CHAPTER FIVE

GENERATING RANDOM VARIABLES

5.1 The Inverse

If f is a function that maps $A \rightarrow B$ and is one-to-one, we can find an inverse function g which maps $B \rightarrow A$. Take ordinary addition. Let $f = x + 2$. The inverse of addition is subtraction. Let $g = x - 2$. Take an arbitrary $x = 10$. $f(10) = 12$ and $g(12) = 10$, and you have gotten the element that you started with. The inverse function of a function f is often written f^{-1} with the obvious property that $f(f^{-1}(x)) = f^{-1}(f(x)) = x$.

Consider some common functions such as $y = x^2$, $y = \sin x$, $y = mx + b$. In order to have an inverse, the first function $y = x^2$ has to be defined for either $x \geq 0$ or $x \leq 0$. Then it is 1-1 and has an inverse.

The inverse function for $f(x) = x^2$ is $g = \sqrt{x}$ if $x \geq 0$. To verify this, $f(g(x)) = f(\sqrt{x}) = (\sqrt{x})^2 = x$; $g(f(x)) = \sqrt{x^2} = x$. Again, if we did not restrict the domain to either $x \geq 0$ or $x \leq 0$, we would not obtain a unique inverse. For example, if $f(x) = x^2 = 4$, x could be either +2 or -2. In order to have a unique value for the inverse function, we restrict the domain.

Similarly, $y = \sin x$ has to be defined on an interval such as $-\pi \leq x < \pi$ or $0 \leq x < 2\pi$ in order for the function to be 1-1 and have an inverse $g = \text{arc sin}(x) = \sin^{-1} x$. (Note that $\sin^{-1} x$ refers to the angle whose sine equals x. If we wish to write the reciprocal of sin x, we use $(\sin x)^{-1} = 1/\sin x$.

The third function, $y = mx + b$, is 1-1 as long as $m \neq 0$. The inverse function is $g = \frac{x - b}{m}$. To check this, calculate $f(g(x)) = f\left(\frac{x - b}{m}\right) = m\left(\frac{x - b}{m}\right) + b = x - b + b = x$.

$\therefore \quad f^{-1} = g = \frac{x - b}{m}$.

Consider one of the most useful functions in statistics and mathematical modeling - $f(x) = y = e^x$. The inverse function is given by $g(x) = \ln x$. Note that $f(g(x)) = e^{\ln x} = x$ and $g (f(x)) = \ln e^x = x$.

EXERCISES 5.1

1) a) Show that $f(x) = x^2 + 1$, $-\infty < x < \infty$, does not have an inverse.

b) How could you restrict the domain to ensure the existence of a unique inverse?

2) Does $f(x) = x^4$ have a unique inverse?

a) over the reals? $-\infty \leq x < \infty$

b) over the complex numbers? $x \in a + bi$, $a, b \in$ reals, $i = \sqrt{-i}$

3) Find the inverse function for $f(x) = \frac{x + 5}{10}$. Show that $f(f^{-1}(x)) = f^{-1}(f(x)) = x$.

4) Derive the inverse function for the function:

$$f(x) = 1 - e^{-x/B}$$

5) Try to solve for the inverse function of the normal density. Highlight the difficulties in obtaining a neat algebraic solution.

5.2 The Inverse Transform Method of Generating Random Variables

We now tie together several chapters of material and show how the computer can generate values for a random variable having a certain probability distribution. As we have seen, we can use a goodness of fit test to determine that the data values conform to the probability distribution. The ***inverse transform*** method is a very useful technique that can be used in either the discrete or the continuous case. The outlined steps are the same and include:

1) Start with the distribution function F(x).

2) Obtain a random number y on [0, 1] (Assume that y is uniformly distributed.)

3) Let $x = F^{-1}(y)$

Then x will have distribution function F(x) or density function f(x).

Continuous Case

Let $f(x) = a x^2$, $0 < x < 2$. Suppose we want to obtain a set of numbers that have distribution function F(x). First we have to solve for a.

$$\int_0^2 a x^2 \, dx = 1 \quad \rightarrow \quad a \frac{x^3}{3} \Big|_0^2 = 1$$

$$a \frac{8}{3} = 1 \quad \rightarrow \quad a = \frac{3}{8}$$

$$\therefore \quad f(x) = \frac{3}{8} x^2 \qquad 0 < x < 2$$

$$= 0 \qquad \text{elsewhere}$$

Follow the steps:

1) $$F(x) = \int_0^x \frac{3}{8} t^2 \, dt = \frac{x^3}{8}$$

2) y = random number on [0, 1].

3) $\frac{x^3}{8} = y \rightarrow x^3 = 8y \rightarrow x = \sqrt[3]{8y} = 2\sqrt[3]{y},$

$F^{-1}(y) = x = 2\sqrt[3]{y}$

Take four random numbers on [0, 1]. Thanks to a pocket scientific calculator, we give them as y values below.

y	$F^{-1}(y) = x$
.247	1.255
.633	1.717
.568	1.656
.548	1.637

$F^{-1}(y)$ has the density function $f(x) = \frac{3}{8} x^2 \quad 0 < x < 2$

$= 0$ elsewhere

This comes from the fact that $F^{-1}(y)$ has the distribution function $F(x) = \frac{x^3}{8}$ and density function $f(x) = F^{-1}(x) = \frac{3}{8} x^2$

We present below an informal proof of why the inverse-transform method works:

1) We want to show that $P(X \leq x) = F(x)$, since this is the definition of the probability distribution function.

2) We know that F has an inverse; otherwise the method could not work.

3) $P(X \leq x) = P(F^{-1}(u) \leq x)$. This is because in our method, $X = F^{-1}(u)$.

4) $P(F^{-1}(u) \leq x) = P(u \leq F(x)) = F(x)$

The last equation holds because any function u is distributed as u (0, 1). For example μ (2, 10) can be thought of as the density function f(x) = 1/b-a = 1/10-2 = 1/8. We also know that if x is a continuous random variable with distributed function F and density function f, that F(x) is uniformly distributed on (0, 1).* Therefore, $0 \leq F(x) \leq 1$ and $P(X \leq x) = F(x)$.

Discrete Case

Suppose a salesperson, Marie, wants to model the number of grenchies that she will sell in a day. The probability of selling zero is 90%, of one grenchy is 5%, of two grenchies is 4%, and of selling three grenchies is 1%. No one has ever sold four or more grenchies in one day. Here is how we generate the random variables, modeling her sales performance. First, we know x is a discrete random variable; for example, Marie can't sell $6\sqrt{3}$ grenchies. The density function f(x) is given in the table below.

x	f(x)
0	.90
1	.05
2	.04
3	.01
4	0

Next, obtain the distribution function F(x), given below.

---

*Hoel, P., Port, S., and Stone, C. *Introduction to Probability Theory*, Houghton Mifflin, Boston, 1971, p. 131.

x	F(x)
0	.90
1	.95
2	.99
3	1.00
4	1.00

Finally, obtain a random number y on [0. 1]. If $0 \leq y \leq .90$, then $x = 0$. If $.90 < y \leq .95$, then $x = 1$. If $.95 < y \leq .99$, then $x = 2$. If $.99 < y \leq 1.0$, then $x = 3$. In this way we can get k random numbers to model Marie's sales performance on k days.

EXERCISES 5.2

1) For the following density function, $f(x) = a\,x^3 \qquad 2 < x < 10$

$= 0 \qquad$ elsewhere

a) Find a.

b) Use the inverse transform method for generating random numbers having F(x) as distribution function. [Input ten random numbers on [0, 1] and substitute to obtain ten values of the given density function.]

