Chapter 1 Introduction

• Sources of Uncertainty
  - randomness: inherent unexplainable variability of nature
  - lack of information/understanding: parameter uncertainty, modeling uncertainty, sampling uncertainty ($\neq$ data uncertainty)
  - error/inaccuracy: data uncertainty, operational uncertainty

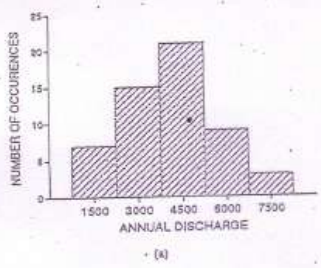
• Definitions of Statistics and Probability
  - Statistics: methods for drawing inferences about the properties of a population based on the properties of a sample from that population
  - Probability: methods for calculating the likelihood of an event given known population characteristics

• Random Variables
  - Definition
  - Discrete or continuous

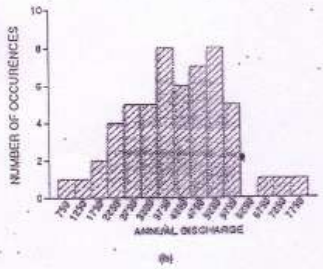
• Populations vs Samples
  - population parameters
  - sample statistics

• Graphical Display of Data
  - histogram
  - box plot
  - quantile plot

• Characteristics of Hydrologic Data
  - observational, not experimental
  - the order of occurrence (and their serial dependence) is often important
  - stochastic



(a)



(b)

STATISTICAL ANALYSIS OF HYDROLOGIC DATA

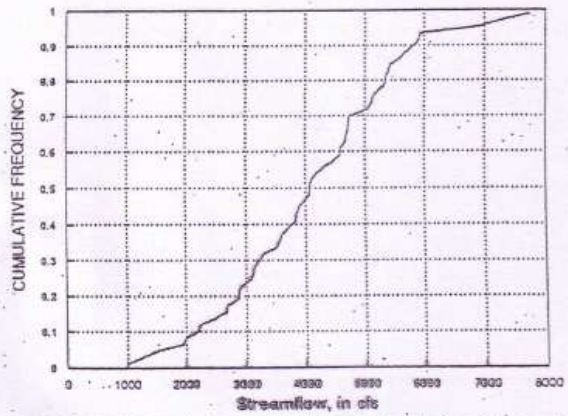


FIGURE 17.2.2 Quantile plot or cumulative distribution function of the annual stream flow, Licking River at Catawba, Ky., 1929 – 1983.

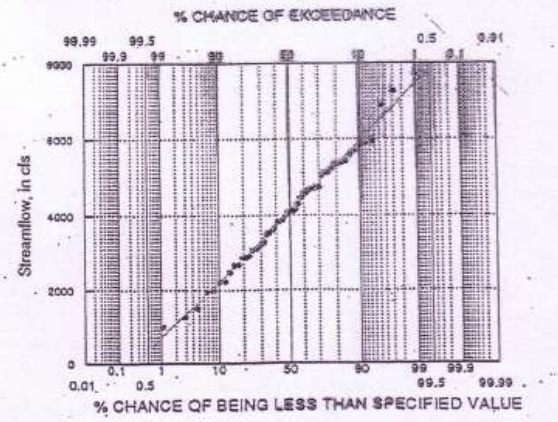


FIGURE 17.2.3 Probability plot of the annual stream flow, Licking River at Catawba, Ky., 1929 – 1983, on normal probability paper.

