AMERICAN SOCIETY OF CIVIL ENGINEERS
HYDROLOGY & HYDRAULIC TECHNICAL GROUP

INVITED LECTURE:

A SHORT COURSE ON PRACTICAL STATISTICS FOR FLOOD CONTROL HYDROLOGY

T.V. Hromadka II

Professor of Mathematics and Environmental Studies,
Department of Mathematics, California State University,
Fullerton, California 92634-9480

THE CENTRAL LIMIT THEOREM

In floodplain management, an underlying technology is the study of statistics. The estimation of 2-year, 10-year, ·100-year and other return frequency peak flow rates are simply applications of well-known statistical concepts. An important fact is if a simple random sample of size n is drawn from a population with mean μ and variance σ^2, then the sample mean $\overline{X}_n$ has approximately a normal distribution with mean μ and variance σ^2/n. That is the distribution function of

$$\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}} \tag{1}$$

is approximately a standard normal. The approximation improves as the sample size increases.

In the above theorem, if the sample were drawn from a population that is in fact a normal distribution, then Equation (1) is exact, and

$$\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}} = Z \tag{2}$$

where Z is the standard normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

Let's see how the Central Limit Theorem applies when we are dealing with a non-normal population.

Example 1.

Samples of size n are drawn from a population having the probability density function

$$f(x) = \begin{cases} \frac{1}{10}e^{-x/10} & x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

(This is the probability density function from an exponential distribution with mean $\mu = 10$, and $\sigma = 10$). The sample mean was computed for each sample. The relative frequency histogram of these mean values for 1000 samples of size $n = 5$ is shown in Figure 1. Figures 2 and 3 show similar results for 1000 samples, but of size $n = 25$ and $n = 100$, respectively. Although all the relative frequency histograms are nearly bell shape, the tendency toward a symmetric normal curve is better for larger n. Also note in Figures 1-3 that the spread of frequency histograms diminishes for larger sample sizes n. A smooth curve drawn through the bar graph of Figure 3 would be nearly the graph of a normal density function with mean 10 and variance $(10)^2/100 = 1$.

The Central Limit Theorem provides a very useful result for statistical inference, for we now know not only that $\overline{X}_n$ has mean μ and variance σ^2/n if the population has mean μ and variance σ^2, but we know also that the probability distribution of $\overline{X}_n$ is approximately normal. For example, suppose we wish to find an interval (a,b) such that

$$P(a \leq \overline{X}_n \leq b) = 0.95 \tag{3}$$

This probability is equivalent to

$$P\left(\frac{a-\mu}{\sigma/\sqrt{n}} \leq \frac{X_n - \mu}{\sigma/\sqrt{n}} \leq \frac{b-\mu}{\sigma/\sqrt{n}}\right) = 0.95 \tag{4}$$

for given values of μ and σ. Since $(\overline{X}_n - \mu)/(\sigma/\sqrt{n})$ has approximately a standard normal distribution, the above equality can be approximated by

$$P\left(\frac{a-\mu}{\sigma/\sqrt{n}} \leq Z \leq \frac{b-\mu}{\sigma/\sqrt{n}}\right) = 0.95 \tag{5}$$

where Z has a standard normal distribution. From probability tables,

$$P(-1.96 \leq Z \leq 1.96) = 0.95 \tag{6}$$

and hence

$$\frac{a-\mu}{\sigma/\sqrt{n}} = -1.96 \qquad \frac{b-\mu}{\sigma/\sqrt{n}} = 1.96 \tag{7}$$

or in a more useful form,

$$a = \mu - 1.96\sigma/\sqrt{n} \qquad b = \mu + 1.96\sigma/\sqrt{n} \tag{8}$$

In flood frequency analysis, the usual procedure is to identify the largest peak flow rate Q_i for each year i, respectively. Next, the logarithm of each Q_i is computed giving $q_i = \log Q_i$. Now, for n years (of zero) runoff data, we have n values of q_i. The mean of the set of values $\{q_i\}$ is the estimate of the log of the 2-year return frequency peak flow rate. Additionally, the q_i frequency histogram typically closely approximates a normal distribution; i.e., a log-normal distribution.

THE SAMPLING DISTRIBUTION OF S^2

Indeed, the beauty of the Central Limit Theorem lies in the fact that $\overline{X}_n$ will have approximately a normal sampling distribution no matter what the shape of the probabilistic model for the population, so long as n is large and σ^2 is finite. For many other statistics additional assumptions are needed before useful sampling distributions can be derived.

First note that if $X_1,....,X_n$ are independent normally distributed random variables with common mean μ and variance σ^2, then $\overline{X}_n$ will be *precisely* normally distributed with mean μ and variance σ^2/n. No approximating distribution is needed in this case since linear functions of independent normal random variables are again normal.

