#277

Floodplain Management Association

1996 Spring Conference and Workshops

A Refresher of Practical Statistics for Flood Control Hydrology

Theodore V. Hromadka II, Ph.D., Ph.D., PE, PH

Boyle Engineering Corporation

1501 Quail Street
Newport Beach, CA 92658-9020

(714) 476-3383

THEOREM *The Central Limit Theorem.* If a random sample of size n is drawn from a population with mean μ and variance σ^2, then the sample mean $\bar{X}$ has approximately a normal distribution with mean μ and variance σ^2/n. That is, the distribution function of

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

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

:. Samples of size n were drawn from a population having the probability density function

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

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 6.12. Figures 6.13 and 6.14 show similar results for 1000 samples of size $n = 25$ and $n = 100$, respectively. Although all the relative frequency histograms have a sort of bell shape, notice that the tendency toward a symmetric normal curve is better for larger n. A smooth curve drawn through the bar graph of Figure 6.14 would be nearly identical to a normal density function with mean 10 and variance $(10)^2/100 = 1$.

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

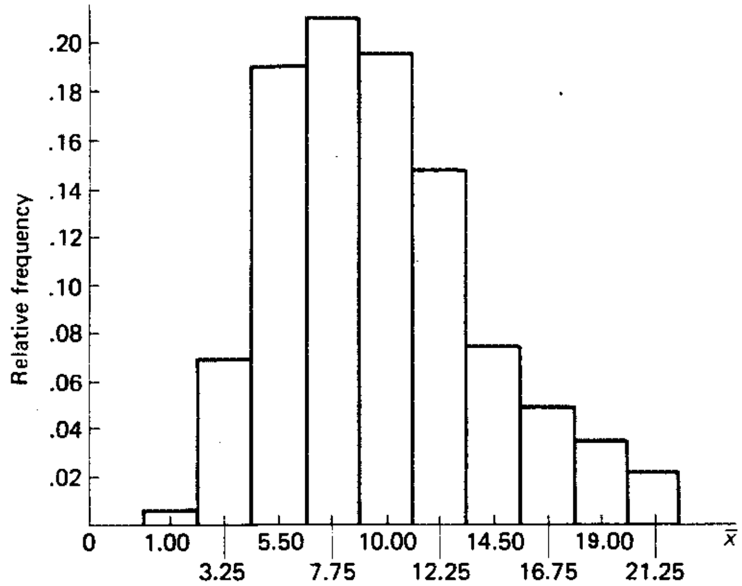
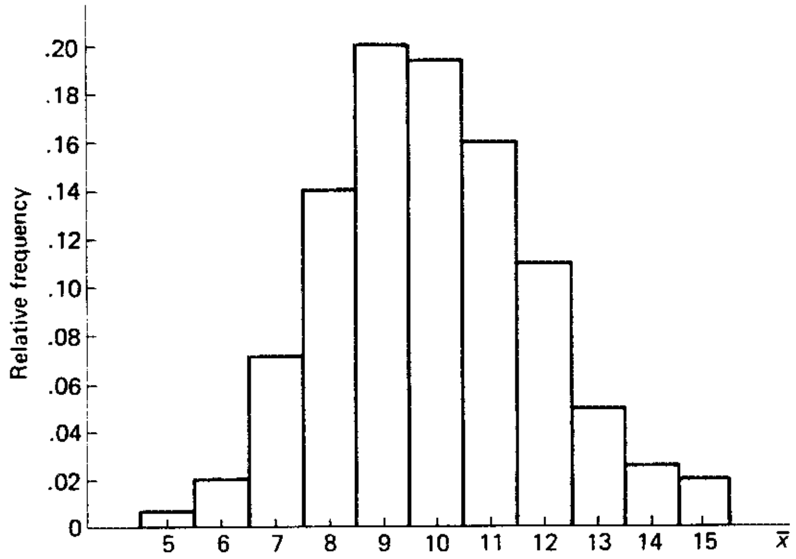


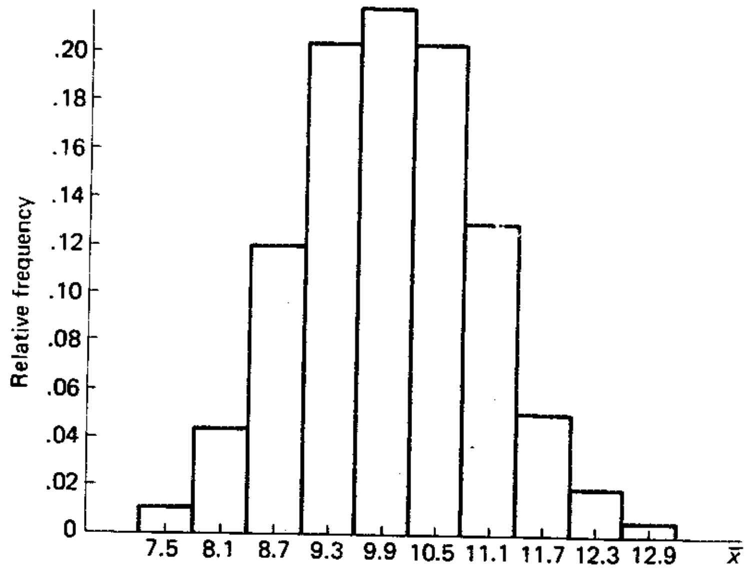
FIGURE 2
Relative Frequency Histogram for $\bar{x}$ from 1000 Samples of Size $n = 25$