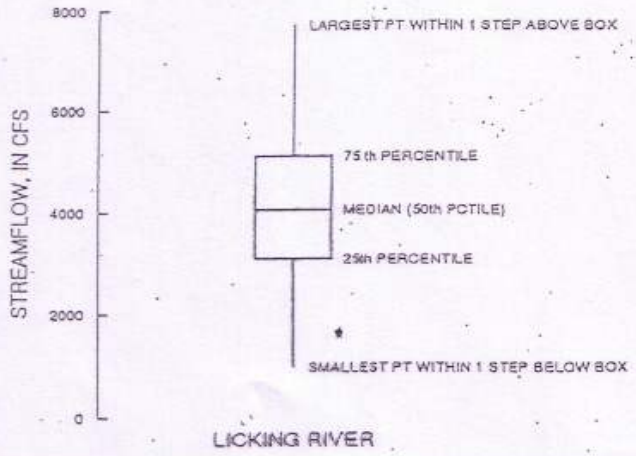
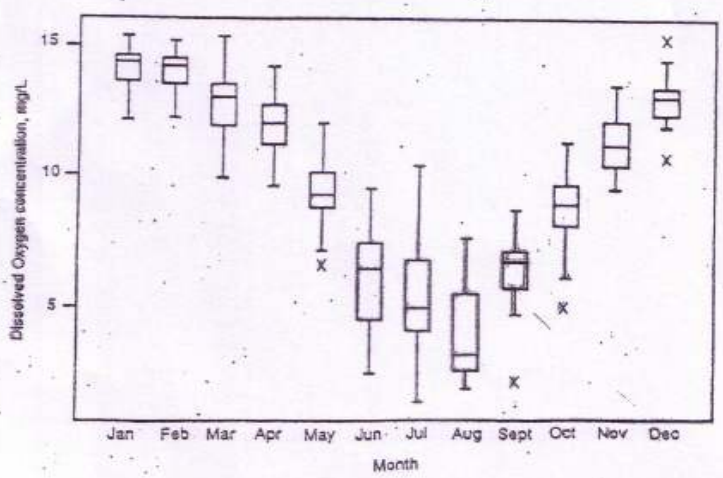


FIGURE 17.2.4 Box plot of the annual stream flow, Licking River at Catawba, Ky., 1929–1983.

STATISTICAL ANALYSIS OF HYDROLOGIC DATA



Chapter 2 Probability

< Basic Probability >

• Axioms of Probability

(i) $P(S) = 1$

(ii) $0 \leq P(A) \leq 1$

(iii) $P(A \cup B) = P(A) + P(B)$ if A & B are mutually exclusive

[note] $P(A \cup B \cup C) =$

• Conditional Probability

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$P(A \cap B) = P(A)P(B)$ if A&B are independent

- Total Probability Theorem

$$P(B) = P(A_1)P(B|A_1) + \ldots + P(A_n)P(B|A_n)$$

- Bayes Theorem

$$P(A_k \mid B) = \frac{P(A_k)P(B|A_k)}{\sum_{j=1}^{n} P(A_i)P(B|A_k)}$$

where $P(A_k)$: prior probability

$P(A_k|B)$: posterior probability

[note] $P(A \cap B \cap C) =$

< Probability Distribution >

• Cumulative Mass/Density Functions (CMF/CDF)

- discrete r.v.

- continuous r.v. $P_X(x) = \text{prob}(X \le x) = \int_{-\infty}^{x} p_X(t)\,dt$

• Probability Mass/Density Functions (PMF/PDF)

- discrete r.v.

- continuous r.v. $dP_X(x) = p_X(x)\,dx$

((note)) the pdf is not a probability and can exceed one.

• Probability Concept for Continuous R.V.

$$\text{prob}(a \le X \le b) = \int_{a}^{b} p_X(t)\,dt = P_X(b) - P_X(a)$$

$$\text{prob}(X = d) = \int_{d}^{d} p_X(t)\,dt = P_X(d) - P_X(d) = 0$$

$$\text{prob}(a \le X \le b) = \text{prob}(a < X \le b) = \text{prob}(a \le X < b) = \text{prob}(a < X < b)$$

• Bivariate and Marginal Distributions

$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x,s)\,ds$$

((note))

(i) $P_{XY}(x,\infty)$ is a cumulative univariate probability function of X only, i.e. the cumulative marginal distribution of X