Under this normality assumption for the population, a sampling distribution can be derived for S^2, but we do not present the derivation here. It turns out that $(n-1)S^2/\sigma^2$ has a sampling distribution that is a special case of the gamma density function. If we let $(n-1)S^2/\sigma^2 = u$, then u will have the probability density function given by

$$f(u) = \begin{cases} \dfrac{1}{\Gamma\left(\frac{n-1}{2}\right) 2^{(n-1)/2}} u^{(n-1)/2 - 1} e^{-u/2} & u > 0 \\ 0 & \text{elsewhere} \end{cases} \qquad (9)$$

The gamma density function with $\alpha = v/2$ and $\beta = 2$ is called a *chi-square density function* with parameter v. The parameter v is commonly known as the *degrees of freedom.* Thus when the sampled population is normal , $(n - 1)S^2/\sigma^2$ has a chi-square distribution with $n - 1$ degrees of freedom.

Definition: An estimator $\hat{\theta}$ is **unbiased** for estimating θ if $E(\hat{\theta}) = \theta$

In studying sampling frequency histograms, we saw that the values of $\overline{X}_n$ tend to center at μ, the true population mean, when random samples are selected from the same population repeatedly. Similarly values of S^2 centered at σ^2, the true population variance. These are demonstrations of the fact that $\overline{X}_n$ is an unbiased estimator of μ and S^2 is an unbiased estimator of σ^2.

For an unbiased estimator $\hat{\theta}$ the sampling distribution of the estimator has mean value θ. How do we want the possible values of $\hat{\theta}$ to spread out to either side of θ for this unbiased estimator? Intuitively it would be desirable for all possible values of $\hat{\theta}$ to be very close to θ. That is, we want the variance of $\hat{\theta}$ to be as small as possible. It is possible to prove that some of our commonly used estimators do indeed have the smallest variance among all unbiased estimators. We will use this variance criterion for comparing estimators. That is, if $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimators of θ, then we would choose as the better estimator the one possessing the smaller variance.

GENERAL DISTRIBUTION: LARGE-SAMPLE CONFIDENCE INTERVAL FOR μ

Suppose we are interested in estimating a mean μ for a population with variance σ^2, assumed, for the moment, to be known. We select a random sample $X_1,...X_n$ from this population and compute $\overline{X}_n$ as a point estimator of μ. If n is large (say, $n \geq 30$ as a rule of thumb), then $\overline{X}_n$ has approximately a normal distribution with mean μ and variance σ^2/n. From these facts we can state that the interval $(\mu - 2\sigma/\sqrt{n}, \mu + 2\sigma/\sqrt{n})$ contains about 95% of the $\overline{X}_n$ values that could be generated in repeated random samplings from the population under study. For convenience let's call this middle 95% the "likely" values of $\overline{X}_n$. Now suppose we are to observe a single sample producing a single $\overline{X}_n$. A question of interest is "What possible values for μ would allow this $\overline{X}_n$ to lie in the likely range of possible sample means?" This set of possible values for μ is the confidence interval with confidence coefficient of approximately 0.95.

The main idea of confidence interval construction is shown on Figure 6.

More formally, under these conditions

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution, approximately. Now for any prescribed α we can find from probability tables a value $z_{\alpha/2}$ such that

$$P\,[- z_{\alpha/2} \leq Z \leq + z_{\alpha/2}] = 1 - \alpha \tag{10}$$

Rewriting this probability statement, we have

$$
\begin{aligned}
1 - \alpha &= P\left[\ -z_{\alpha/2} \leq \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq + z_{\alpha/2}\right] \\
&= P\left[\ -z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{X}_n - \mu \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \\
&= P\left[\ \overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \qquad (11)
\end{aligned}
$$

The interval

$$
(\ \overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})
$$

forms a realization of a large-sample confidence interval for μ with confidence coefficient approximately $(1 - \alpha)$.

It is of interest to consider the peak flow rate data we are given for a flood frequency analysis. Unlike a simple random sample, our peak flow data are clustered together according to the time period we have gauged the watercourse. Consequently, the sample typically may not demonstrate the breadth of variability that the population truly has if there are background cycles present in the weather patterns.

MODELS

In rainfall-runoff hydrology, we are aware of the hundreds of so-called deterministic models, and of probabilistic or statistical models. We will consider some practical concepts of both.

Figure 7 shows a possible set of responses for the same values of x when we are using a probabilistic model. Note that the deterministic part of the model (the straight line itself) is the same. Now, however, the inclusion of a random error component allows the peak loads to vary from this line. Since we believe that will vary randomly for a given value of x, the probabilistic model provides a more realistic model of Y than does the deterministic model.

General Form of Probabilistic Models

Y = deterministic component + random error

where Y is the random variable to be predicted. We will always assume that the mean value of the random error equals zero. This is equivalent to assuming that the mean value of Y, $E(Y)$, equals the deterministic component of the model:

$E(Y)$ = deterministic component

Figure 8 considers the most primitive model, Y = constant, along with the data available.

The Straight-Line Probabilistic Model

$$Y = \beta_o + \beta_1 x + \varepsilon$$

where Y = dependent variable (variable to be modeled)

x = independent variable (variable used as a predictor of Y)

ε = random error component (see Figure 10)

β_o = y intercept of the line, that is, point at which the line intercepts or cuts through the y axis (see Figures 9, 11)

β_1 = slope of the line, that is, amount of increase (or decrease) in the mean Y for every 1 unit increase in x (see Figure 9)

Note that we use the Greek symbols β_o and β_1 to represent the Y intercept and slope of the model, as we used the Greek symbol μ to represent the constant mean in the Model $Y = \mu + \varepsilon$. In each case these symbols represent population parameters with numerical values that will need to be estimated using sample data. Consequently, β_o and β_1 are themselves statistics, which have a distribution of outcomes that depend upon the underlying population.

