. When $p = 0.90$, 0.95, 0.975 and 0.99, the values are read from the tables:

$$f_{0.90} = 2.38, \quad f_{0.95} = 3.07, \quad f_{0.975} = 3.85, \quad f_{0.99} = 5.06.$$

2. When $p = 0.10$, 0.05, 0.025, and 0.01, the quantiles are computed using reciprocals. Specifically, since

$$P(F \le x) = P\left(\frac{1}{F} \ge \frac{1}{x}\right) \quad \text{for every } x,$$

to obtain the 0.10, 0.05, 0.025, and 0.01 quantiles of the distribution with 8 degrees of freedom in the numerator and 10 degrees of freedom in the denominator, we use the reciprocals of the 0.90, 0.95, 0.975, and 0.99 quantiles of the f ratio distribution with 10 degrees of freedom in the numerator and 8 degrees of freedom in the denominator.

Thus,

$$f_{0.10} = \frac{1}{2.54} = 0.39, \; f_{0.05} = \frac{1}{3.35} = 0.30, \; f_{0.025} = \frac{1}{4.30} = 0.23, \; f_{0.01} = \frac{1}{5.81} = 0.17.$$

1.3.9 Confidence Interval Procedures

Sampling from one normal distribution. Let $\overline{X}$ be the sample mean and S^2 be the sample variance of a random sample of size n from a normal distribution with mean μ and standard deviation σ.

1. $100(1-\alpha)\%$ *CI for* μ*, when* σ *is known:*

$$\overline{X} \; \pm \; z(\alpha/2) \sqrt{\frac{\sigma^2}{n}},$$

where $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

2. $100(1-\alpha)\%$ *CI for* μ*, when* σ *is estimated:*

$$\overline{X} \; \pm \; t_{n-1}(\alpha/2) \sqrt{\frac{S^2}{n}},$$

where $t_{n-1}(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the Student t distribution with $n-1$ *df*.

3. $100(1-\alpha)\%$ *CI for* σ^2*, when* μ *is known:*

$$\left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_n(\alpha/2)}, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_n(1-\alpha/2)}\right],$$

where $\chi^2_n(p)$ is the $100(1-p)\%$ point of the chi-square distribution with n *df*.

4. $100(1-\alpha)\%$ *CI for* σ^2*, when* μ *is estimated:*

$$\left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\chi^2_{n-1}(\alpha/2)},\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\chi^2_{n-1}(1-\alpha/2)}\right],$$

where $\chi^2_{n-1}(p)$ is the $100(1-p)\%$ point of the chi-square distribution with $(n-1)$ *df.*

Note that $\sum_i(X_i-\overline{X})^2$ is the same as $(n-1)S^2$.

5. $100(1-\alpha)\%$ *CI for* σ*:*

If $[L,U]$ is a $100(1-\alpha)\%$ CI for σ^2, then $[\sqrt{L},\sqrt{U}]$ is a $100(1-\alpha)\%$ CI for σ.

Large sample approximate intervals based on Fisher's theorem. Under the conditions of Fisher's theorem, an approximate $100(1-\alpha)\%$ confidence interval for θ has the following form:

$$\widehat{\theta}\ \pm\ z(\alpha/2)\sqrt{\frac{1}{nI(\widehat{\theta})}}$$

where $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

Several special cases are of interest:

1. Bernoulli Distribution. Let $Y=\sum_{i=1}^{n}X_i$ be the sample sum of a random sample of size n from a Bernoulli distribution with parameter p, where $p\in(0,1)$. If n is large and $0<Y<n$, then an approximate $100(1-\alpha)\%$ confidence interval for p has the following form:

$$\widehat{p}\ \pm\ z(\alpha/2)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \qquad \text{where } \widehat{p}=\frac{Y}{n}$$

and $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

Note that the half-width of the confidence interval, $z(\alpha/2)\sqrt{\widehat{p}(1-\widehat{p})/n}$, is often called the *margin of error* in survey analysis; unless otherwise specified, $\alpha=0.05$ is used to calculate the margin of error.