The Central Limit Theorem provides a very useful result for statistical inference, for we now know not only that $\bar{X}$ 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 for $\bar{X}$ is approximately normal. For example, suppose we wish to find an interval $(a; b)$ such that

$$P(a \leqslant \bar{X} \leqslant b) = 0.95$$

FIGURE 3
Relative Frequency Histogram for $\bar{x}$ from 1000 Samples of Size $n = 100$



This probability is equivalent to

$$P\left(\frac{a-\mu}{\sigma/\sqrt{n}} \leqslant \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leqslant \frac{b-\mu}{\sigma/\sqrt{n}}\right) = 0.95$$

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

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

where Z has a standard normal distribution. From Table 4 in the Appendix we know that

$$P(-1.96 \leqslant Z \leqslant 1.96) = 0.95$$

and hence

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

or

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

THE SAMPLING DISTRIBUTION OF S^2

The beauty of the Central Limit Theorem lies in the fact that $\bar{X}$ 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. A common assumption is that the probabilistic model for the population is itself normal. That is, we assume that if the population of measurements of interest could be viewed in histogram fashion, that histogram would have roughly the shape of a normal curve. This, incidentally, is not a bad assumption for many sets of measurements one is likely to come across in real-world experimentation.

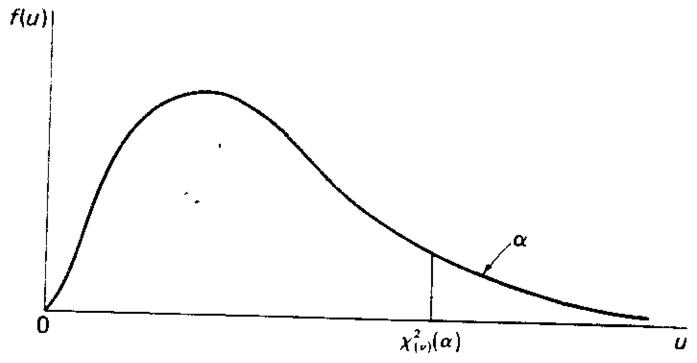
First note that if $X_1, \ldots, X_n$ are independent normally distributed random variables with common mean μ and variance σ^2, then $\bar{X}$ will be *precisely* normally distributed with mean μ and variance σ^2/n. No approximating distribution is needed in this case since linear functions of independent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

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

The gamma density function with $\alpha = \nu/2$ and $\beta = 2$ is called a *chi-square density function* with parameter ν. The parameter ν 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.

FIGURE 4
A χ^2 Distribution (Probability Density Function)



DEFINITION

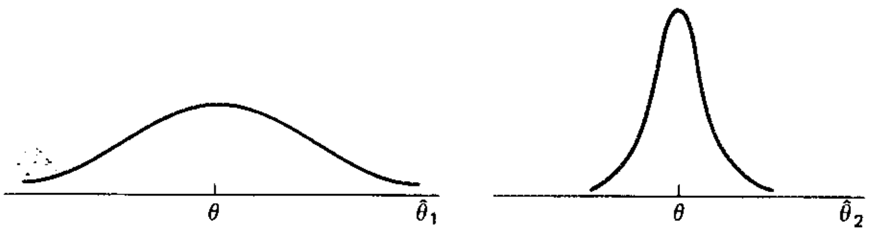
An estimator $\hat{\theta}$ is **unbiased** for estimating θ if

$$E(\hat{\theta}) = \theta$$

In the sampling distributions we saw that the values of $\bar{X}$ 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 $\bar{X}$ 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 for $\hat{\theta}$ 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. At the level of this text it is not possible to prove that some of our estimators do indeed have the smallest variance among all unbiased estimators, but 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.

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

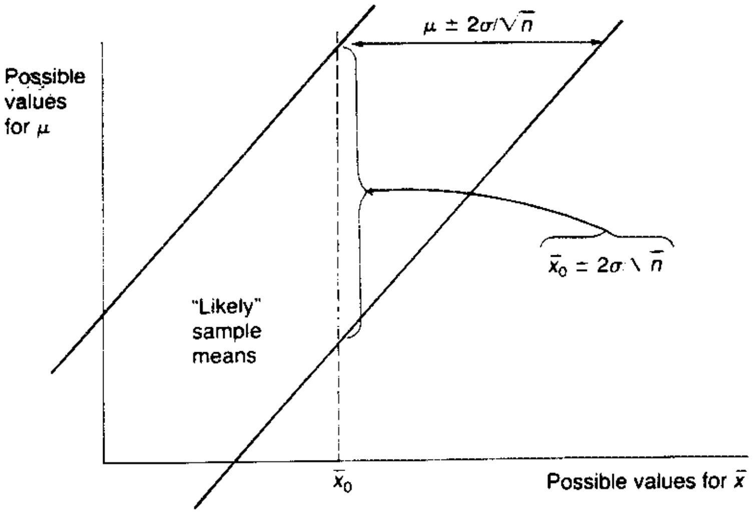


General Distribution: Large-Sample Confidence Interval for μ

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 $X_1, \ldots, X_n$ from this population and compute $\bar{X}$ as a point estimator of μ. If n is large (say, $\boxed{n \geqslant 30}$ as a rule of thumb), then $\bar{X}$ has approximately a normal distribution with mean μ and variance σ^2/n. From these facts we can state that the interval $\mu \pm 2\sigma/\sqrt{n}$ contains about 95% of the $\bar{x}$ 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 $\bar{x}$. Now suppose we are to observe a single sample producing a single $\bar{x}$. A question of interest is "What possible values for μ would allow this $\bar{x}$ 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 in the diagram.