2) Use the inverse transform method to generate random variables with the Cauchy density:

$$f(x) = \frac{1}{\pi B}\left[1 + \frac{(x - \alpha)^2}{B}\right]^{-1}$$

$-\infty < \alpha < \infty$

$B > 0$

$-\infty < x < \infty$

3) Use elementary probability theory and the inverse transform method for discrete random variables to obtain random variables for the sum of two dice in "craps."

[P(1) = 0, P(2) = 1/36, . . .]

4) An expert in "cridge" sales estimates the following projections in sales for the coming year:

x	P(x)
less than 100	0
less than 200	1/10
less than 300	3/10
less than 600	1/2
less than 800	3/4
less than 1000	1

Write a discrete inverse transform algorithm to model sales for the coming year.

5) For the exponential distribution f(x) , use the inverse transform method and ten random numbers on [0, 1] to obtain ten values of a random variable having density f(x).

$$f(x) = \frac{1}{B} e^{-x/B} \qquad x \geq 0$$

$$= 0 \qquad \text{elsewhere}$$

Use $\bar{x} = \frac{\sum x}{n}$ from your ten random numbers to estimate B

$$F(x) = \int_0^x \frac{1}{B} e^{-x/B} \qquad x \geq 0$$

5.3 The Acceptance-Rejection Method

Sometimes the inverse transform method is cumbersome since complicated methods from numerical analysis are often necessary. To illustrate, consider:

$$f(x) = a\ x^2\ (1+x) \qquad 0 \le x \le 1$$

$$= 0 \qquad \text{elsewhere}$$

First, to compute a

$$\int_0^1 (a\ x^2 + a\ x^3)\ dx = 1 \quad \rightarrow \quad a\ \left[\frac{x^3}{3} + \frac{x^4}{4} \right] \Bigg|_0^1 = 1$$

$$a = 12/7$$

$$\rightarrow\ f(x) = 12/7\ x^2\ (1+x) \qquad 0 \le x \le 1$$

$$= 0 \qquad \text{elsewhere}$$

To use the inverse transform method,

$$F(x) = \int_0^x \frac{12}{7}\ (t^2 + t^3)\ dt, \qquad 0 \le x \le 1$$

$$F(x) = 0 \qquad x < 0$$

$$F(x) = 1 \qquad x \ge 1$$

$$F(x) = \frac{12}{7} \left[\frac{x^3}{3} + \frac{x^4}{4} \right] \qquad 0 \le x \le 1$$

$$F(x) = 0 \qquad x < 0$$

$$F(x) = 1 \qquad x \ge 1$$

Our next step is to obtain a random number y on [0, 1], for example y = 1/2. Then we solve for $x = F^{-1}(1/2)$:

$$\frac{12}{7} \left[\frac{x^3}{3} + \frac{x^4}{4} \right] = 1/2$$

and all other computations of F^{-1} (y) would require numerical analysis. [Refer to any elementary text on Numerical Analysis.]

The method that generates values of x with desired distributions and avoids cumbersome Numerical Analysis procedures is called the ***acceptance-rejection method***. It is the method of choice when direct methods are inefficient. It has certain requirements that include the following:

(a) f (x) has an upper bound, call it c, over its range.

(b) x has a lower bound and upper bound for the range that we consider, $a \leq x \leq b$, a and b are real numbers. This means that we can't consider intervals such as $0 \leq x < \infty$.

(c) We must select the scale of the graph to make $c\,(b - a) = 1$. It is not required that f (x) take on the value of the bound c. For our picture, it does.

The summary of steps involved in the ***acceptance-rejection*** method are:

1) Find the maximum value c of f (x) on $a \leq x \leq b$.

$\therefore$ $f(x) \leq c \;\forall\; x \in [a, b]$

2) Compute two values μ_1, μ_2 of the uniformly distributed variables, both defined on $[a, b] = [0, 1]$.

3) Let $x_0 = a + \mu_1 (b - a)$

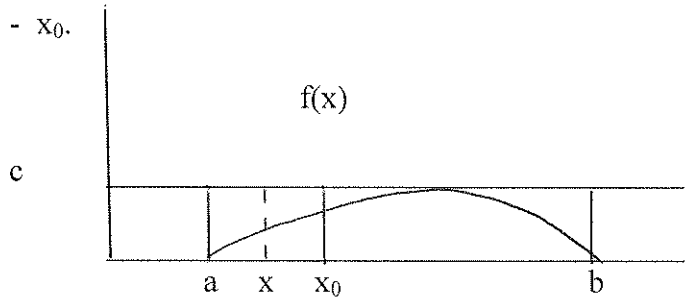
4) Let $y_0 = c\ \mu_2$

5) If $y_0 \leq f(x_0)$, accept x_0 as having the given distribution function F(x); otherwise reject x_0 and repeat the process.

Rationale for the Acceptance-Rejection Method

Though not a formal proof, consider f(x) below:

Let $\Delta x = x - x_0$.



1) We know that $c(b - a) = 1$ [formula for the area of rectangle]. Also $f(x)$ is a density function where the area $= 1$. As previously mentioned, the scale of our graph must permit $c(b - a) = 1$.

2) Consider Δx. If y_0 falls within the rectangle formed by Δx under the curve, then an output $x_0 - \Delta x$ to x_0 will be given.

3) If y_0 falls on or below the curve between 0 and x_0, then outputs will be taken in the range a to x_0.

4) $P(x \leq x_0) =$ prob that y_0 falls under the curve to the left of x_0, given $x \leq x_0$.

5) $P(x < x_0) = F(x_0)$ Prob. y_0 falling on or below the curve to the left of x_0 is the ratio of the area under that part of the curve to the area of the rectangle with sides $(x_0 - a)$ and c.

6) x_0 is uniformly distributed between a and $b \rightarrow$ Prob. that x will be between a and x_0 is $\frac{x_0 - a}{b - a}$.

7) $$F(x_0) = \frac{\int_a^{x_0} f(x)\,dx}{c(x_0 - a)} \cdot \frac{(x_0 - a)}{b - a}$$

Now $c(b - a) = 1 \quad \therefore \quad F(x_0) = \int_a^{x_0} f(x)\, dx$ which completes our informal proof.*

To illustrate the method, consider generating random numbers corresponding to the following probability density function. We will use the acceptance-rejection method.

Example 1: Let $f(x) = 1 + 2x - 2x^2 \qquad 0 \leq x \leq 1$

$= 0 \qquad$ elsewhere

$f'(x) = 2 - 4x = 0 \qquad \rightarrow x = 1/2$

$f(1/2) = 1 + 1 - 2(1/4) = 3/2$

$f''(1/2) = -4$. Therefore, $3/2 = c =$ a maximum.

For our example $a = 0$, $b = 1$, and $c = 3/2$.

To complete our demonstration of the acceptance-rejection method, take two values from $u(0, 1)$. Use previous methods to compute these values. Call the two numbers u_1 and u_2. With a pocket calculator, we could obtain values such as $u_1 = .25$ and $u_2 = .10$. Then consider the following steps:

1) Let $x_0 = a + u_1 (b - a) = .25$

2) Let $y_0 = c\, u_2 = 3/2\,(.10) = .15$

3) $f(x_0) = f(.25) = 1.375$

4) $.15 \leq 1.375$, so accept .25 as x_1 having distribution function $F(x)$. If $y_0 > f(x_0)$, reject x_0 and repeat the process.

---

* Gordon, G. *System Simulation.* Prentice-Hall, Englewood Cliffs, 1978, pp. 138-139.