In order to make progress, some assumption is inserted into the analysis regarding the character of the error between the data and the deterministic model estimates. Usually, we will assume a normal distribution.

Probability Distribution of the Random Error Component ε

The error component is normally distributed with mean zero and constant variance σ^2. The errors associated with different observations are independent.

Sampling of Distribution of $\hat{\beta}_1$

If we assume that the error components are independent normal random variables with mean zero and constant variance σ^2, the sampling distribution of the least-squares estimate $\hat{\beta}_1$ of the slope will be normal, with mean β_1 (the true slope) and standard deviation

$$\sigma_{\beta 1} = \frac{\sigma}{\sqrt{SS_{xx}}}$$

MODEL ERRORS

How good are our models? How well do the models estimate the T-year peak flowrate? How well do our models fit the stream gauge data? The following Figures 13 and 14 demonstrate these important concepts. Remember: highly complex rainfall-runoff models have these difficulties; are you aware of them?

Sampling Errors for the Estimator of the Mean of Y and the Predictor of an Individual Y

1. The standard deviation of the sampling distribution of the estimator $\hat{Y}$ of the mean value of Y at a fixed x is

$$\sigma\hat{Y} = \sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SS_{xx}}}$$

where σ is the standard deviation of the random error ε.

2. The standard deviation of the prediction error for the predictor $\hat{Y}$ of an individual Y value at a fixed x is

$$\sigma_{(Y-\hat{Y})} = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_{xx}}}$$

where σ is the standard deviation of the random error ε.

The true value of σ will rarely be known. Thus we estimate σ by s and calculate the estimation and prediction intervals as shown next.

A 100(1-α) Percent Confidence Interval for the Mean Value of Y at a Fixed *x*

$$\hat{y} \pm t_{\alpha/2}(n - 2) \text{ (estimated standard deviation of } \hat{Y})$$

or

$$\hat{y} \pm t_{\alpha/2}(n - 2)s\sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{SS_{xx}}}$$

REFERENCES

1. Schaeffer, Richard L., and McClave, James T., "Probability and Statistics for Engineers", 3rd Edition, PWS-Kent Publishing Company, Boston.

2. Phillips, John L., Jr., "How to Think About Statistics", Revised Edition, W.H. Freeman and Company, New York.

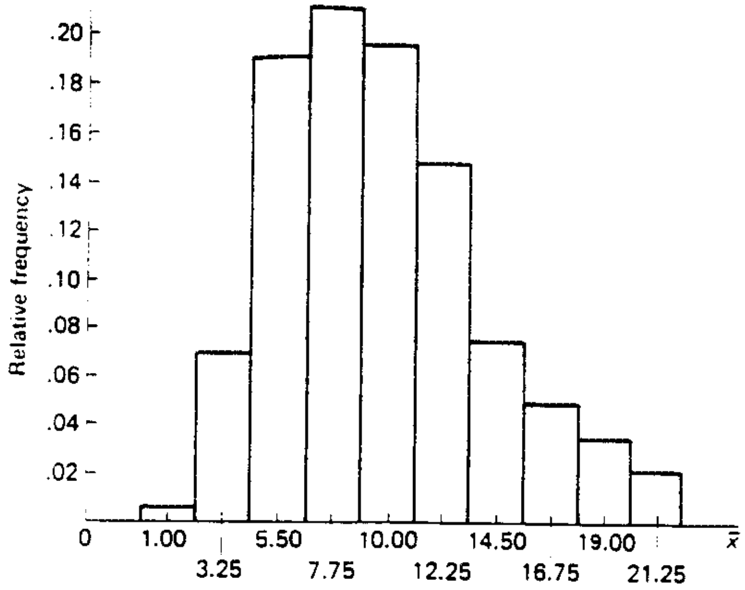


Figure 1. Relative Frequency Histogram for $\bar{x}$ from 1000 Samples of Size n = 5.

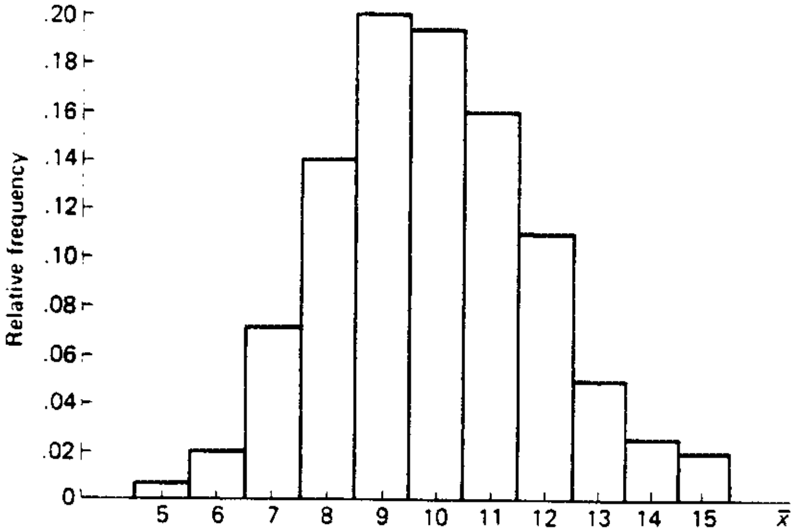


Figure 2. Relative Frequency Histogram for $\bar{x}$ from 1000 Samples of Size n = 25.