FIGURE 6
The Construction of a Confidence Interval



More formally, under these conditions

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

has a standard normal distribution, approximately. Now for any prescribed α we can find from Table 4 in the Appendix a value $z_{\alpha/2}$ such that

$$P[-z_{\alpha/2} \leqslant Z \leqslant +z_{\alpha/2}] = 1 - \alpha$$

Rewriting this probability statement, we have

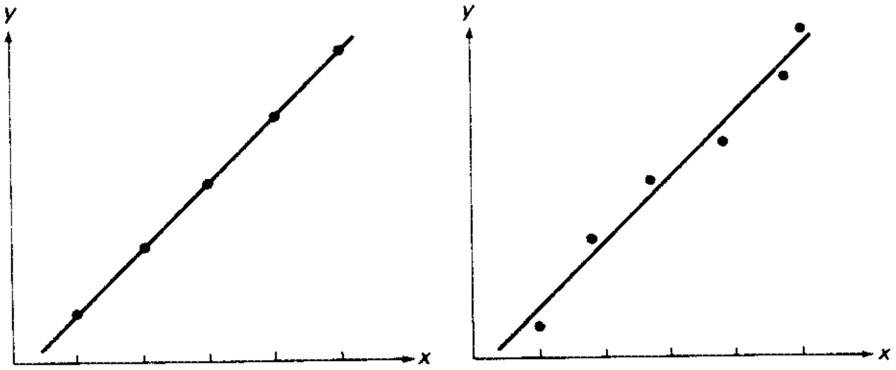
$$\begin{aligned} 1 - \alpha &= P\left[-z_{\alpha/2} \leqslant \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leqslant +z_{\alpha/2}\right] \\ &= P\left[-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \bar{X} - \mu \leqslant +z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \\ &= P\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \end{aligned}$$

The interval

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

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

FIGURE 7
Deterministic and Probabilistic Models



(a) Deterministic model
$y = 5x$

(b) Probabilistic model
$y = 5x$ + random error

Figure 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 the will vary randomly for a given value of x, the probabilistic model provides a more realistic model for Y than does the deterministic model.

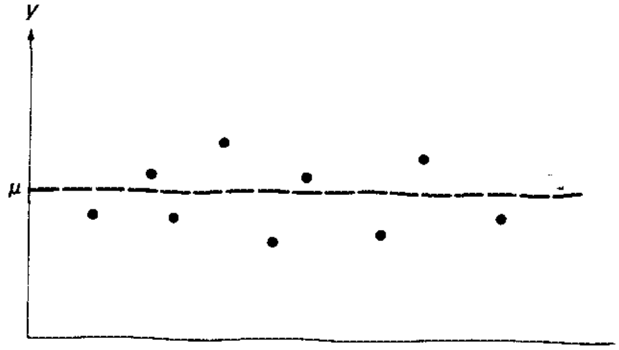
GENERAL FORM OF PROBABILISTIC MODELS

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:

$$E(Y) = \text{deterministic component}$$

FIGURE 8
The Probabilistic Model $Y = \mu + \varepsilon$

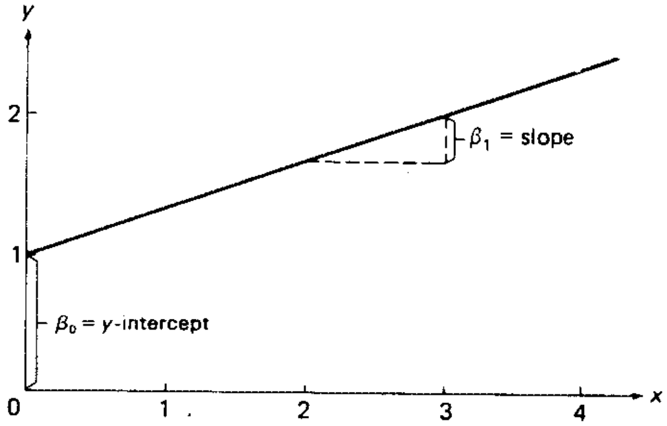


THE STRAIGHT-LINE PROBABILISTIC MODEL

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

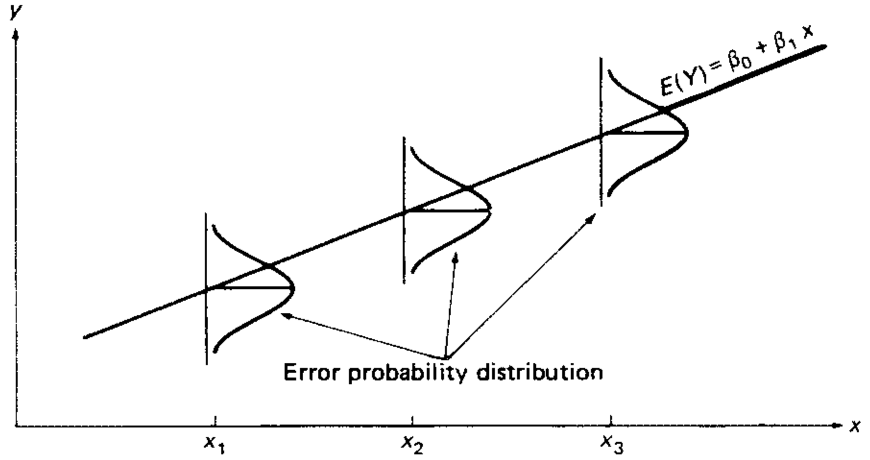
where Y = dependent variable (variable to be modeled)
x = independent variable (variable used as a predictor of Y)
ε = random error component
β_0 = y intercept of the line, that is, point at which the line intercepts or cuts through the y axis (see Figure 9.3)
β_1 = slope of the line, that is, amount of increase (or decrease) in the mean of Y for every 1 unit increase in x (see Figure 9.3)

FIGURE 9
Straight-Line Probabilistic Model



Note that we use the Greek symbols β_0 amd 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.