2. Poisson Distribution. Let $Y=\sum_{i=1}^{n}X_i$ be the sample sum of a random sample of size n from a Poisson distribution with parameter λ. If n is large and $Y>0$, then an approximate $100(1-\alpha)\%$ confidence interval for λ has the following form:

$$\widehat{\lambda}\ \pm\ z(\alpha/2)\sqrt{\frac{\widehat{\lambda}}{n}} \qquad \text{where } \widehat{\lambda}=\frac{Y}{n}$$

and $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

3. Exponential Distribution. Let $Y=\sum_{i=1}^{n}X_i$ be the sample sum of a random sample of size n from an exponential distribution with parameter λ. If n is large, then an approximate $100(1-\alpha)\%$ confidence interval for λ has the following form:

$$\widehat{\lambda}\ \pm\ z(\alpha/2)\sqrt{\frac{(\widehat{\lambda})^2}{n}} \qquad \text{where } \widehat{\lambda}=\frac{n}{Y}$$

and $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

Independent sampling from two normal distributions. Let $\overline{X}$ and S_x^2 be the sample mean and sample variance, respectively, of a random sample of size n from a normal distribution with mean μ_x and standard deviation σ_x. Let $\overline{Y}$ and S_y^2 be the sample mean and sample variance, respectively, of a random sample of size m from a normal distribution with mean μ_y and standard deviation σ_y. Assume that the samples were chosen independently.

1. $100(1-\alpha)\%$ CI for $\mu_x - \mu_y$, when σ_x, σ_y are known:

$$\left(\overline{X} - \overline{Y}\right) \;\pm\; z(\alpha/2)\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$$

where $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

2. $100(1-\alpha)\%$ CI for $\mu_x - \mu_y$, when $\sigma_x = \sigma_y$ is estimated:

$$\left(\overline{X} - \overline{Y}\right) \;\pm\; t_{n+m-2}(\alpha/2)\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}\,, \quad \text{where} \quad S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

and $t_{n+m-2}(\alpha/2)$ is the $100(1-\alpha/2)\%$ point on the Student t distribution with $(n+m-2)$ df.

3. Approximate $100(1-\alpha)\%$ CI for $\mu_x - \mu_y$, when $\sigma_x \neq \sigma_y$ are estimated:

$$\left(\overline{X} - \overline{Y}\right) \;\pm\; t_{df}(\alpha/2)\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}\,, \quad \text{where} \quad df = \frac{\left((S_x^2/n) + (S_y^2/m)\right)^2}{\left((S_x^2/n)^2/n + (S_y^2/m)^2/m\right)} - 2$$

and $t_{df}(\alpha/2)$ is the $100(1-\alpha/2)\%$ point on the Student t distribution with df df.

4. $100(1-\alpha)\%$ CI for σ_x^2/σ_y^2, when μ_x and μ_y are estimated:

$$\left[\frac{S_x^2/S_y^2}{f_{n-1,m-1}(\alpha/2)},\ \frac{S_x^2/S_y^2}{f_{n-1,m-1}(1-\alpha/2)}\right]$$

where $f_{n-1,m-1}(p)$ is the $100(1-p)\%$ point of the f ratio distribution with $(n-1)$ and $(m-1)$ *df*.

1.3.10 Hypothesis Test Procedures

Sampling from one normal distribution. Let $\overline{X}$ be the sample mean and S^2 be the sample variance of a random sample of size n from a normal distribution with mean μ and standard deviation σ.

1. Size α tests of $H_O : \mu = \mu_o$:

	σ Known	*σ Estimated*
Test Statistic	$Z = \dfrac{\overline{X} - \mu_o}{\sqrt{\sigma^2/n}}$	$T = \dfrac{\overline{X} - \mu_o}{\sqrt{S^2/n}}$
RR for $H_A : \mu < \mu_o$	$Z \leq -z(\alpha)$	$T \leq -t_{n-1}(\alpha)$
RR for $H_A : \mu > \mu_o$	$Z \geq z(\alpha)$	$T \geq t_{n-1}(\alpha)$
RR for $H_A : \mu \neq \mu_o$	$\lvert Z\rvert \geq z(\alpha/2)$	$\lvert T\rvert \geq t_{n-1}(\alpha/2)$

$z(p)$ is the $100(1-p)\%$ point of the standard normal distribution; $t_{n-1}(p)$ is the $100(1-p)\%$ point of the Student t distribution with $n-1$ degrees of freedom. Each test is a likelihood ratio test; the one-sided tests are uniformly most powerful.