Consider as a second example the following function:

Example 2: $f(x) = a\ x^4 \qquad 0 \le x \le 2$

$= 0 \qquad$ elsewhere

First calculate a.

$$\int_0^2 a\ x^4\ dx = 1 \quad \rightarrow \quad a\ \frac{x^5}{5}\Big|_0^2 = 1$$

$\therefore\ a\ (32/5) = 1 \quad \rightarrow \quad a = 5/32$

$\therefore$ The density of interest for our acceptance-rejection method is:

$f(x) = (5/32)\ x^4 \qquad 0 \le x \le 2$

$= 0 \qquad$ elsewhere

Obtain two uniform values from u (0, 1). Let $u_1 = .27$ and $u_2 = .09$. Next, we observe that the maximum of f(x) on [0, 2] occurs at $x = 2$. The maximum $c = 5/32 \cdot (16) = 5/2$. Then substitute as follows: [a, b] = [0, 2]

1) Let $x_0 = a + u_1\ (b - a) = 0 + (.27)\ (2) = .54$

2) Let $y_0 = c\ u_2 = 5/32\ (.09) = .01$

3) $f(x_0) = 5/32\ (.54)^4 = .01$

4) $y_0 \le f(x_0)$ since $.01 \le .01$.

Therefore, we accept $x_0 = .54$ as an appropriate random value corresponding to probability density function, $f(x) = (5/32)\ x^4$.

The application of the acceptance-rejection method requires that the probability that x takes on values below a and above b be equal to zero. For probability density functions that are defined with ∞ or $-\infty$ as lower or upper limits, we can truncate at an arbitrary number with large absolute value to minimize error, which is usually small. For example, if we cut off values of x for the

standard normal distribution that are 4 or more standard deviations above the mean, we would neglect less than .0001 of the probability density. To illustrate this notion, refer to a table of the standard normal distribution.

For the standard normal curve above, $P(Z > 3.6) = .0001$. We have lost little, by restricting Z values to a maximum of four standard deviations from the mean.

Note that we changed the interval for our second example so x is defined for $0 \leq x \leq 2$. Our value of c is 5/32. To follow the informal proof, we need to make adjustments so that $c \cdot (b - a) = 1$. In our case $c(b - a) = 5/32\ (2 - 0) = 10/32$. Consequently, we would alter the scale on the x and y axis to ensure that $c(b - a) = 1$.

Discrete Case

For x as a discrete random variable with $P(x_i)$, $i = 0, 1, 2, \ldots$, we let a function (called a ***majorizing function***) $t(x_i)$ be defined. We require that $t(x_i) \geq P(x_i)$ for all

$i = 0, 1, 2, \ldots$ We further let $c = \sum_{i=0}^{\infty} t(x_i)$ and define $r(x_i) = \frac{t(x_i)}{c}$ for $i = 0, 1, 2, \ldots$

We generate our random variables with density $P(x)$ by the following three steps:

1) Generate y with probability mass function r.

2) Generate u_1 from $u(0, 1)$

3) If $u_1 \leq \frac{P(y)}{t(y)}$ accept $x = y$ as a suitably distributed random variable.

Otherwise, repeat the process.* The selection of $t(x_i)$ to fit closely and above $p(x_i)$ plays an important role in the efficiency of the algorithm. For each trial, the probability of acceptance

---

*Law, A. and Kelton, W. *Simulation Modeling and Analysis*. McGraw Hill, New York, 1991, p. 517.

becomes 1/c. Therefore, the probability of rejection becomes 1 - (1/c). By selecting t fitting closely above f, we have significantly improved generation of random variables.* The inverse transform method may be easier to apply to most discrete generation problems.

---

*Law, A. and Kelton, W. *Simulation Modeling and Analysis*. McGraw Hill, New York, 1991, p. 481.

EXERCISES 5.3

1) Use the acceptance-rejection method for generating the following probability density:

[first solve for a]

$$f(x) = a\ x^2 \qquad 2 \leq x < 3$$

$$= 0 \qquad \text{elsewhere}$$

2) Obtain random variables for the density for exercise #1 above using the inverse transform method. Compare the methods for efficiency.

3) Is the inverse transform method or acceptance-rejection method preferable to model the following exponential density?

$$f(x) = e^{-x/B} \qquad x \geq 0$$

$$= 0 \qquad \text{elsewhere}$$

4) How can we adjust the acceptance-rejection method to the following discrete case? Hint:

For $j = 0, 1, 2, \ldots,$ $P(x_j)$ = probability density function. Let $t(x_j) \geq P(x_j)$ for all j.

$$P(x) = p(1 - p)^x \qquad x = 0, 1, 2, \ldots$$

$$= 0 \qquad \text{elsewhere}$$

Let $p = .4$

5) For the example below, use the acceptance-rejection method to generate the given density:

$$f(x) = 60\ x^3 (1 - x)^2 \qquad 0 \leq x \leq 1$$

$$0 \qquad \text{elsewhere}$$

5.4 The Relationship Between the Discrete Poisson Distribution and Continuous Exponential Distribution

One of the most useful relationships in Statistical Modeling is the Poisson-Exponential Distribution connection. The Poisson Distribution is discrete. The exponential distribution is continuous (Calculus based). However, if we consider the interval between consecutive Poisson events, this distribution is exponential.

To show this, consider:

1) Let an event happen at time t

2) Next consider the interval (t, t+x)

3) The Poisson Distribution has mean λ

4) Probability that no Poisson outcome occurs in the interval (t, t+x) $= e^{-\lambda x} \frac{(\lambda x)^0}{0!} = e^{-\lambda x}$

Let x be defined as the interval between Poisson events.

There are no Poisson events in (t, t+x) if $X > x$.

$$P(X > x) = e^{-\lambda x} \quad \text{or} \quad P(X \leq x) = 1 - e^{-\lambda x}$$

Define $P(X \leq x) = \int_0^x f_x(t)\, dt$

$$\frac{d}{dx} \int_0^x f_x(t)\, dt = \frac{d}{dx} (1 - e^{-\lambda x})$$

$$f_x(x) = \lambda e^{-\lambda x} \qquad x > 0$$

Now let us illustrate the relationship between the Poisson Distribution and Exponential Distribution with an excellent real world application.*

---

* Richard J. Larsen and Morris L. Marx, *An Introduction to Mathematical Statistics and Its Applications*, (Prentice Hall, 2001), p. 262.

Over "short" geological periods, a volcano's eruptions are believed to be Poisson events - that is, they are thought to occur independently and at a constant rate. If so, the pdf describing the intervals between eruptions should have the form $f_y(y) = \lambda e^{-\lambda y}$. Collected for the purpose of testing that presumption are the data in the Table 4.2.6, showing the intervals (in months) that elapsed between 37 consecutive eruptions of Mauna Loa, a 14,000 foot volcano in Hawaii (97). During the period covered - 1832 to 1950 - eruptions were occurring at the rate of $\lambda = 0.027$ per month (or once every 3.1 years).