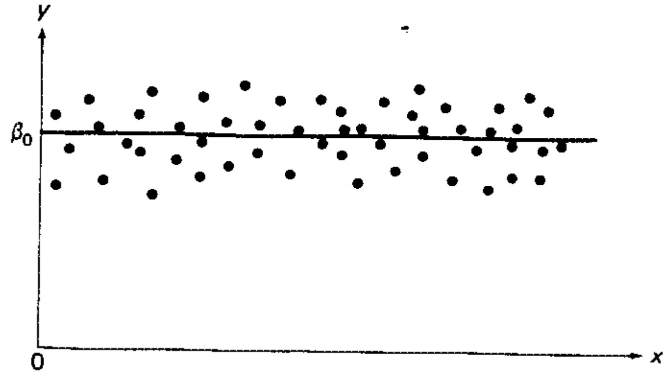
FIGURE 10 *The Probability Distribution of ε*



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.

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



SAMPLING 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 estimator $\hat{\beta}_1$ of the slope will be normal, with mean β_1 (the true slope) and standard deviation

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{\mathrm{SS}_{xx}}} \qquad \text{(see Figure 9.8)}$$

[*Note*: Proof of the unbiasedness of $\hat{\beta}_1$ and a derivation of its standard deviation are given next.]

FIGURE 12
Sampling Distribution of $\hat{\beta}_1$

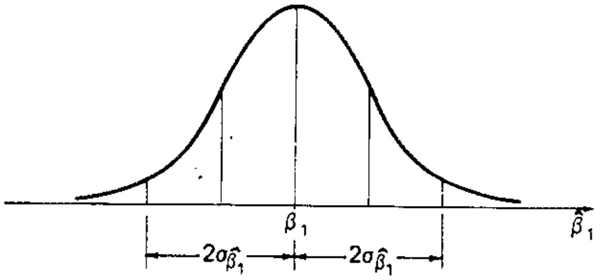


FIGURE 13
Error of Estimating the Mean Value of Y for a Given Value of x

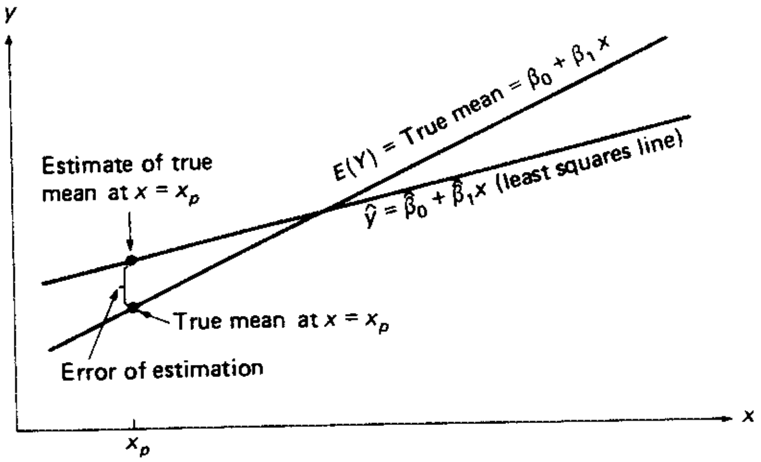
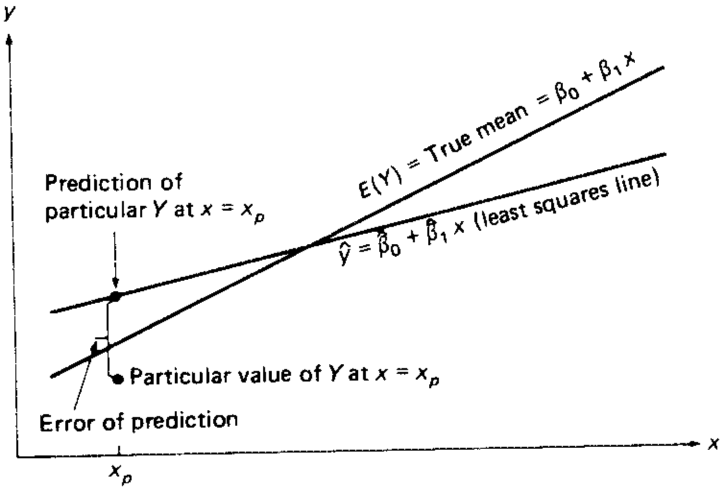


FIGURE 14
Error of Predicting a Future Value of Y for a Given Value of x



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

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

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

FIGURE 15
A 95% Confidence Interval for Mean Peak Power Load and a Prediction Interval for Peak Power Load When $x = 90$