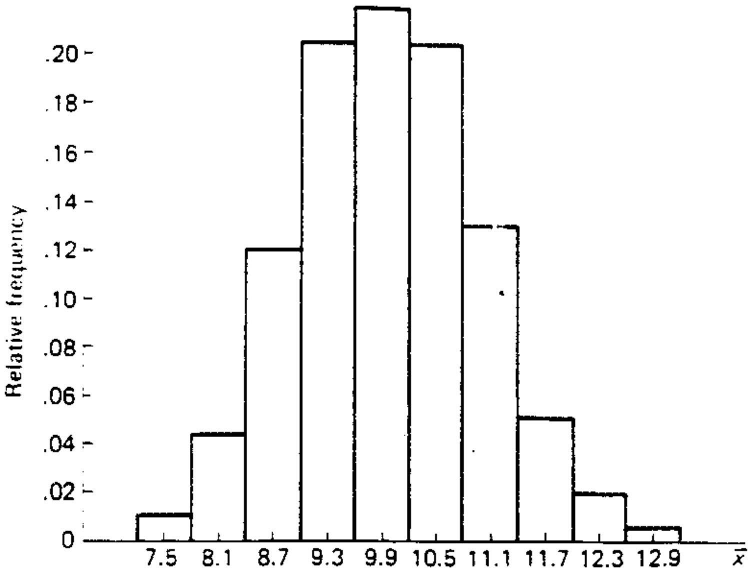


Figure 3. Relative Frequency Histogram for $\overline{x}$ from 1000 Sampes of Size n = 100.

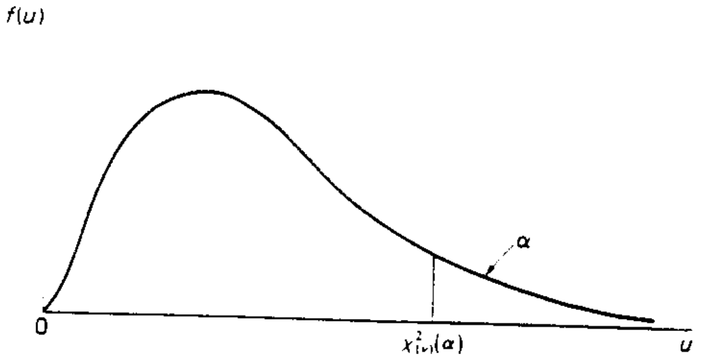


Figure 4. A Chi-Square Distribution.

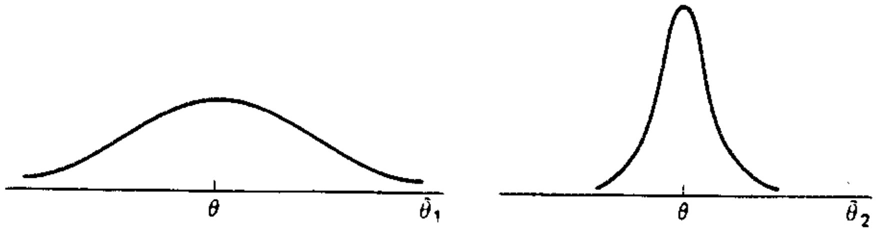


Figure 5. Distributions of Two Unbiased Estimators, $\widehat{\theta}_1$ and $\widehat{\theta}_2$, with $V(\widehat{\theta}_2) < V(\widehat{\theta}_1)$.

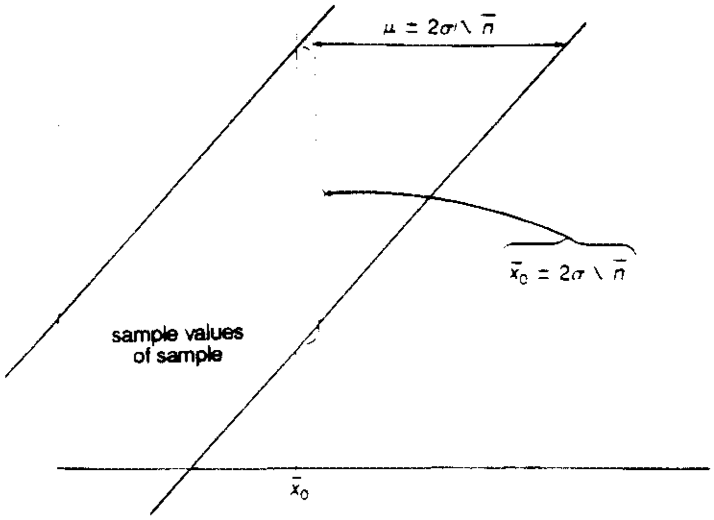


Figure 6. The Construction of a Confidence Interval.