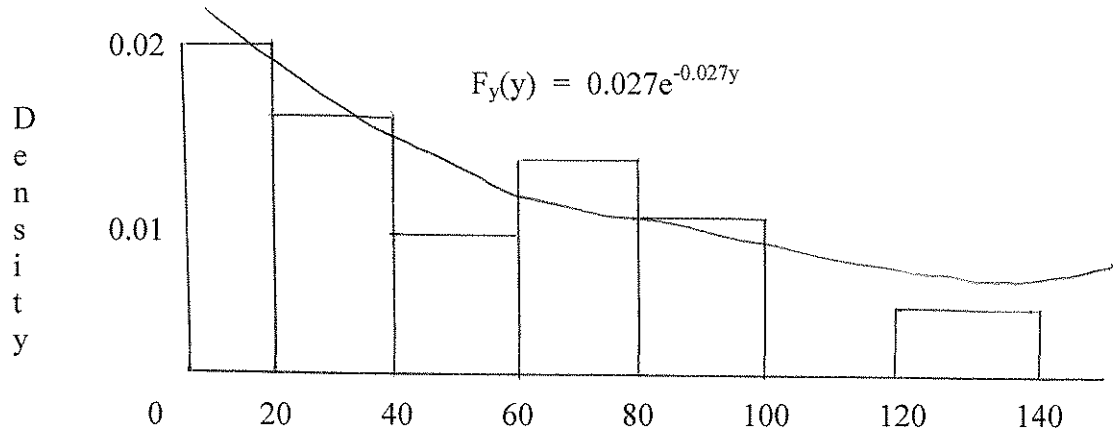
Table 4.2.6

126	73	3	6	37	23
73	23	2	65	94	51
26	21	6	68	16	20
6	18	6	41	40	18
41	11	12	38	77	61
26	3	38	50	91	12

The next step is to create a frequency distribution for the intervals in months between the hypothesized Poisson events. Please note Marx and Larsen's frequency distribution table below:

Interval (mos), y	Frequency	Density
$0 \leq y < 20$	13	0.0181
$20 \leq y < 40$	9	0.0125
$40 \leq y < 60$	5	0.0069
$60 \leq y < 80$	6	0.0083
$80 \leq y < 100$	2	0.0028
$100 \leq y < 120$	0	0.0000
$120 \leq y < 140$	$\frac{1}{36}$	0.0014

Marx and Larsen provided below a graph of the predicted exponential distribution together with the histogram, based upon the frequencies and time interval.



Now let us use calculus and the X^2 Goodness of Fit Test to determine whether the exponential density is a good fit for the data.

Since there were 37 consecutive eruptions of the volcano in 118 years, we can equate $(118) \cdot (12) = 1416$ months. Then divide 1416 by 37, and eruptions occur on average (mean) $\frac{1416}{37} = 38.2703$ months $= 3.1$ years. Eruptions occur at the rate of

$$\lambda = \frac{1}{38.2703} \text{ per month.}$$

$$\lambda = .0261$$

We now use the exponential density $f(y) = \lambda e^{-\lambda y} \quad y > 0$ to compute the expected probabilities for each of the intervals (0,20), (20,40),……(120,140).

A. $P(0 < y < 20) = \lambda e^{-\lambda y}$

$$= \int_0^{20} .0261\ e^{-.0261y}\ dy$$

$$= -e^{-.0261y} \Big|_0^{20}$$

$$= -(e^{-(.0261)(20)} - 1)$$

$$= 1 - .593 = .407$$

B. $P(20 < y < 40) = \int_{20}^{40} .0261\ e^{-.0261y}\ dy$

$$= -[\, e^{-.0261y} \Big|_{20}^{40} \,]$$

$$= -[\, e^{-1.044} - e^{-.522} \,]$$

$$= -[\, .3520 - .5933 \,] = .2413$$

C. $P(40 < y < 60) = -[e^{-.0261(60)} - e^{-.0261(40)}]$

$= -[.0289 - .3520] = .1431$

D. $P(60 < y < 80) = -[e^{-.0261(80)} - e^{-.0261(60)}]$

$= -[.1239 - .2089] = .085$

E. $P(80 < y < 100) = -[e^{-.0261(100)} - e^{-.0261(80)}]$

$= -[.0735 - .1239] = .0504$

F. $P(100 < y < 120) = -[e^{-.0261(120)} - e^{-.0261(100)}]$

$= -[.0436 - .0735] = .0299$

G. $P(120 < y < 140) = -[e^{-.0261(140)} - e^{-.0261(120)}]$

$= -[.0259 - .0436] = .0177$

We can check the soundness of the probabilities by adding the seven probabilities:

$.407 + .2413 + \ldots + .0177 = 1.048$

We have a probability of 105%, not the perfect 100% we would prefer. But the real world rarely coincides with the textbooks. A cumulative probability of 105% simply illustrates the uncertainty inherent in statistical modeling.

Next we can create a table of observed and expected number of eruptions. To illustrate, the observed number of eruptions within time interval (0,20) months is 13. To calculate the expected number of eruptions, multiply the calculated probability $P(0 < y < 20) = .407$ by 37, the total number of eruptions. The expected number for $0 < y < 20$ is $\ell_1 = .407\ (37) = 15.059$. The table for the observed and expected number of eruptions details the seven categories as follows:

	0-20	20-40	40-60	60-80	80-100	100-120	120-140
Observed	13	9	5	6	2	0	1
Expected	15.059	8.9281	5.2947	3,145	4,5843	.9213	.6549

$\ell_1 = .407\,(37) = 15.059$

Also note, it is impossible to have .059 of an eruption. Statistical Modeling always requires a certain element of compromise.

$\ell_2 = .2413\,(37) = 8.9281$

$\ell_3 = .1431\,(37) = 5.2947$

$\ell_4 = 3.145$

$\ell_5 = 4.5843$

$\ell_6 = .9213$

$\ell_7 = .6549$

Next, use the X^2 Goodness of Fit formula: $X^2 = \Sigma \frac{(obs - exp)}{Exp}$

$$X^2 = \frac{(13 - 15.059)^2}{15.059} + \frac{(9 - 8.9281)^2}{8.9281} + \ldots\ldots$$

$X^2 = .2815 + .0006 + .0164 + 2.5917 + 1.4568 + .9213 + .1819 = 5.4502$

Next we look up the critical X^2 value, (K - 1) degrees of freedom, where K = 7 categories.

$K - 1 = 6 = df$

Critical X^2 .05, 6df = 12.6

$5.4502 < 12.6$

Therefore, we accept the null hypothesis. The data is exponentially distributed with 95% confidence or a Type I error of 5%. Another real world issue is the number of intervals. We used

7. You could use 5 or 10 or 20. It is possible to accept the null hypothesis with one arbitrary set of intervals and reject the null hypothesis with a different set - using the same data.

We next turn to the Kolmogorov-Smirnov goodness of fit test. As previously noted, the intervals are not arbitrary, and there is no need to group data. A marvelous feature of the K-S goodness of fit test is that we an simply enter the data and let SPSS do all the work. Let us enter the previous data in SPSS and perform the K-S test.

The reader may feel that a K-S goodness of fit test with the same data is redundant. After all, we have used Calculus, our knowledge of the exponential distribution, and the X^2 goodness of fit test to confirm that the volcano eruption data is exponentially distributed. But, and this is a big BUT, we used 7 intervals. We could have used 10; we could have used 4. Each different set of intervals could result in a different Z or P value.

Redundancy is fine. My multiple regression equations connected with The Sigfluence Generation were complemented by splitting the data into two sets, computing the different multiple regression equations and using Statistical Modeling to test whether the two equations were statistically significantly different. Also, The Sigfluence Generation used TRIAD methodology. We employed data mining, statistics, statistical modeling, and focus groups to conclude that our 18-25 year olds are The Sigfluence Generation. Now let us use SPSS and the K-S goodness of fit test to confirm or reject the goodness of fit of the exponential distribution to our volcano data.

K-S Goodness of Fit Test

First, enter the volcano eruption data with 36 numbers into the first column. Then go to:

Analyze

Nonparametric Tests

1 - Sample K-S

Next, check the test distribution box for normal, uniform and exponential. Click OK. You then will have the following results:

One-Sample Kolomogorov-Smirnov Test

		VAR00001
N		35
Normal Parameters[a,b]	Mean	36.7222
	Std. Deviation	30.54422
Most Extreme Differences	Absolute	.165
	Positive	.165
	Negative	-.128
Kolomogorov-Smirnov Z		.990
Asymp. Sig. (2-tailed)		.281

a. Test distribution is Normal
b. Calculated from data.