(ii) $P_{XY}(-\infty, y) = P_{XY}(x, -\infty) = 0$

• Conditional Distributions

$p_{X|Y}(x|Y=y_0) = p_{XY}(x, y_0)/p_Y(y_0)$

((note))

(i) independence

$p_{X|Y}(x|y) = p_X(x)$

$p_{XY}(x, y) = p_X(x)p_Y(y)$

(ii) when the vectors **V** and **W** have a joint multivariate normal distribution, the conditional distribution of a vector **V** given the value of a vector **W** is

$$\mathbf{V}|\mathbf{W} \sim N\ [\mu_V + \Sigma_{VW}\Sigma_W^{-1}(\mathbf{w}-\mu_W),\ \Sigma_V - \Sigma_{VW}\Sigma_W^{-1}\Sigma_{VW}^T]$$

where Σ_{VW} is the covariance matrix of the vectors **V** and **W**,

Σ_W is the covariance matrix of **W** with itself, and

Σ_V is the covariance matrix of **V** with itself.

• Transformation

$p_U(u) = p_X(x)|dx/du|$

$p_{UV}(u, v) = p_{XY}(x, y)|J(\frac{x, y}{u, v})|$

• Return Period

- Definition: the average interval in years between the occurrence of a flood of specified magnitude and an equal or larger flood (= recurrence interval)

$T_X(x) = 1/[1 - P_X(x)] = 1/\text{prob}(X \geq x) = 1/p$

- the probability that at least one event that equals or exceeds the T-year event will occur in any series of N years:

$1 - (1 - p)^N$

- the probability that the first exceedance of the T-year event occurs in year k:

$p(1 - p)^{k-1}$

Chapter 3 Properties of Random Variables

< Statistics >

• Summary Statistics

- measure of central tendency

(i) arithmetic mean:

$\mu_x = E(X)$

$\bar{x} =$

((note)) E(a+bX) =

E(X+Y) =

(ii) others: median, mode, geometric mean, weighted mean

- measure of dispersion

(i) variance

Var(X) = σ^2 =

s_X^2 =

((note)) Var(a+bX) =

Var(X±Y) =

(ii) others: standard deviation, coefficient of variation, range

- measure of symmetry: skewness coefficient

- measure of peakness: kurtosis

- measure for jointly distributed random variables

(i) covariance

Cov(X,Y) = σ_{XY} =

s_{XY} =

((note)) Cov(aX+b, cY+d) =

(ii) correlation coefficient

Corr(X,Y) = ρ_{XY} =

r_{XY} =

((note)) Corr(aX+b, cY+d) =

Observed Annual Statistics. The overall sample statistics $\bar{y}$, s^2, c_v, g, and r_k are usually determined for annual hydrologic time series. Coefficients of variation c_v of annual flows are typically smaller than one, although they may be close to one or greater than one in streams in arid and semiarid regions. From an analysis of the annual flows of 126 rivers, McMahon and Mein[117] report a median value of c_v of 0.25. Coefficients of skewness g of annual flows are typically greater than zero. In some streams, small values of g are found, suggesting that annual flows are approximately normally distributed. On the other hand, in some streams of arid and semiarid regions, g can be greater than one. A range of g between -0.4 and about 2.0 and a median value of 0.40 has been reported.[117] Similarly, r_1 of annual flows are generally small but positive, although in some cases, because of sample variability, the r_1's are negative. It is quite typical to find values of r_1 in the range of $+0.0$ to 0.4 for annual stream-flow series. Yevjevich[202] found that, for a large number of rivers worldwide, the average value of r_1 was about 0.15, while McMahon and Mein[117] found a range of r_1 between -0.2 and 0.8 with a mean value of 0.23. Large values of r_1 for annual flows can be found for a number of reasons, including the effect of natural or manmade surface storage such as lakes, reservoirs, or glaciers, the effect of slow groundwater storage response, and the effect of nonstationarities. Figure 19.2.2 shows a slow-decaying correlogram r_k for the annual flows of the White Nile River at Mongalla and a fast decaying r_k for the Blue Nile River at Khartoum, while the r_k for the Nile River at Aswan lies between the other two.