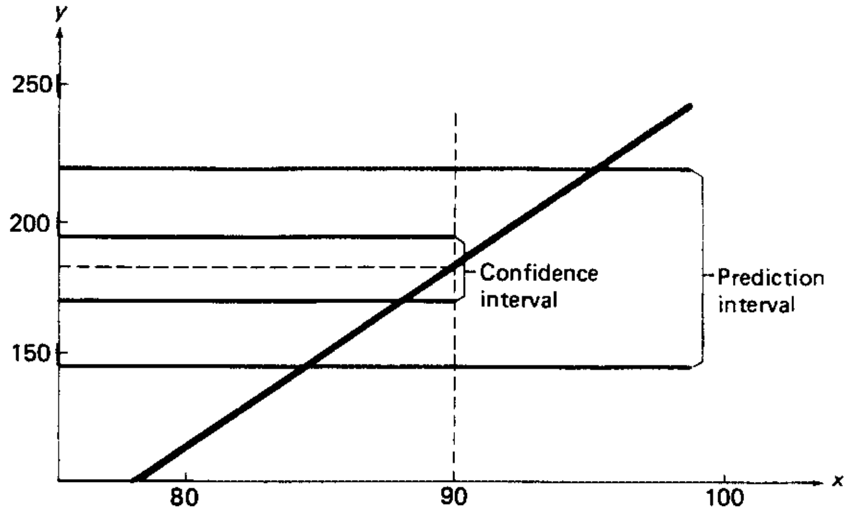
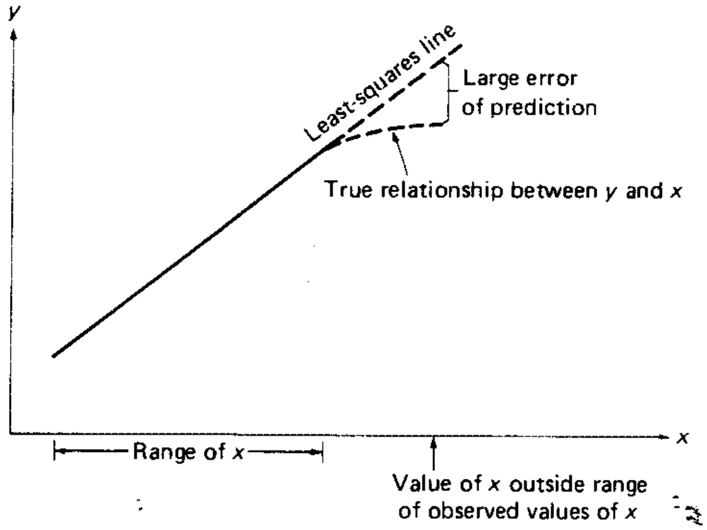


FIGURE 16
The Danger of Using a Model to Predict Outside the Range of the Sample Values of x



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

$$N = 10{,}000$$

$$S = 10.2$$

Of course, you can't calculate the standard deviation unless you know the mean, so it, too, is essential:

$$\overline{X} = 115.5$$

$$SE_{\overline{X}} = 0.102$$

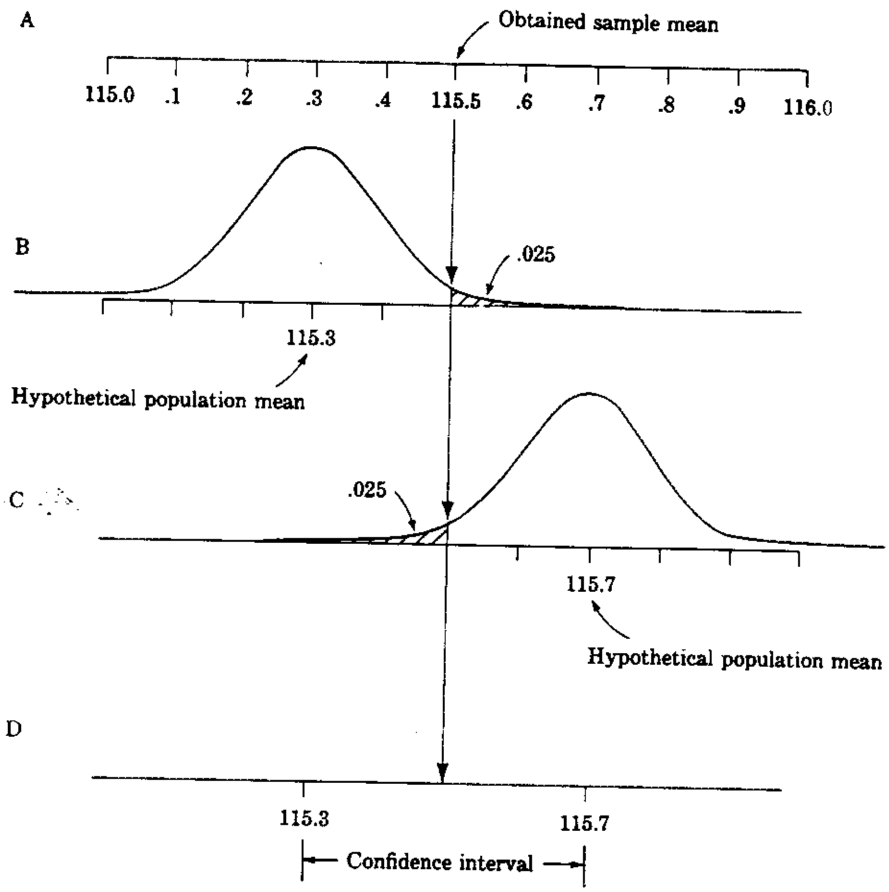


Figure 17 Establishing a confidence interval at .95 level of confidence.