One-Sample Kolomogorov-Smirnov Test 2

		VAR00001
N		36
Uniform Parameters[a,b]	Minimum	2.00
	Maximum	126.00
Most Extreme Differences	Absolute	.380
	Positive	.380
	Negative	-.028
Kolomogorov-Smirnov Z		2.280
Asymp. Sig. (2-tailed)		.000

a. Test distribution is Uniform
b. Calculated from data.

One-Sample Kolomogorov-Smirnov Test 4

		VAR00001
N		36
Exponential Parameters[a,b]	Mean	36.7222
Most Extreme Differences	Absolute	.107
	Positive	.050
	Negative	-.107
Kolomogorov-Smirnov Z		.643
Asymp. Sig. (2-tailed)		.803

a. Test distribution is Exponential
b. Calculated from data

The key item is Asymp. Sig. (2-tailed). The high p value of .803 for the exponential distribution confirms that the null hypothesis holds. Our data is exponentially distributed. Note the Asymp. Sig. level of .000 for the uniform distribution. This means that we can confirm with alpha level (Type I error) less than 1% the Alternative Hypothesis. Our data is not uniformly distributed. A full explanation of these essential SPSS features is available in *SPSS for Windows* by Paul Kinnear ad Colin Gray, Psychology Press, 2000.

This is a lot of work, but we can safely use the exponential distribution to model volcano eruption intervals. But be careful. Build redundancy into mathematical modeling. A good fit today does not mean a good fit tomorrow. The real world stubbornly resists our most rigorous efforts to build valid and reliable statistical models.

A promising research avenue in Statistical Modeling is studying the characteristics of data that may yield different goodness of fit results based upon the test used or arbitrary selection of intervals. The real world frequently resists the elegance and logic of pure mathematics.

We next turn to an elementary programming example that can be used to model real world data, that is a good fit to a hypothesized probability density.

5.5 Elementary Programming Example

Generation of Poisson Distributed Arrival Times

The first program is a very short BASIC program to generate, using the inverse transform method, a set of numbers (random variables) that have the Poisson probability distribution. It is assumed that a goodness of fit test has shown that the Poisson distribution is a good fit and that the computer has a random number generator (as the TANDY PC-7 pocket computer which was used for this example).

First, we set the random number $\mu \in (0, 1)$ equal to the distribution function

$$F(x) = \int_0^x \frac{1}{B} e^{-x/B} dx$$

$$\mu = - e^{-x/B} \Big|_0^x = - e^{-x/B} - (-1)$$

$$\mu = F(x) = 1 - e^{-x/B}$$

Solve for x $\rightarrow$ $e^{-x/B} = 1 - \mu \rightarrow x = - B \ln (1 - \mu)$

Take a specific case where $B = .5$.

```
10 LET μ = R (0, 1)
15 LET I = 0
20 LET I = I + 1
25 IF I = 99, GO TO 50
30 LET x = - .5 ln (1 - μ)
35 PRINT x
40 GO TO 10
50 END
```

This elementary program will generate 98 numbers with the Poisson distribution. The first ten random numbers together with their conversion into a Poisson distributed random variable is given.

μ	x
.876	1.044
.763	.720
.878	1.052
.768	.731
.879	1.056
.768	.731
.883	1.073
.487	.334
.936	1.374
.091	.048

$x = -.5 \ln(1 - \mu)$. We could have easily used μ instead of $(1 - \mu)$ since if either is uniformly distributed, so is the other.

Admittedly, the random number function of the TANDY PC-7 is far from random. It is probably based upon linear congruential relationships and a relatively small seed.

5.6 Research Area (Mathematical Statistics)

Dependence and the Central Limit Theorem in Statistics

By this time you have covered the necessary mathematical prerequisites for a current very difficult problem in statistics. In most statistical results, we assume that the sample values are independent of one another. This is very rarely the case in the real world, where people's opinions, results of surveys, and typical measures are dependent on each other and external variables.

Consider the Central Limit Theorem in statistics. This key result states that if $x_1, x_2, \ldots, x_n$ are independent random variables having the same distribution with mean μ, variance σ and moment generating function $M_x(x)$, then if $n \geq 30$, the limiting distribution of the random variable $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is the standard normal distribution.

The moment generating function of X is defined by $M_x(t) = E(e^{tx})$. This gives the following two formulas:

(1) Discrete Random Variables - $M_x(t) = \sum_{i=1}^{n} e^{tx_i} f(x_i)$

(2) Continuous Random Variables - $M_x(t) = \int e^{tx} f(x)\, dx$

This remarkably general result allows us to test whether the mean of a sample is significantly different from the population mean, regardless of the distribution of the random variables. The result also extends to comparing large sample means from two distinct populations and several other important hypothesis tests.

But what happens to the Central Limit Theorem if one of the samples $x_1, x_2, \ldots, x_n$ is a dependent random variable? How do we adjust our α level of Type I error to the typical real world situations of 2-sample hypothesis testing in which one or both samples are dependent?

Tim Sheehan, my graduate student at Iona College, and I collaborated in the following research study. To prepare for reading the following article, review the concepts of:

a) α level - Type I error

b) uniform (0, 1) distribution

c) normal distribution

d) Monte Carlo generation of normal and uniform random variables

e) autocorrelation

f) large sample difference of means hypothesis tests

After reviewing these six topics, write a short program that generates two samples of random numbers on (0, 1), thirty in each sample. For each pair of samples have the program compute the z value:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/30 + \sigma_2^2/30}}$$

$\bar{x}_1$ = mean sample 1; σ_1^2 = variance sample 1;

$\bar{x}_2$ = mean sample 2; σ_2^2 = variance sample 2.

This is the large sample ($n_1 \geq 30$, $n_2 \geq 30$) difference of means hypothesis test. We want to add a line to the program to determine if the mean of population 1 is greater than the mean of population 2 at the α level of .05. This means to count whenever $z > 1.64$.

Perform this procedure 10,000 times as Tim and I did. Compare the percentage of times (using a counter) that $z > 1.64$. We found a value of .054, which was in fundamental agreement with our level of .05. After all, this brief experiment is an applied exercise of what Type I error is. We know that the means of the two samples are equal, since they both come from the same random

number generator. However, our α level is the probability of rejecting H_o and concluding that $\mu_1 > \mu_2$ if $\mu_1 = \mu_2$.

Now that you have tried this simple computer exercise, you are ready for one other experiment that demonstrates the central issue of our article - dependence. Go back to your program and insert a line that multiplies the random numbers by the cheating factor, say 1.1, if the random number $x \in (0, 1)$ is greater than or equal to .5. This line invalidates the Central Limit Theorem and induces autocorrelation in the dependent sample. Leave the other sample intact. Now observe the increase in α level. This is the focus of our article. Go back and measure the average autocorrelation between A_n and A_{n+1} in the dependent sample. This is what Tim and I are discussing in our article. Now you are ready to read our article. Please spend time with each section, since it took us many hours of research to gather results and write the article.

The dependence problem in real world testing is widely known and has resisted a breakthrough. Dr. E. Lehmann (University of California - Berkeley) has proved a remarkable result that you may wish to peruse in his book in greater detail (see the bottom of page 1 of my article for his cited book). Perhaps one of you may contribute to a general result someday. Dr. Lehmann sent me a note discussing the problem and assisting us with references and the advice that we cannot get very far in adjusting hypothesis testing without assuming something about the nature of the dependence. We followed his advice and were very specific in the selection of certain well known probability distributions. We also chose cheating factors which led to the measure of autocorrelation for each dependent sample. We then computed the average autocorrelation for the 10,000 pairs of samples and linked autocorrelation and adjusted α level. The major problem besides the absence of a general result is that in addition to violating the independence assumption

underlying the Central Limit Theorem, we also biased the mean and variance in the dependent sample. This greatly complicated and confounded the issue. This is why we need our future researchers like you to read this article and to become participants. Please read our article which immediately follows.