< Parameter Estimation >

• Precision/Accuracy

- precision: the ability of an estimator to provide repeated estimates that are close together.

((note)) 1. due to random error
2. measured with the variance of a estimator

- accuracy: precision + unbiasedness

((note)) measured with the MSE = variance + bias^2

• Properties of Estimators

- unbiasedness: A point estimator $\hat{\Theta}$ is an unbiased estimator of the population parameter Θ if $E[\hat{\Theta}]=\Theta$. If the estimator is biased, the bias = $E[\hat{\Theta}]-\Theta$.

- consistency: An estimator $\hat{\Theta}_n$, based on a sample size n, is a consistent estimator of a parameter Θ if, for any positive number ε, $\lim_{n\to\infty} Pr[|\hat{\Theta}_n - \Theta| \le \varepsilon] = 1$.

- efficiency: An estimator that has minimum MSE among all possible unbiased estimators is called an efficient estimator.

- sufficiency: Let a sample X_1, X_2, ..., X_n be drawn randomly from a population having a probability distribution with unknown parameter Θ . Then the statistic $T=f(X_1, X_2, ..., X_n)$ is said to be sufficient for estimating Θ if the distribution of X_1, X_2, ..., X_n conditional to the statistic T is independent of Θ .

◎ 한반도의 수자원 특성

표 1 유량자료의 기본 통계 특성(연최대 일유량)

지점	자료 기간 (year)	표준편차 (cms)	평균 (cms)	변동계수 (cms)
영월	18	909.82	934.07	0.97
이목정	20	132.33	123.40	1.07
백옥포	20	583.14	330.86	1.76
상안미	20	496.53	539.93	0.92
하반정	20	181.67	150.84	1.20
방림교	20	464.29	551.49	0.84
영춘	17	2986.17	4298.53	0.69
괴산	25	298.07	476.82	0.63
경안	19	429.58	405.44	1.06
내린천	11	721.58	1100.80	0.66
원통	11	577.21	787.04	0.73
소양강 댐	33	1650.91	2122.30	0.78
서면	13	1223.23	1342.45	0.91
충주댐	21	3422.23	5345.54	0.64
고로	11	481.57	194.32	2.48
동곡	11	385.60	135.13	2.85
무성	12	324.95	395.34	0.82
미성	11	250.46	347.35	0.72
병천	12	264.04	143.64	1.83
효령	11	11.52	20.26	0.57
안동댐	30	652.17	1089.77	0.60
임하댐	15	606.76	1081.65	0.56
합천댐	18	340.04	645.62	0.53
산성교	17	86.38	55.18	1.57
이평교	17	87.21	79.88	1.09
탄부교	18	30.40	47.99	0.63
기대교	18	394.12	315.97	1.25
산계교	18	496.12	441.64	1.12
대청댐	26	1225.34	2721.28	0.45
부안댐	10	30.00	69.24	0.43
남평	28	177.34	361.73	0.49
섬진강 댐	32	311.04	616.06	0.50
평균	17.77	642.57	859.85	1.00
표준편차	5.99	781.82	1228.19	0.57
변동계수	0.34	1.22	1.42	0.57

* 출처 : 국내 홍수빈도해석 지침서 작성을 위한 연구(성장현, 2007)

• Moment Generation Function

$$M_x(t) = E[e^{tX}]$$
$$= \int_{-\infty}^{\infty} e^{tx} f_x(x)dx \qquad \text{for continuous r.v.}$$

- If the m.g.f exists, its m-th derivative at the origin (t=0) is the m-th order central moment of X.

$$\frac{dM_x(0)}{dt} = \mu_1'$$

$$\frac{d^2M_x(0)}{dt^2} = \mu_2'$$

$$\frac{d^mM_x(0)}{dt^m} = \mu_m'$$

- sample moments

$$M_1 = 0$$
$$M_2 = M_2' - \overline{X}^2$$
$$M_3 = M_3' - 3\overline{X}M_2' + 2\overline{X}^3$$
etc.

where M_r' is the r-th sample moment about the origin

M_r is the r-th sample moment about the sample mean.