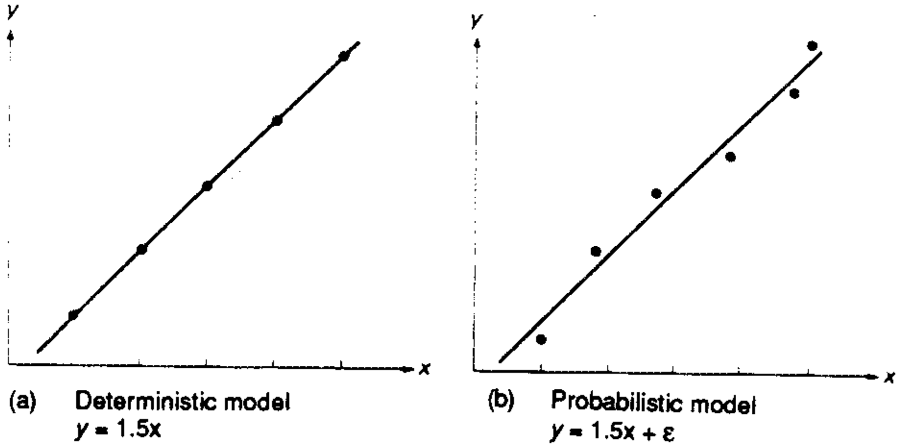


Figure 7. Deterministic and Probabilistic Models.

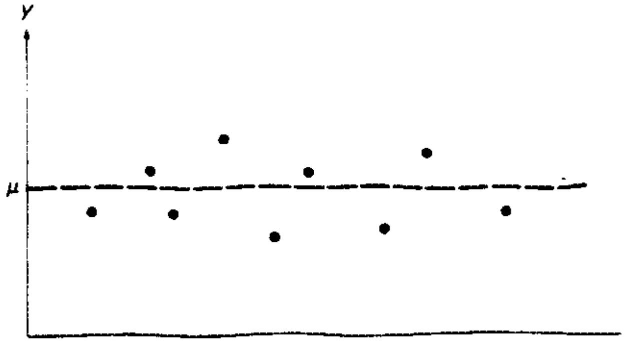


Figure 8. The Probabilistic Model $Y = \mu + \varepsilon$.

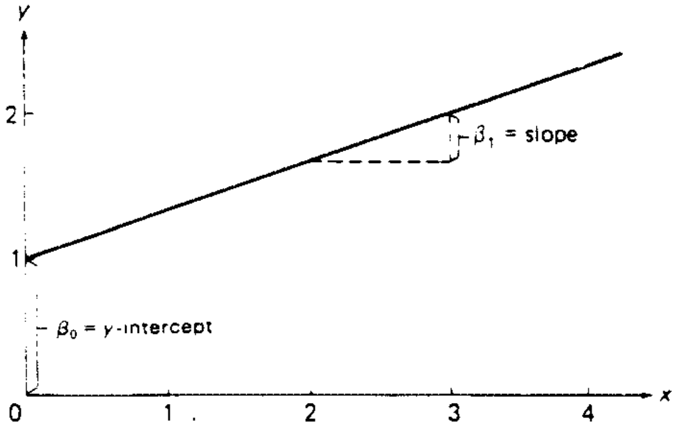


Figure 9. Straight-Line Probabilistic Model.

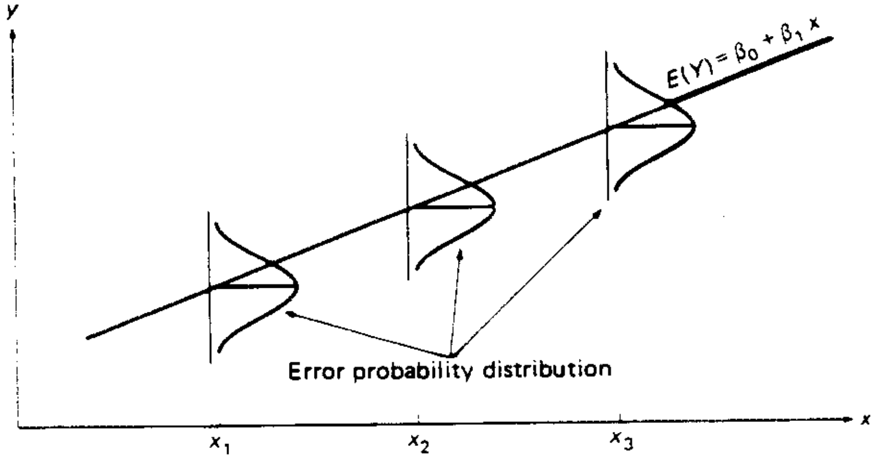


Figure 10. The probability Distribution of ε.