Permission was obtained from Pergamon Press for the following reprint of the March, 1989 *International Journal of Mathematical and Computer Modeling* article.

DEPENDENT RANDOM VARIABLES AND HYPOTHESIS TESTING - A MONTE CARLO PERSPECTIVE

J. F. Loase[1,2] and T. Sheehan[2]

[1]Westchester Community College, SUNY, Valhalla, NY 10595, USA
[2]Iona College, New Rochelle, NY 10801, USA

(Received February, 1989; accepted for publication March 1989)

Communicated by E. Y. Rodin

Abstract - This research paper simulated hypothesis testing of the differences of means, when the conventional assumption of independence within one of the samples had been violated. The study ran separate Monte Carlo simulations in which both samples came from uniform and normal populations. Dependence was introduced by multiplying the randomly generated scores within one sample by a predetermined factor. Then the simulation collected data on 10,000 paired samples with factors ranging from 1.0 (independence) to 2.0 (the highest level of dependence). A separate study calculated the mean autocorrelation associated with different conditions of dependence and linked this autocorrelation to the adjusted level of Type I error (α level). The results demonstrated a systematic increase in Type I error as the level of autocorrelation increased. The α level that our study found for certain levels of dependence (with $n = 30$) far exceeded the asymptotic level of adjusted α, suggesting that we further explore the effects of autocorrelation on conventional hypothesis testing.

INTRODUCTION

> "Thus one comes to perceive, in the concept of independence, the first germ of the true nature of problems in probability theory." (Kolmogorov)

The purpose of this paper is to simulate hypothesis testing in the real world in a situation where conventional assumptions related to independence and equal means have been violated. To illustrate, we usually test the differences of means of two populations, assuming that the two samples are independent, identically distributed, with equal means and variances. Our study ran Monte Carlo simulations in which both samples came from a uniform distribution and a normal distribution. However, we introduced bias and dependence in one sample and left the other sample intact. Our study then observed the effects on Type I error - the α level of the test - when we employed the conventional t-test for the difference of means between the independent and dependent sample.

Our study used the real world context of cheating to serve as the basis for our model. We considered two groups of students. In the first group, the scores were independent, and there was no cheating. However, in the second group, we constructed a model of cheating and increased selected students' grades accordingly. Of course, conventional hypothesis testing procedures are derived from the Central Limit Theorem (CLT) and most applications of the CLT require

independence.*

Table 1

Nominal Level	Autocorrelation 0	0.1	0.2	0.3	0.4
0.05	0.05	0.0679	0.0896	0.1138	0.1408
0.02	0.02	0.0314	0.0465	0.0655	0.0890
0.01	0.01	0.0163	0.0269	0.0415	0.0610

*In his recent book, Lehmann [2] devoted a section to the effects of dependence on the one-sample t-test. He proved that if $x_1, \ldots, x_n$ were jointly normally distributed with common marginal distribution $N(0, \sigma^2)$ and with correlation coefficients $\rho_y = \text{corr}(x_i, x_j)$. As $n \to \infty$, suppose

(a) $\text{var}\ \bar{x} = \frac{\sigma^2}{n^2} \sum_{i=1}^{n} \sum_{j=1} \rho_y \to \infty$

(b) $\text{var}\ \frac{1}{n} \sum x_i^2 \to 0$ and

(c) $\frac{1}{n} \sum_{i \neq j} \sum \rho_y \to r$

The distribution of the t-statistic tends to the normal distribution $N(0, 1 + r)$.

Lehmann concluded that even under rather weak dependence the previous assumptions are satisfied with $r \neq 0$ and therefore the level of the t-test is quite sensitive to the assumption of independence.

The dependence that we induced was measured by calculating the mean difference between the autocorrelation of the dependent and independent samples. We know that dependence within one sample influences the Type I error of a hypothesis test. We also know from Table 1 above [1]

that the t- test is very sensitive to dependence for a first-order regressive process. The asymptotic levels only give us the adjusted α level as $n \rightarrow x$. We do not know the adjusted α level if n were very large, even if n were 100,000.

Now let us consider our study, where the number of elements within each sample is 30.

CHEATER'S SIMULATION

Our first three studies simulated the dependence that occurs in hypothesis testing for the differences of means where one group is independent and the other group is dependent. To give our study a real-world context, consider a test given to 30 students. Assume the $(n + 1)$th student can only cheat from the nth student if the nth student is above average ($x > 0.5$ from a uniform distribution).

The Monte Carlo program utilized random numbers and simulated the comparison of test scores of two groups of 30 students. The first group scored according to the uniform μ (0, 1) distribution. The second group scores also followed μ (0, 1) except that dependence was introduced. The dependence may be thought as a cheating factor (CF). Each below-average student (score < 0.5) was allowed to cheat if the preceding student was above average. The cheating student increased his/her grade by the multiplication of his score by a variable CF.

The scores of the two groups were then compared. The difference of the means was calculated and tested for statistical significance ($\alpha = 0.05$). The simulation collected data on 10,000 paired samples with CFs of 1.0 - 2.0 with increment 0.1. Both distributions were uniform [0, 1] We refer to this simulaton as Dependence 1.

Dependence 2

Then the model of the CF was modified in that, instead of multiplying the below-average score by a factor, the below-average score is made equal to the preceding above-average score. A simulation of 10,000 runs utilizing this criterion was recorded.

Dependence 3

A modification of this model simply changed the definition of cheater to those who were in the bottom 10 percentile of all students, with the μ (0, 1) distribution. This meant a score < 0.1.

Dependence 4

Finally, we simulated the hypothesis testing of two groups of 30 students in which the scores came from the normal distribution N (50, 20). The dependence was modeled as a CF in which each below-average student (score < 50) was allowed to multiply his/her score by varied CFs, if the preceding score was above average (score > 50).

The difference of means was tested at the $\alpha = 0.05$ level. The simulation ran for 10,000 paired samples, $n = 30$ in each sample. The CF ranged from 1.0 to 2.0 with increment 0.2. The results are given in Tables 2-5.

SIMULATION RESULTS

Table 2 - Dependence 1

CF	Trials with significant positive difference (%)
1.0	5.440
1.1	7.100
1.2	8.210
1.3	9.360

1.4	10.840
1.5	13.330
1.6	14.570
1.7	16.580
1.8	18.220
1.9	20.530
2.0	23.320

Table 3 - Dependence 2

CF	Trials with significant positive difference (%)
Summary of 10,000 trials with copy model of cheating	53.040

Table 4 - Dependence 3

CF	Trials with significant positive difference (%)
Summary of 10,000 trials with copy model of cheating	17.260

Table 5 - Dependence 4

CF	Trials with significant positive difference (%)
1.0	5.780
1.2	10.900
1.4	19.520
1.6	28.040
1.8	37.310
2.0	47.890

COMMENTS - SIMULATION 1

One observes a systematic increase in Type I error as the CF increased. Indeed the 5.44% statistical significance obtained in our first simulation verified the soundness of the program, since we were assuming $\alpha = 0.05$.

A Monte Carlo approach to blackjack and to many statistical questions is appropriate when a neat, elegant mathematical approach is precluded. A Monte Carlo approach to scrutinize the effects of dependence upon Type I error was appropriate due to the unwieldiness of adjusting the CLT to levels of autocorrelation for conventional samples well below asymptomatic levels. The next inquiry was to perform a *post hoc* analysis of the autocorrelation between successive elements in our simulation of dependent variables and attempt to develop a systematic method for linking autocorrelation to the adjusted level.