2. Size α Tests of $H_O : \sigma^2 = \sigma_o^2$:

	μ Known	*μ Estimated*
Test Statistic	$V = \dfrac{1}{\sigma_o^2}\sum_{i=1}^{n}(X_i - \mu)^2$	$V = \dfrac{1}{\sigma_o^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$
RR for $H_A : \sigma^2 < \sigma_o^2$	$V \leq \chi^2_n(1-\alpha)$	$V \leq \chi^2_{n-1}(1-\alpha)$
RR for $H_A : \sigma^2 > \sigma_o^2$	$V \geq \chi^2_n(\alpha)$	$V \geq \chi^2_{n-1}(\alpha)$
RR for $H_A : \sigma^2 \neq \sigma_o^2$	$V \leq \chi^2_n(1-\alpha/2)$ or $V \geq \chi^2_n(\alpha/2)$	$V \leq \chi^2_{n-1}(1-\alpha/2)$ or $V \geq \chi^2_{n-1}(\alpha/2)$

$\chi^2_{df}(p)$ is the $100(1-p)\%$ point of the chi-square distribution with *df* degrees of freedom. Each test is a likelihood ratio test; one-sided tests are uniformly most powerful.

Sampling from Bernoulli and Poisson distributions. In each case, $Y = \sum_{i=1}^{n} X_i$ is the sample sum of a random sample of size n from the X distribution. Since Y is discrete, the cutoffs c_p are approximate only. In large samples, the cutoffs from the standard normal distribution are also approximate only.

1. Bernoulli/Bionomial Distributions: Approximate size α tests of $H_O : p = p_o$.

	Small Sample Test	*Large Sample Test*
Test Statistic	$Y = \sum_{i=1}^{n} X_i$	$Z = \frac{Y - np_o}{\sqrt{np_o(1-p_o)}}$
RR for $H_A : p < p_o$	$Y \leq c_\alpha$	$Z \leq -z(\alpha)$
RR for $H_A : p > p_o$	$Y \geq c_{1-\alpha}$	$Z \geq z(\alpha)$
RR for $H_A : p \neq p_o$	$Y \leq c_{\alpha/2}$ or $Y \geq c_{1-\alpha/2}$	$\lvert Z \rvert \geq z(\alpha/2)$

X has a Bernoulli distribution with parameter p, and Y has a binomial distribution with parameters n and p. The approximate cutoffs c_q in the first column satisfy $P_{p_o}(Y \leq c_q) \approx q$; in particular, c_q is computed assuming that $p = p_o$. Each test is a likelihood ratio test; one-sided tests are uniformly most powerful.

If n is large enough so that $E(Y) \geq 10$ and $E(n - Y) \geq 10$ when $p = p_o$, then the large sample test can be used. The cutoff $z(q)$ is the $100(1-q)\%$ point of the standard normal distribution.

2. Poisson Distributions: Approximate size α tests of $H_O : \lambda = \lambda_o$.

	Small Sample Test	*Large Sample Test*
Test Statistic	$Y = \sum_{i=1}^{n} X_i$	$Z = \frac{Y - n\lambda_o}{\sqrt{n\lambda_o}}$
RR for $H_A : \lambda < \lambda_o$	$Y \leq c_\alpha$	$Z \leq -z(\alpha)$
RR for $H_A : \lambda > \lambda_o$	$Y \geq c_{1-\alpha}$	$Z \geq z(\alpha)$
RR for $H_A : \lambda \neq \lambda_o$	$Y \leq c_{\alpha/2}$ or $Y \geq c_{1-\alpha/2}$	$\lvert Z \rvert \geq z(\alpha/2)$

X has a Poisson distribution with parameter λ, and Y has a Poisson distribution with parameter $n\lambda$. The approximate cutoffs c_q in the first column satisfy $P_{\lambda_o}(Y \leq c_q) \approx c_q$; in particular, c_q is computed assuming that $\lambda = \lambda_o$. Each test is a likelihood ratio test; one-sided tests are uniformly most powerful.