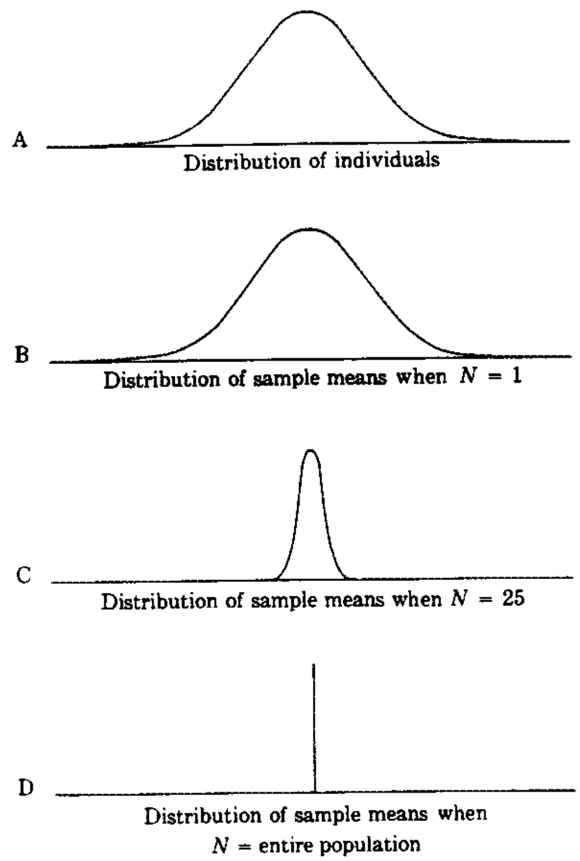


Figure 18 Distributions of individuals and of the means of samples of three different sizes.

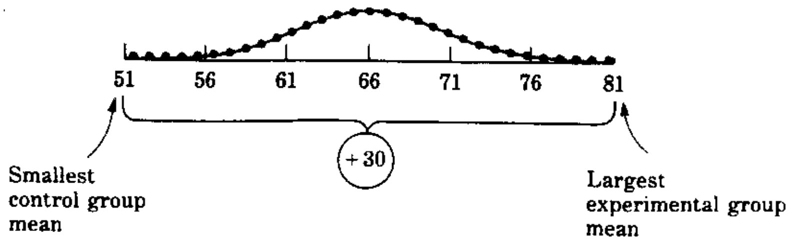


Figure 19 Sample means, $SE_{\bar{X}} = 5$, when control (———) and experimental (· · ·) populations are identical.

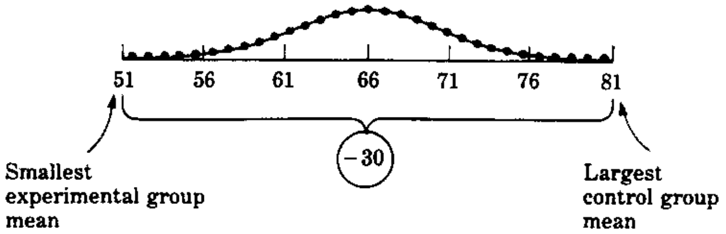


Figure 20 Same as Figure 7-5 except that different extreme means are identified (control, ———; experimental, · · ·).

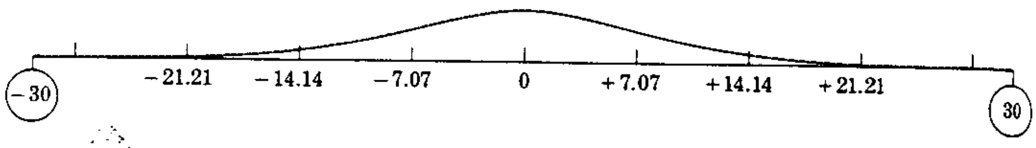


Figure 21 Differences between means, $SE_{\bar{X}_e - \bar{X}_c} = 7.07$, on a scale of raw scores.

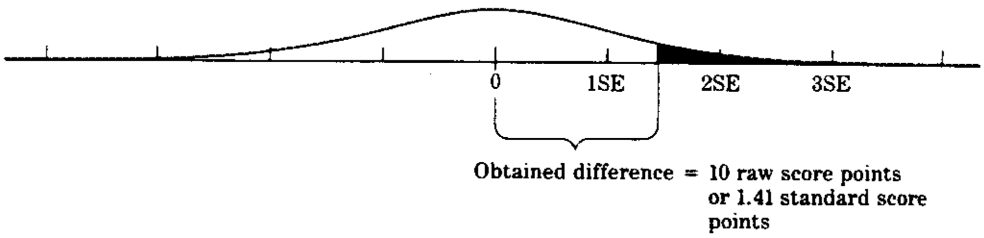


Figure 22 Same as Figure 7-7 but on a scale of standard error units ($SE_{\bar{X}_e - \bar{X}_c}$).

THIRD EDITION

Probability and Statistics for Engineers

Richard L. Scheaffer
University of Florida

James T. McClave
Infotech, Inc.

PWS-KENT PIBLISHING COMPANY, BOSTON

How to Think about Statistics

Revised Edition

John L. Phillips, Jr.
Boise State University

W. H. Freeman and Company
New York

Floodplain Management Association

1996 Spring Conference and Workshops

PRACTICAL STATISTICS FOR FLOODPLAIN MANAGMENT DECISIONS

Alan J. Nestlinger, PE, MS

Senior Civil Engineer

Orange County Environmental Management Agency

County of Orange/EMA

300 North Flower Street
Santa Ana, CA 92702-4048

(714) 834-2425

Practical Statistics for Floodplain Management Decisions

I. The experience of a large flood control district with regional calibration of its hydrology manual.

A. Management needs and requests for certainty.

B. What variance tells you about the strength of your calibration.

C. Management use of confidence intervals.

1. Design of structures to be maintained by the agency.
2. Non-structural evaluation
3. Existing conditions floodplains
4. Environmental mitigation

II. Why calibrate?

A. Prescriptive hydrology may be strongly biased.

B. Strengthens agency position in enforcing standards.

C. Reduces agency vulnerability to litigation (opponents typically cannot produce a superior calibration to that of an agency).