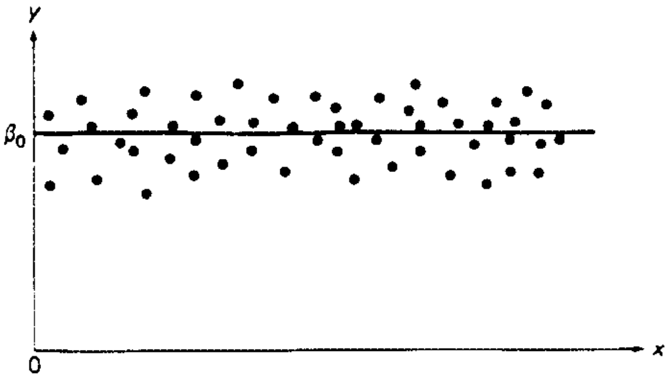


Figure 11. Graph of the Model $Y = \beta_0 + \varepsilon$.

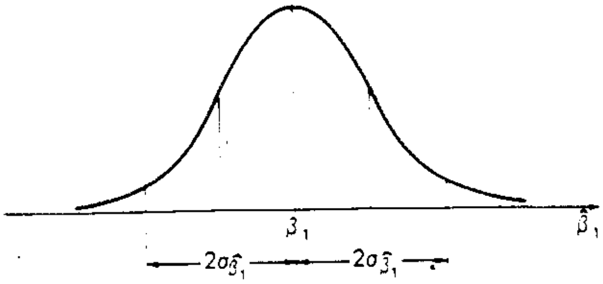
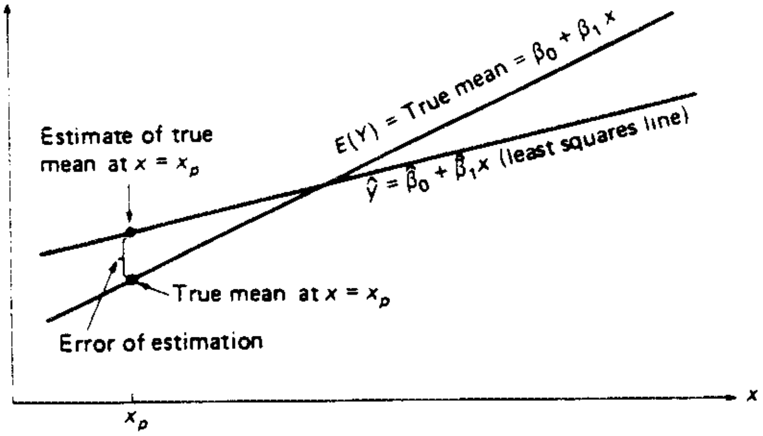
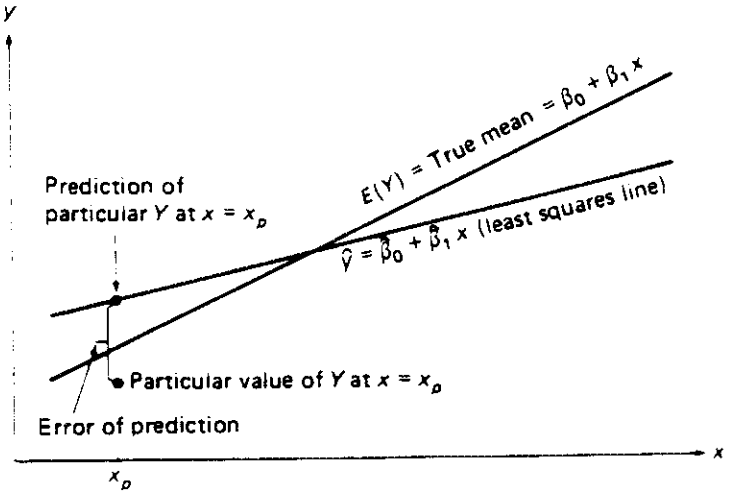


Figure 12. Sampling Distribution of $\hat{\beta}_1$



$E(Y)$ = True mean = $\beta_0 + \beta_1 x$
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (least squares line)
Estimate of true mean at $x = x_p$
True mean at $x = x_p$
Error of estimation
x
x_p

Figure 13. Error of Estimating the Mean Value of Y for a given Value of x.



y
$E(Y)$ = True mean = $\beta_0 + \beta_1 x$
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (least squares line)
Prediction of particular Y at $x = x_p$
Particular value of Y at $x = x_p$
Error of prediction
x
x_p

Figure 14. Error of Predicting a Future Value of Y for a Given Value of x.