SIMULATION II

A second Monte Carlo simulation was performed to study the relationship of mean autocorrelation between the pairs of elements in the dependent sample and the eventual level of Type I error. We made a slight variation of the model so that the following represent the dependence relationships that we introduced:

Model 1 - The lower score in the dependent sample was increased by a given percentage ranging from 0 to 1, step 0.1.

Model 2 - The lower score in the dependent sample was increased by a percentage of the difference between the higher and lower score, the percentage ranging from 0 to 1, step 0.1.

The four programs were defined in the following way.

Dependence 5 - Model 1 utilizing scores generated from a uniform distribution.

Dependence 6 - Model 2 using scores generated from a uniform distribution.

Dependence 7 - This program used model 1 and scores obtained from a normal distribution.

Dependence 8 - This simulation utilized model 2 and scores generated from a normal distribution.

SUMMARY OF DATA ANALYSIS

The results in Table 6 below reflect a mean difference of autocorrelations since the simulation of the independent sample results in a consistent (-0.03) autocorrelation. This was likely due to the non-random character of most linear congruential random number generators.

Table 6

CF	Mean difference of means (dependent-independent)	Mean difference of autocorrelation (dependent-independent)	% Type I error in comparison of groups
		Dependence 5	
1.0	0.0013	0.001	5.40
1.1	0.0057	0.011	6.29
1.2	0.0120	0.033	7.44
1.3	0.0185	0.051	8.86
1.4	0.0244	0.072	10.84
1.5	0.0308	0.086	12.55
1.6	0.0357	0.009	14.56
1.7	0.0435	0.112	16.68
1.8	0.0484	0.126	18.34
1.9	0.0533	0.133	20.78
2.0	0.0601	0.143	22.48
		Dependence 6	
0.0	0.0002	-0.004	5.28
0.1	0.0174	0.047	9.02
0.2	0.0334	0.093	13.64
0.3	0.0538	0.151	21.10
0.4	0.0732	0.209	31.55

0.5	0.0986	0.272	44.81
0.6	0.1254	0.340	60.65
0.7	0.1575	0.423	76.99
0.8	0.2008	0.514	90.97
0.9	0.2606	0.627	98.31
1.0	0.3977	0.826	99.95
		Dependence 7	
1.0	-0.0196	-0.003	5.17
1.1	0.7869	0.035	7.43
1.2	1.6818	0.062	10.95
1.3	2.4776	0.095	14.32
1.4	3.3071	0.112	18.51
1.5	4.1865	0.133	23.90
1.6	4.9442	0.144	28.40
1.7	5.7715	0.148	32.95
1.8	6.4587	0.156	37.27
1.9	7.4329	0.137	43.44
2.0	8.1973	0.133	48.10
		Dependence 8	
0.0	-0.0209	0.000	4.84
0.1	1.1876	0.046	8.55
0.2	2.3335	0.097	13.53

0.3	3.6646	0.157	20.85
0.4	4.9516	0.226	30.55
0.5	6.5661	0.297	42.54
0.6	8.4789	0.377	58.66
0.7	10.6870	0.461	74.46
0.8	13.8119	0.567	89.28
0.9	18.3932	0.690	97.28

COMMENTS - SIMULATION II

As Lehmann [2] noted and we have shown, the effect of dependence on the level of the t-test is a serious problem in mathematical statistics. Indeed, Professor Lehmann noted in a letter to the present authors, that we cannot get very far in adjusting hypothesis testing to various levels of dependence without assuming something about the structure of the dependence.

In our second simulation we noted a dramatic increase in Type I error as the mean difference in autocorrelation increased. However, one anomaly that deserves further scrutiny is our model Dependence 7 in which the autocorrelation rises to a maximum and then actually decreases as the CF approaches 2.0. This is probably accounted for by our reliance on a Monte Carlo method which uses random numbers that appear to operate non-randomly, indeed quite systematically. Indeed, the Type I error (Dependence 7) rose predictably, but the drop in autocorrelation appears to be mathematically inexplicable. Note that none of our levels of Type I error were very close to the asymptotic levels of Type I error for first-order autoregressive Gaussian processes, highlighted by Gastwirth and Rubin [1]. This is as expected, since the asymptotic levels of adjusted α were computed as $n \rightarrow \infty$. Our study only utilized samples of 30, which more accurately reflects

conventional hypothesis testing in the real world. Of course, our CF biased the mean and variance of our dependent sample.

Note that the link between mean autocorrelation is systematic in Model 2, in which the lower score is increased by a percentage of the difference between the higher and lower score. Whether the scores were generated from a uniform or normal distribution, approximately equal autocorrelations yielded close adjusted α levels. However, in Model 1, there was a significant discrepancy between the mean autocorrelation and adjusted α level between samples generated from normal or uniform distributions. Of course, Model 2 more accurately reflects cheating in the real world, where cheating from a better student increases your score by a percentage of the difference between the two scores.

DISCUSSION

In our simulation, we have constructed our dependent sample so that only the first observation would have a $\mu\,(0, 1)$ distribution. The remainder would not unless our CF were 1.

To show this, consider the "cheater's simulation" with Dependence 1. To construct the dependent sample, generate a sample of independent $\mu\,(0, 1)$ observations $Y_1, Y_2, \ldots Y_n$ as follows:

$$Y_1 = X_1 - \mu\,(0, 1)$$

$$Y_{i+1} = X_{i+1} \text{ if } X_i \leq 1/2$$

$$= X_{i+1} \text{ if } X_i \geq 1/2, \quad X_{i+1} > 1/2\,, \; i = 1, \ldots n\,(=29)$$

Let $c = CF$

$$Y_{i+1} = c \bullet X_{i+1} \text{ if } X_i > 1/2, \quad X_{i+1} \leq 1/2, \; 1 \leq c \leq 2$$

Consider the distribution of Y_2 which will equal $Y_i\,, \; i = 1, 2, 3, \ldots, n$

Now

$$P(Y_2 < X) = P(X_2 < X \mid X_1 \leq 1/2)\ P(X_1 \leq 1/2)$$

$$+ P(X_2 < X \mid X_1 > 1/2,\ X_2 > 1/2)\ P(X_1 > 1/2,\ X_2 > 1/2)$$

$$+ P(X_2 < X/c \mid X_1 > 1/2,\ X_2 \leq 1/2)\ P(X_1 > 1/2,\ X_2 \leq 1/2)$$

$$= P(X_2 < X)\ 1/2 + P(X_2 < X) / X_2 > 1/2)\ 1/4$$

$$+ P(X_2 < X/c \mid X_2 \leq 1/2)\ 1/4 \qquad (1)$$

Note c is the CF: $P(X_2 < X/c) = P(cX_2 < X)$:

$$P(X_2 < X \mid X_2 > 1/2) = 0 \qquad \text{if } X < 1/2$$

$$= 2\,(X - 1/2) \qquad \text{if } X > 1/2$$

$$P(X_2 < X/c \mid X_2 \leq 0.5) = 1 \qquad \text{if } X/c > 1/2$$

$$= 2X/c \qquad \text{if } X/c \leq 1/2$$

Substitute into expression (1).

$$F_2(X) = P(1/2 < X) = X/2\,(1 + 1/c) \qquad 0 \leq X \leq 1/2$$

$$= X\,(1 + 1/2c) - 1/4 \qquad 1/2 < X \leq c/2$$

$$= X \qquad c/2 < X < 1$$

This distribution function is continuous and is the $\mu\,(0, 1)$ distribution if $c = 1$. Since $Y_2 \geq 0$, compute the expected value:

$$E(Y_2) = \int_0^x [1 - F_2(x)]\ dx$$

$$= \int_0^{1/2} [1 - X/2\,(1 - 1/c)]\ dx + \int_{1/2}^{c/2} [1 - X\,(1 + 1/2c) + 1/4]\ dx$$

$$= \frac{7 + c}{16}$$

Our first sample has mean 1/2; our second sample only has mean 1/2 if c = 1. The factor c introduces a positive bias in sample 2 so that the observed scores from the cheater's group is higher than the mean of the random group.†

A recommended follow-up study would be to design simulations with varying levels of correlation within one sample without biasing the mean or variance of the dependent sample. With the use of Monte Carlo studies of dependent samples, we may be able to discover new relationships between autocorrelations, variance and Type I error. Most importantly, we may be able to adjust hypothesis testing to more accurately model real world phenomena, that are typically dependent.