III. Can you calibrate your hydrology (HEC-1, SCS TR-20, etc.) to rain and streamgage data?

A. "At site" calibration?

B. Regional calibration?

C. Quality and quantity of gage data.

FMA Statistics Workshop

I.A. A flood control district's managers (all professional engineers) wanted to base the district's new hydrology manual on streamflow data rather than the previous manual's reliance on rain data combined with prescriptive models. The rain data and prescriptive model method essentially "asserted" the frequency of the calculated runoff with no way of knowing whether or not the assertion was correct to any degree at all. Managers thought that basing the flood frequency of model results in the new manual on stream flow data would provide something close to certainty in the computed results. Did they get what they wanted? Let's look at what was possible and what was actually done to meet their perceived need for certainty.

The availability of a large data set composed of stream gage and rain gage data collected by the U.S. Army Corps of Engineers for selected watersheds in and around Los Angeles made the calibration effort possible. The "LACDA" data set has been described as the best regional calibration data set available in the Southwestern United States.

I.B. In spite of the large LACDA data set the results of frequency analysis displayed a disturbing degree of variance. Variance is a measure of how much the data "spreads" from it's central values (mean, median, mode). A large amount of variance is very typical of stream flow data. We are all familiar with the "bell curve" or Gaussian "normal" distribution (see figures 1 and 2).

We might remember here that the rain and stream gage data used to develop the famous SCS Curve Numbers (CN) had about as many large runoff events on dry watersheds as they had small runoff events on wet watersheds. This counter intuitive result is unfortunately common in data derived from field data.

Consider the sharp peak curve (small variance) that plots frequency on the y-axis and data on the x-axis (fig. 1). This curve might be typical of measurements made on precision ball bearings. The diameters of the bearings might vary no more than 0.00005" for 95% of those taken in any sample. The y-axis tells you how frequently you would find bearings of particular dimensions.

Now look at broad peak curve (large variance-fig. 2). We are using the bell curve to illustrate a point that is not lost by avoiding the more complicated Log-Normal (LN) or log Pearson III (LP3) curve that is widely used to represent actual stream flow data.

Figure 2 plots frequency (y-axis) versus the range of possible 100-year Q's (say in cfs/sq.mi.). Note that Q's that vary ±20% have very similar frequencies. That means that the data is telling you that the best calibration data gives a 40% range in magnitude for Q's that are very close to the same likelihood of occurrence. E.g., a discharge of 800 cfs or 1200 cfs is nearly as likely to be the Q-100 as the central value of 1000 cfs in this example.

For a floodplain manager this is not good news. There is a large amount of uncertainty in the calibration results in spite of a large amount of high quality gage data. The message to be received and understood here is that the "true" Q100 is not only unknown it is unknowable. The statistics are not wrong but they are weak and cannot be made better than they are (data is data).

In the distant future, with a thousand years of additional data, we will still find that there remains a substantial amount of variance in flood frequency analysis. The rainfall-runoff process is inherently complex and will not become simpler because we measure it a lot. We are unlikely to find comprehensive solutions to the problems of sparse data collection from large watersheds, spatial and temporal incoherence of data, the inhomogeneity of the record (watershed development causing large changes in watershed response to rainfall), the heterogeneity (non-uniformity of the soil data, rain data, land use, etc.) and, considering the thousand year period suggested above, large scale climatic change that essentially invalidates all prior data.

If we use the mean value of the distribution (50% confidence interval) on a regional basis half of our designs will be undersized (by potentially dangerous margins) and the other half oversized but we wont know which is which!

Well! What to do with all this uncertainty that cannot be removed? Statisticians deal with this problem all the time so let's look at their methods. If you want to capture the "true" Q100 but cannot confidently pin it down you can increase your chances of using a discharge that is at least as large as the unknowable Q100 by going to higher confidence intervals. See figure 2. The district chose to go one standard deviation higher than the mean (85% confidence interval, often called high confidence design - HC) for design discharges used for channel improvements, levees, dams and other hydraulic structures for which the district would be liable for the proper performance and facility maintenance. That's easy. Bigger is always better, right?

I.C. Wrong! What about evaluating a floodplain under existing conditions for which no conveyance facilities are to be built for insurance purposes? The mean value (50% confidence interval) is also known as the "expected value" (EV). It is your best guess, the one equally likely to be high or low, sometimes called the maximum likelihood estimator. This is the value that would help to balance insurance premiums with flood damage losses.

Whenever you want to evaluate existing conditions, say to determine the level of protection provided by an old city storm drain system, EV is the best choice.

Consider environmental mitigation. Let's say a new residential subdivision is going in near the headwaters of a watershed and there is a large open space preserve downstream of the subdivision. In order to minimize the development impact on the natural stream course a retarding basin is required. It is decided to mitigate the development impacts from 2-year to 100-year discharges. Due to the strongly non-linear nature of the the frequency curve and the expanding envelope of the 85% confidence interval values the HC 2-year flow might be 10x greater than the EV 2-year flow (Niguel Creek in Laguna Canyon, Orange County).

Consider the effects on the natural stream if the outlet of the retarding basin is sized to mitigate the development impact on the 2-year flow to the HC value. At Niguel Creek that value was 300 cfs. The EV 2-year flow was 30 cfs. Result: many frequent storms were not mitigated at all! The basin outlet was re-sized to the EV discharges for all recurrence intervals and the downstream channel impacts were essentially eliminated. In this case using a SMALLER discharge was more CONSERVATIVE.

II.A. Prescriptive hydrology may be strongly biased. Taking a formula out of a book written in New Jersey in 1969 and using it in Southern California in 1996 gives little assurance that the desired recurrence interval of discharge will be produced.