((example))

< Parameter Estimation >

• Precision/Accuracy

- precision: the ability of an estimator to provide repeated estimates that are close together.

((note)) 1. due to random error
2. measured with the variance of a estimator

- accuracy: precision + unbiasedness

((note)) measured with the MSE = variance + bias^2

• Properties of Estimators

- unbiasedness: A point estimator $\hat{\Theta}$ is an unbiased estimator of the population parameter Θ if $E[\hat{\Theta}]=\Theta$. If the estimator is biased, the bias = $E[\hat{\Theta}]-\Theta$.

- consistency: An estimator $\hat{\Theta}_n$, based on a sample size n, is a consistent estimator of a parameter Θ if, for any positive number ε, $\lim_{n\to\infty} Pr[|\hat{\Theta}_{n-\Theta}|\leq\varepsilon]=1$.

- efficiency: An estimator that has minimum MSE among all possible unbiased estimators is called an efficient estimator.

- sufficiency: Let a sample X_1, X_2, ..., X_n be drawn randomly from a population having a probability distribution with unknown parameter Θ . Then the statistic $T=f(X_1, X_2, ..., X_n)$ is said to be sufficient for estimating Θ if the distribution of X_1, X_2, ..., X_n conditional to the statistic T is independent of Θ .

• Method of Moments Estimation
  - estimates population parameters using the moments of samples
  - equates the first m parameters of the distribution to the first m sample moments
  - computationally simple

  ((example))

• Maximum likelihood Estimation
  - The best value of a parameter of a probability distribution should be that value which maximizes the likelihood or joint probability of occurrence of the observed sample
  - likelihood function

  - log likelihood function

  ((example))

5.1 A REVIEW OF TERMS RELATED TO RANDOM SAMPLING

A *population* consists of all conceivable observations of a process or attribute of a component (such as the density of a batch of concrete listed in Table E.1.2). A population may consist of elements that do not exist (in a physical sense); it is then said to be conceptual. A *sample*, such as the values listed in Table E.1.1, is a subset of a population. A *random* sample is one that is representative of the population.[1] A *random variable* is a real-valued function defined on a sample space. Whether a random variable is continuous or discrete depends on how the sample space is defined.

If the population is known or assumed to have a distribution such as the normal distribution discussed in Chapter 4 but the value of a parameter θ is unknown, then we need a random sample of observations, say $X_1, X_2, \ldots, X_n$ of size n, to *estimate* θ. The joint distribution of $X_1, X_2, \ldots, X_n$ is known as the *sampling distribution* of $X_1, X_2, \ldots, X_n$. Any function of the observations that is quantifiable and does not contain any unknown parameter is called a *statistic*. A statistic is a random variable that gives us a means of *estimation*. We can determine a single number to represent θ or we can determine two numbers, which include θ within their range at a given level of probability. These procedures are discussed in the next two sections. It is also important to distinguish between an *estimator* and an *estimate*. The first is the rule or method of estimation; for example, the sample mean $\overline{X}$ is a point estimator of μ, the population mean; the second is the value that the estimator yields in a particular application.

5.2 PROPERTIES OF ESTIMATORS

An important field of statistical inference is the estimation of parameters. Alternative types of estimators, which have properties that are more or less desirable than others, can be used for such a purpose, as discussed initially in Subsection 3.2.3. In this section we summarize and exemplify these properties.