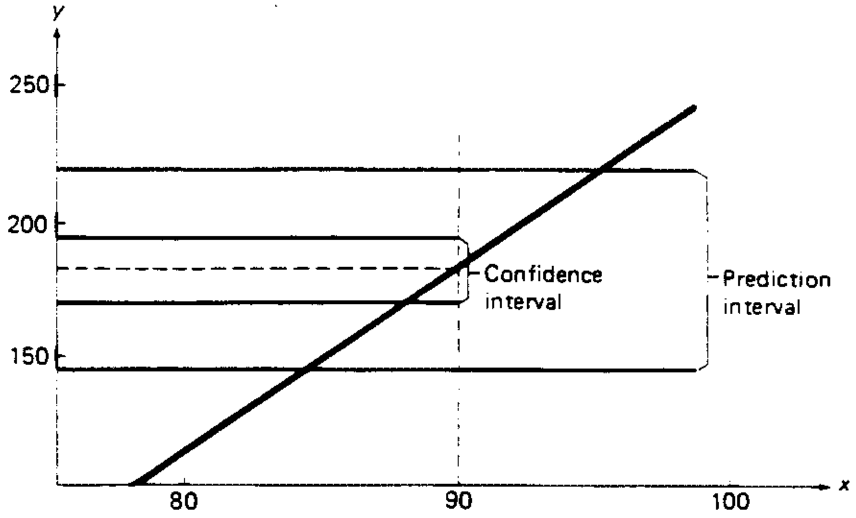


Figure 15. A 95% Confidence Interval for Mean and a Prediction Interval for Peak when $x = 90$.

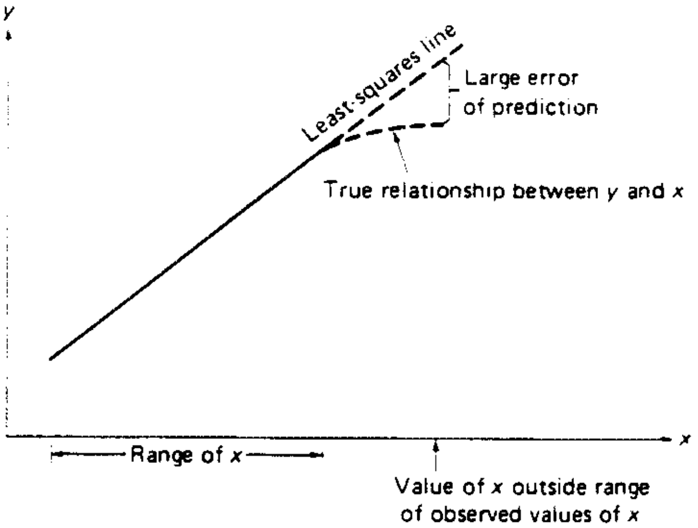

y
Least-squares line
Large error of prediction
True relationship between y and x
Range of x
x
Value of x outside range of observed values of x

Figure 16. Using a Model to Predict Outside the Range of Sample Values of x.

Example 2

To calculate the standard error of a mean, you need the *size* of the sample and its *standard deviation*:

$$N = 10{,}000$$
$$S = 10.2$$
$$\overline{X} = 115.5$$
$$SE_{\overline{X}} = 0.102$$

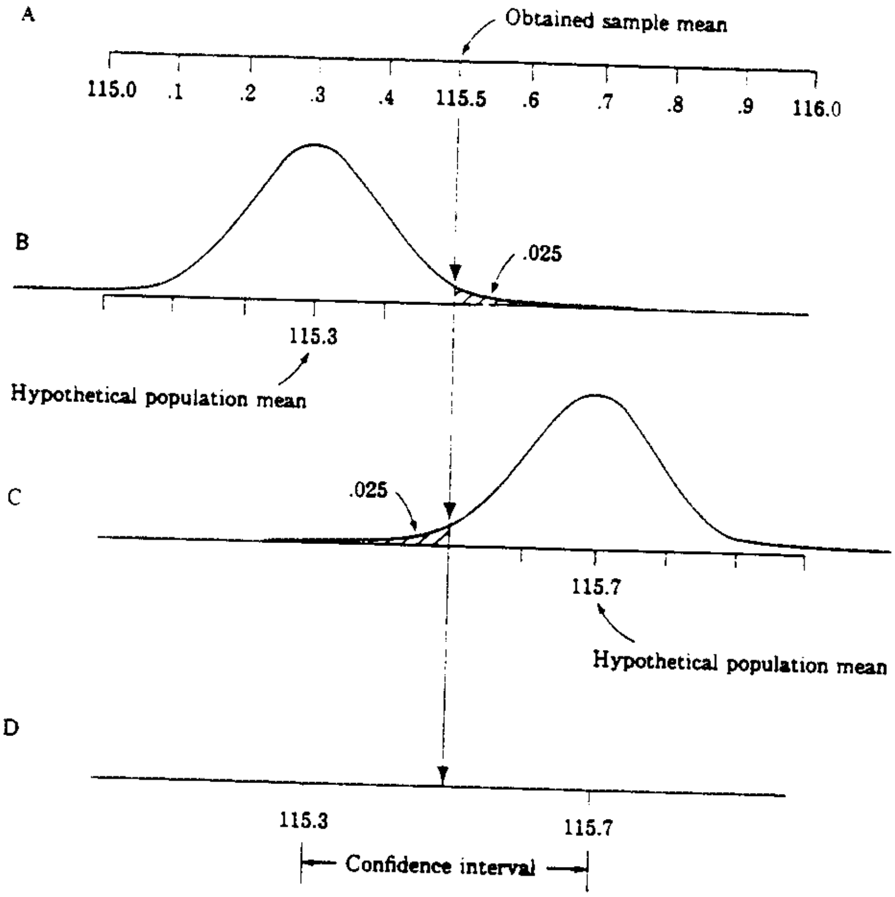


Figure 17. Establishing a Confidence Interval at .95 Level of Confidence.