REFERENCES

1. J. Gastwirth and J. Rubin. Effect of dependence on the level of some one-sample tests. J. Am. Statist. Ass., 66, 336, 818 (1971).

2. E. Lehmann. Testing Statistical Hypotheses, pp. 209-213. Wiley, New York (1986).

---

†The authors wish to thank an anonymous referee for most helpful suggestions for revision of this part of the manuscript.

Student Activity

Now that you have finished reading our article, consider trying to extend our knowledge of adjusting α levels for dependence in one or both samples. Use the inverse transform method to generate samples that are normal, uniform, Poisson, and any other distribution that you care to try. Test your procedure with two samples of thirty and independence. Verify by hypothesis testing that your α level under independence assumptions is close to 5%. Then induce dependence as we did in one or both samples and observe the effect on α level. We need a generalizable result to more

appropriately model real world phenomena, that are typically dependent. But keep in mind that this deep and difficult problem may forever resist a neat solution.

INTERMEDIATE STATISTICS MULTIPLE REGRESSION ANALYSIS OF VARIANCE AND GOODNESS OF FIT TESTS

Monte Carlo Reliability Analysis

You should have already taken my Sigfluence Survey, obtained your three sigfluence scores and better understand the fifty pruned variables. The pruned data set of interest was created after a data mining analysis of the 2500 correlations together with additional analyses. In short, we used SPSS to analyze the most salient dependent variable (y = Potential for Sigfluence) as a multiple regression equation based on response to the following dependent variables:

X1 = Economic status.

X2 = Actual sigfluence.

X3 = Need for sigfluence.

X4 = Level of satisfaction with life as measured by Loase (2002).

We used backward elimination and relied upon the process stopping when the individual parameters partial correlation with the dependent variable became sufficiently small. Economic status was eliminated as a significant variable. Our final set of coefficients were .16 for the constant, .44 for x_2, .23 for x_3, and .01 for x_4.

Next we split the sample into the first 270 cases and repeated the process. Economic status was again deleted. The respective coefficients became .218, .012, .442, and .145. We repeated the process with the second half of the 541 sample data set. Economic status was deleted and the individual coefficients were computed as .112, .013, .440, and .297.

Please visit my website sigfluence.com. You may download my 8th book, *The Sigfluence Generation: Our Young People's Potential to Transform America.* The book, which won a silver

medal in the 2012 Benjamin Franklin National Book Contest (non-fiction), was based upon twenty years of statistical research. You may request the data set of 50 variables / 541 cases that took my two graduate students eighteen months simply to enter

It was natural to speculate as to whether the three results with subsets of the same data set were yielding the same multiple regression equation or more precisely, was there no statistically significant difference in the three multiple regression results. This question has no simple solution in Mathematical Statistics. The deletion of economic status in the three multiple regression analyses suggest that this variable does not enhance the partial correlation with the dependent variable. However, how do we compare multiple regression equations?

Monte Carlo Method 1

We suggest the following. The art and science of Mathematical Modeling and Monte Carlo analysis has advanced so that it is fairly easy to build in a reliability check to multiple linear regression. We pseudo-randomly deleted 10 cases from our data set and obtained the multiple regression equation:

$$Y = .158 + .44\,X2 + .24\,X3 + .01\,X2.$$

We next performed a standard chi-square goodness of fit test using the ten deleted cases as our ten observed values of y and as the values of the independent variables to substitute to obtain the ten expected values of y. Our computed chi-square value was .281, which was extremely small, suggesting a good fit at any conventional level of statistical significance, $p < .01$.

We used the chi-square goodness of fit test, since:

1) Multiple regression assumes that the errors are normally distributed with a mean of zero.

2) If X has the standard normal distribution, then X squared has the chi-square distribution.

As Dr. Lehmann (University of California at Berkeley) discussed with this author, complex psychometric isues, such as the effect of levels of auto-correlation on Type I error, can only be addressed by a partnership among computer scientists, pschometricians, and statisticians. Our two proposed Monte Carlo reliability methods are easy to administer and offer a check of the soundness of multiple regression results. These tools are offered as a preliminary to an eventual software solution to the checking of multiple regression equations, best fitting a given data set.

We have to take advantage of the dramatic recent advances in Mathematical Modeling, data mining, and random number generation to simplify the work of the typical researcher using multiple regression. We also suggest another method to test the reliability of multiple regression equations.

Monte Carlo Method II

The researcher can split the sample data into odd and even cases and obtain two different multiple regression results. Then it is necessary to use Monte Carlo analysis to obtain a random sample of the hypothesized normally distributed independent variables. Each set of pseudo-random independent variables yields a predicted y value for the dependent variable. Then we can employ the simple two sample difference of means formula to determine whether the two distinct sets of y values have equal means (the multiple regression equation replicates) or different means (the two equations are significantly different).

We split our sample in half - 270 and 271 cases. Then we tested whether the independent variables were normally distributed. They were not. If they had been normally distributed, we could have performed a modified inverse transform method, mapping random numbers from zero to

one to an interval (-3, 3), which would have yielded a z score from (-3, 3) and covered 99.7% of the cases. We would have been able to generate sets of randomly generated independent variables. Since the independent variables were not normally distributed, we used SPSS to generate 35 sets of random independent cases and substituted the values in each of the two multiple regression equations. We arrived at 35 values of y, the dependent variable, for each equation. Then we performed an elementary different of means test and calculated a z score of -.95. We accepted the null hypothesis that there was no significant difference in the sets of y values. Both results confirmed the reliability of our computed multiple regression equation and did it simply. We may need a transformation of statistics to provide a theoretical foundation for the two methods this article advances. Future partnerships among Mathematical Modeling, Mathematical Statistics, Computer Science, and Psychometrics are recommended to adjust research results to departures from underlying assumptions.

One caution related to the proposed second method is that the inverse transform of a standard normal distribution is non-trivial to develop without Numerical Analysis. Also Numerical methods require sophisticated error analysis of the associated probabilistic estimates. The proposed partnership of the future should involve computer scientists with expertise in numerical methods. The future beckons with software developed out of a partnership among researchers in Computer Science, Statistics, and Psychometrics to facilitate increasingly rigorous and simple means of testing the reliability and validitiy of our research results.

Student Activity

Please go to the SPSS program. Open the SPSS 50 variable data set and define a different multiple regression set of variables. As a preliminary example, define Quest 6 as the dependent variable. Include EDUCAT and HEALTH as independent variables.

The resulting coefficients are constant = 6.639, EDUCAT 1.909, and HEALTH .627. The small B = .05 for HEALTH, suggests that HEALTH is not a major contributor to the regression equation.

Find a replacement variable for HEALTH and construct the resulting multiple regression equation. Use SPSS to split the 542 data set into 2 sets of 271. Use the earlier described method to test the validity of your multiple regression equation by splitting the data set, constructing separate multiple regression equations, and obtain random sets of numbers for the independent variable, and test whether there is no statistically significant difference between the multiple regression equations.

You can easily delete a small set of data and use (say 10 cases) to substitute for the observed and expected values of y - the dependent variable, utilizing the standard chi-square test. The chi-square test offers another method to test reliability and validity of your multiple regression equation.