5.2.1 Unbiasedness

Given a sample of observations, our objective here is to estimate the value of a parameter θ. The observations are random variables, say, $X_1, \ldots, X_n$; hence an estimate of the parameter obtained from them, which is a statistic and a function of the observations, is also a random variable. In most cases, such a statistic can differ considerably from the true value of the parameter regardless of the method of estimation. However, we seek to find an estimator that will, on average (that is, after repeated sampling), give satisfactory results. That is, the estimator will produce statistics that are distributed according to a certain law. This law is the sampling distribution to which we referred earlier. Some types of these sampling distributions will be considered in this chapter. The law must have some desirable attributes if the estimator is to be acceptable for our purpose. For instance, if the mean value of this distribution is θ, then the estimator has the property of *unbiasedness*.

[1]More formally, a random sample is a collection $X_1, X_2, \ldots, X_n$ of random variables taken from a population with density $f(\cdot)$ if the joint density $f_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n)$.

Definition and properties. A point estimator $\hat{\theta}$ is an **unbiased estimator** of the population parameter θ if $E(\hat{\theta}) = \theta$. If the estimator is biased, the **bias** $= E(\hat{\theta}) - \theta$.

Example 5.1. Mean and variance of the sample mean. It can be shown that the sample mean $\overline{X}$ and the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

are unbiased estimators of μ and σ^2.

The first result follows immediately by taking expectations of a random sample of size n,

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n),$$

which yields

$$E(\overline{X}) = \frac{1}{n}(nE[X_i]) = \frac{1}{n}(n\mu) = \mu.$$

For the variance [as in Eqs. (1.2.6) and (1.2.7)]

$$E[S^2] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}E\left[(X_i - \mu)^2\right] - nE\left[(\overline{X} - \mu)^2\right]\right] = \frac{1}{n-1}\left\{\sum_{i=1}^{n}\sigma^2 - n\,\mathrm{Var}\left[\overline{X}\right]\right\}$$

$$= \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) = \sigma^2.$$

Unfortunately, many estimators are biased but have other desirable properties. Methods of correcting or reducing the bias, such as the jackknife and bootstrap, were discussed in Subsection 3.2.3. There are also three other properties that our ideal estimator should have. These are consistency, efficiency, and sufficiency, concepts introduced by the British statistician Fisher.

5.2.2 Consistency

A *consistent estimator* of a parameter θ produces statistics that converge to θ, in terms of probability. Thus we can define consistency as follows:

Definition and properties. An estimator $\hat{\theta}_n$, based on a sample size n, is a consistent estimator of a parameter θ if, for any positive number ϵ,

$$\lim_{n\to\infty}\Pr[|\hat{\theta}_n - \theta| \le \epsilon] = 1. \tag{5.2.1}$$

One finds, however, that sometimes an unbiased estimator may not be consistent. This case is illustrated as follows:

Example 5.2. Unbiasedness and consistency. A simple example of a consistent estimator that does not necessarily have the property of unbiasedness is found in Subsection 1.2.2 in which we considered two methods of estimating a variance σ^2. The

estimators are written as random variables, firstly by

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

and secondly by

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

As previously noted, the second equation gives the unbiased estimator of the variance. However, it can be shown (by considering the entire population as implied by the original Fisher definition of consistency) that the first equation gives a consistent estimator. Because inconsistency in this case is not considered to be a serious shortcoming, we prefer to use the second equation.

5.2.3 Minimum Variance

In practice we seldom have more than one sample, but if we have a number of samples with high variability, we may find that a single statistic that gives an estimate of a population parameter θ is quite different from the true value even if the estimator is unbiased. So we must seek an estimator that is also comparatively low in variance. Among unbiased estimators, the one with the smallest variance is called the *minimum variance unbiased estimator.*