If n is large enough so that $E(Y) \geq 100$ when $\lambda = \lambda_o$, then the large sample test can be used. The cutoff $z(q)$ is the $100(1-q)\%$ point of the standard normal distribution.

Independent sampling from two normal distributions. Let $\overline{X}$ and S_x^2 be the sample mean and sample variance, respectively, of a random sample of size n from a normal distribution with mean μ_x and standard deviation σ_x. Let $\overline{Y}$ and S_y^2 be the sample mean and sample variance, respectively, of a random sample of size m from a normal distribution with mean μ_y and standard deviation σ_y. Assume that the samples were chosen independently.

1. Size α tests of $H_O : \delta = \delta_o$, where $\delta = \mu_x - \mu_y$:

	σ_x, σ_y *Known*	$\sigma_x = \sigma_y$ *Estimated*
Test Statistic	$Z = \dfrac{(\overline{X} - \overline{Y}) - \delta_o}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$	$T = \dfrac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}}$
RR for $H_A : \delta < \delta_o$	$Z \leq -z(\alpha)$	$T \leq -t_{n+m-2}(\alpha)$
RR for $H_A : \delta > \delta_o$	$Z \geq z(\alpha)$	$T \geq t_{n+m-2}(\alpha)$
RR for $H_A : \delta \neq \delta_o$	$\lvert Z \rvert \geq z(\alpha/2)$	$\lvert T \rvert \geq t_{n+m-2}(\alpha/2)$

$S_p^2 = ((n-1)S_x^2 + (m-1)S_y^2)/(n+m-2)$ is the pooled estimate of the common variance.

$z(p)$ is the $100(1-p)\%$ point of the standard normal distribution; $t_{n+m-2}(p)$ is the $100(1-p)\%$ point of the Student t distribution with $n+m-2$ degrees of freedom.

Each test is a likelihood ratio test; one-sided tests are uniformly most powerful.

2. Approximate size α tests of $H_O : \delta = \delta_o$, where $\delta = \mu_x - \mu_y$:

	$\sigma_x \neq \sigma_y$ *Estimated*
Test Statistic	$T = \dfrac{(\overline{X} - \overline{Y}) - \delta_o}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$
RR for $H_A : \delta < \delta_o$	$T \leq -t_{df}(\alpha)$
RR for $H_A : \delta > \delta_o$	$T \geq t_{df}(\alpha)$
RR for $H_A : \delta \neq \delta_o$	$\lvert T \rvert \geq t_{df}(\alpha/2)$

$df = -2 + \Big((S_x^2/n) + (S_y^2/m)\Big)^2 \Big/ \Big((S_x^2/n)^2/n + (S_y^2/m)^2/m\Big)$, and $t_{df}(p)$ is the $100(1-p)\%$ point of the Student t distribution with df degrees of freedom.

3. $100\alpha\%$ *tests of* $H_O : r = r_o$*, where* $r = \sigma_x^2/\sigma_y^2$*:*

	μ_x*,* μ_y *Estimated*
Test Statistic:	$F = \dfrac{S_x^2/S_y^2}{r_o}$
RR for $H_A : r < r_o$*:*	$F \le f_{n-1,m-1}(1-\alpha)$
RR for $H_A : r > r_o$*:*	$F \ge f_{n-1,m-1}(\alpha)$
RR for $H_A : r \neq r_o$*:*	$F \le f_{n-1,m-1}(1-\alpha/2)$ or $F \ge f_{n-1,m-1}(\alpha/2)$

$f_{n-1,m-1}(p)$ is the $100(1-p)\%$ point of the f ratio distribution with $(n-1)$ and $(m-1)$ *df*.

Each test is a likelihood ratio test; one-sided tests are uniformly most powerful.