II.B. When negotiating subdivision approvals with developers an agency having calibrated methods has a much stronger position for rejecting non-calibrated generic hydrology submittals.

II.C. Because an agency regional calibration would typically use the best available data there can no other calibration that is superior to the agency's. If litigation results from overbank flows it is easy to assert that that the storm must have been larger than the required protection level (typically 100-year) if your design discharge is calibrated.

III. It has been shown in several journal articles that regional calibration is more "robust" than "at-site" calibration (i.e., a calibration specific to one channel location). Being robust means that the results have lower variance and, as explained in (I) above variance is the biggest problem with calibration.

Any calibration rests on the quality, quantity and representativeness of the data used in the statistical analysis.

If you do not have access to a gage network start your own. Be careful with streamgage site selection.

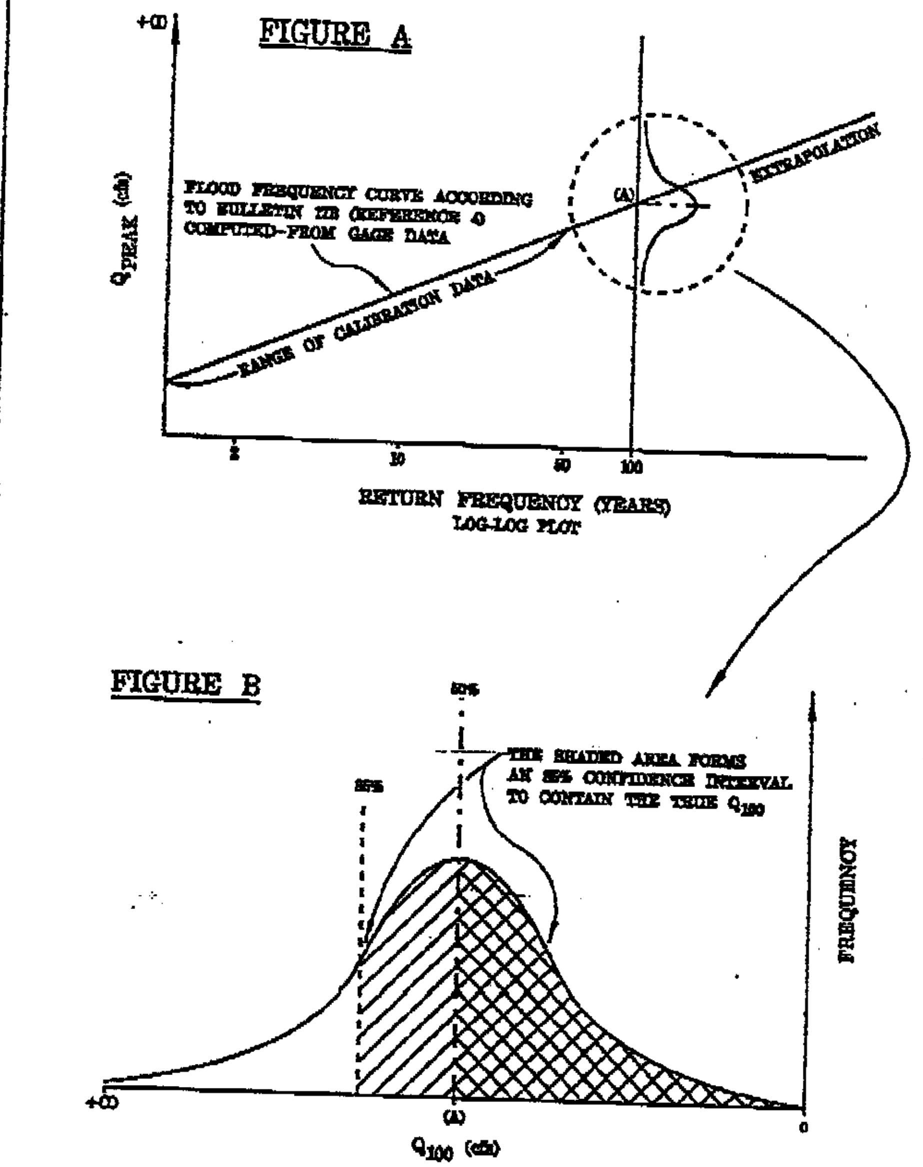
It is very important to have a "ratable" site so that stream depth can be confidently converted into discharge.

Give careful thought to the homogeneity of the statistical record at a stream gage site. If the watershed has gone from 100% bean fields to 100% urban development over the last 20 years you will get misleading results from applying standard methods to the record. Try to use watersheds whose development has changed little over the last 20 years at least.

Plot Transformations - Lagrangian to Eulerian

Remember that a streamgage is a nearly perfect integrator of the hundreds of parameters that mediate between rainfall and runoff but raingages, by definition, provide point data. The point data from raingages typically samples less than one part per billion (often one part per 10 billion) of the watershed area. For watersheds sensitive to storm durations of 3 hours or less widely spaced rain gages may not adequately sample the rain while the streamgage is likely to measure 100% of the runoff. This problem is at the heart of successful calibration. You cannot have too many raingages.

In the coming decade we may see be able to use the new National Weather Service WSR-88d (NEXRAD) radars to develop CAD-like rainfall maps for use in calibration studies. We will still need mechanical gages to calibrate the radar images so do not postpone raingage installations while waiting for NEXRAD to improve our calibrations.



FREQUENCY DISTRIBUTION OF Q_{100} VALUES
(VARIATION IN Q_{100} VALUES DUE TO SAMPLING ERROR)