Furthermore, it has been found that some types of estimators have a bound that is exceeded by the variance. This type is known as a minimum variance bound (mvb) estimator. The lower bound is found by what is known as the Cramer-Rao inequality.[2] Hence we obtain the following relationship:

$$\frac{\partial \ln L}{\partial \theta} = g(\theta)\{\hat{\theta} - f(\theta)\}, \tag{5.2.2}$$

where $\ln L$ is the log-likelihood function discussed in Section 3.2, $g(\theta)$ and $f(\theta)$ are functions independent of the sample of observations, and $f(\theta)$ is in a simple form such as θ or θ^2, which is relevant to the sampling distribution. Thus if an equation of the form of Eq. (5.2.2) can be obtained, $\hat{\theta}$ is a minimum variance bound estimator of $f(\theta)$. It can also be shown from Eq. (5.2.2) that

$$\mathrm{Var}[\hat{\theta}] = f'(\theta)/g(\theta), \tag{5.2.3}$$

so that if $f(\theta) = \theta$, $\mathrm{Var}[\hat{\theta}] = 1/g(\theta)$.

Definition and properties. **A minimum variance unbiased estimator** is the estimator with the smallest variance out of all unbiased estimators.

If the derivative of the log-likelihood function $\ln L$ can be put in the form

$$\frac{\delta \ln L}{\delta \theta} = g(\theta)\{\hat{\theta} - f(\theta)\},$$

then $\hat{\theta}$ is a *minimum variance bound estimator* of $f(\theta)$, with variance

$$\mathrm{Var}[\hat{\theta}] = f'(\theta)/g(\theta).$$

[2] See, for example, Stuart and Ord (1991, pp. 614–616).

Example 5.3. Minimum variance bound of the location parameter and the square of the scale parameter of the normal distribution. The pdf of the normal distribution is given by

$$\phi(x) = \frac{1}{b\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-a}{b}\right)^2\right] \qquad \text{for } -\infty < x < +\infty.$$

The estimator of the square of the scale parameter b is the variance as seen from Example 3.23. Also,

$$\frac{\delta \ln L}{\delta a} = \frac{n}{b^2}(\bar{X} - a).$$

Thus from Eq. (5.2.2), $f(\theta) = f(a) = a$; and $\bar{X}$ is an mvb estimator of a with variance b^2/n, from Eq. (5.2.3).

To estimate the variance statistic of b^2, we use the estimator for a just given. From Example 3.23 and Eq. (5.2.2), with $f(\theta) = f(b) = b^2$,

$$\frac{\delta \ln L}{\delta b} = -\frac{n}{b} + \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{b^3} = \frac{n}{b^3}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 - b^2\right].$$

Then from Eq. (5.2.2) we find that

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is an mvb estimator of b^2. Its variance is $2b^4/n$ from Eq. (5.2.3).

5.2.4 Efficiency

The term *efficiency* is used as a relative measure of the variance of the sampling distribution, with the efficiency increasing as the variance decreases. One may search unbiased estimators to find the one with the smallest variance and call it the most efficient. It seems, however, desirable to combine the properties of unbiasedness and minimum variance because an estimator can have minimum variance but still be biased, albeit to a small degree. This combination can be accomplished by means of the *mean square error* (mse) criterion. Thus if A is an estimator of θ, the mse is

$$\begin{aligned} E[(A-\theta)^2] &= E[\{(A - E[A]) - (\theta - E[A])\}^2] \\ &= E[(A - E[A])^2] + (\theta - E[A])^2 \\ &= \mathrm{Var}[A] + (\mathrm{bias})^2. \end{aligned}$$

(Note that the terms of the cross-product, $2E[(A - E[A])(E[A] - \theta)]$, sum to zero.) Thus the estimator becomes more efficient as the mse decreases.

Definition and properties. An estimator that has minimum mean square error among all possible unbiased estimators is called an efficient estimator. The mean square error of an estimator, which is equivalent to the sum of its variance and the square of its bias, can be used as a relative measure of efficiency when comparing two or more estimators.