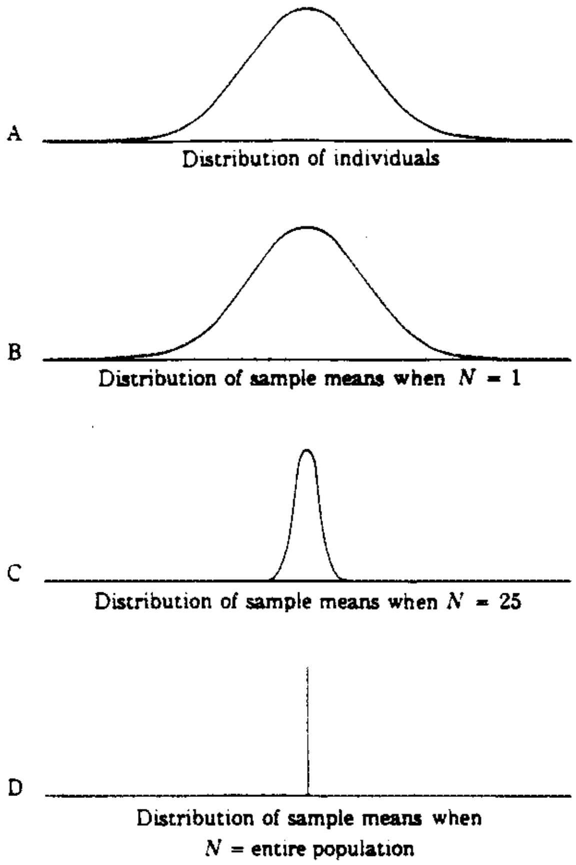

A
Distribution of individuals
B
Distribution of sample means when $N = 1$
C
Distribution of sample means when $N = 25$
D
Distribution of sample means when
N = entire population

Figure 18. Distributions of Individuals and of the Means of Samples of Three Different Sizes.

Example 3

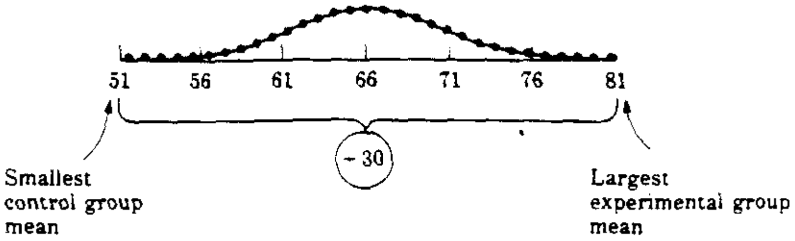


Figure 19. Sample Means, $SE_{\overline{X}} = 5$, when control (——) and experiment (···) Populations are Identical.

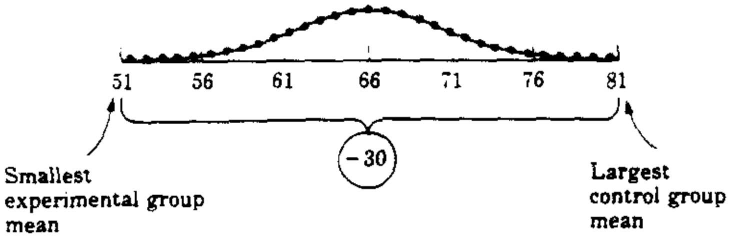


Figure 20. Same as Figure 19 except that different extreme means are identified (control,——; experimental, ···).

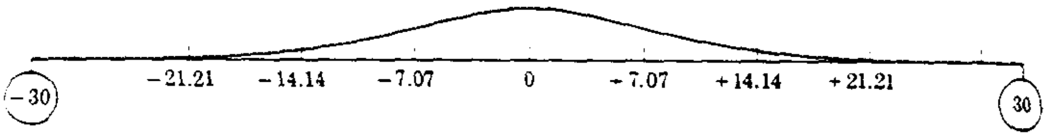


Figure 21. Differences between means, $SE_{\overline{X}_e-\overline{X}_c} = 7.07$, on a scale of raw scores.

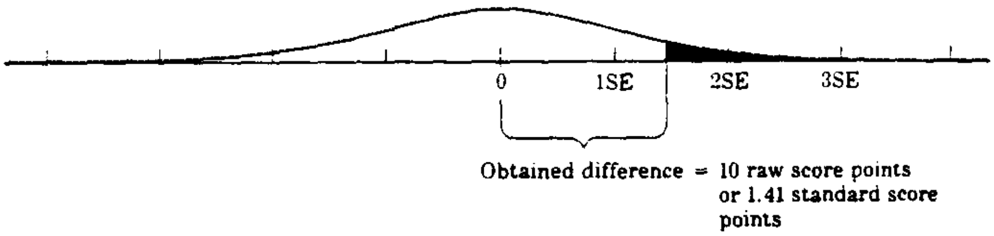


Figure 22. Same as Figure 21 but on a scale of standard error units ($SE_{\overline{X}_e-\overline{X}_c}$).