Example 5.4. Relative efficiencies of the estimators of the mean of concrete densities. From Tables 1.2.1 and 1.2.2, the mean of the densities of 40 concrete test cubes is 2445 kg/m^3. However, if we had only the first five test cubes, the estimated mean

would be 2431 kg/m^3. Both estimators are unbiased as seen in Example 5.1. Hence the relative efficiencies, as given by the ratio of the mse values, are equivalent to the ratio of the variances; that is,

$$\frac{\sigma^2/40}{\sigma^2/5} = \frac{1}{8}$$

This result merely confirms what we already know; that is, the large-sample estimator for the mean is more efficient than that based on a smaller sample. Also, efficiency is inversely proportional to the sample size n.

The outcome of Example 5.4 notwithstanding, the minimization of variance will generally give different results from the minimization of mse.

5.2.5 Sufficiency

Properties such as unbiasedness, consistency, and minimum mean square error guide us to select the most suitable estimators, but the discussion is not complete without an explanation of sufficiency. A *sufficient estimator* gives as much information as possible about a sample of observations so that no additional information can be conveyed by any other estimator. This can also be defined more formally as follows:

Definition and properties. Let a sample $X_1, X_2, \ldots, X_n$ be drawn randomly from a population having a probability distribution with unknown parameter θ. Then the statistic $T = f(X_1, X_2, \ldots, X_n)$ is said to be sufficient for estimating θ if the distribution of $X_1, X_2, \ldots, X_n$ conditional to the statistic T is independent of θ.

We can see, for example, that the median, taken as a measure of mean density or central tendency as discussed in Subsection 1.2.1, does not contain all the information in a sample. The median is the middle value of the sample; if any other value is changed, the mean changes but the median is unaltered. It is therefore not a sufficient statistic for the purpose, unlike the mean which is discussed in Example 5.5.

Example 5.5. Normal and uniform variates. From Table 1.2.2, the mean and standard deviation of the compressive strengths are 60.14 and 5.02 N/mm^2. We also made the hypothesis (in Example 4.28) that the concrete strengths are normally distributed (which is subject to verification later in this chapter but is confirmed by numerous other studies). If the sample variance is the true value of the variance σ^2, then the sample mean $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ is a sufficient statistic for the location parameter of the normal distribution, which is the population mean μ. On the other hand, if the sample mean is the true value of μ, the sample variance $S^2 = (1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a sufficient statistic for the square of the scale parameter of the normal distribution, which is $= \sigma^2$. In practice, both parameters are unknown. However, if $\bar{X}$ and S^2 are considered jointly, these two statistics are jointly sufficient for μ and σ^2. This is because no other estimators can provide any more information for the population mean and variance.

Also consider a uniform $(0, \theta)$ distribution. Let us draw a random sample $X_1, X_2, \ldots, X_n$ from this distribution. Then for estimating θ, $X_{\max} = \max[X_1, X_2, \ldots, X_n]$ is sufficient.[3]

[3]We present results here without proof. See, for example, Casella and Berger (1990, pp. 252–254).

Example 5.6. Poisson variates. Suppose $X_1, X_2, \ldots, X_n$ is a random sample of Poisson (λ) random variables. Then it can be shown that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for λ.

The joint sampling pdf of the variate is

$$f(x_1, x_2, \ldots, x_n \mid \lambda) = \prod_{i=1}^{n}\left(\lambda^{x_i} e^{-\lambda}/x_i!\right)$$

$$= \lambda^{S_n} e^{-n\lambda}/M_n$$

where

$$S_n = \sum_{i=1}^{n} x_i$$

$$M_n = \prod_{i=1}^{n} x_i!.$$

The sum of n Poisson (λ) variables is Poisson ($n\lambda$) distributed. Thus the pdf of T is $p(t \mid \lambda) = (n\lambda)^t e^{-n\lambda}/t!$ for $t = 0, 1, 2, \ldots$.

If $S_n = t$, the conditional pdf of the summed sample values is

$$h(x_1, x_2, \ldots, x_n \mid t, \lambda) = \frac{\lambda^{S_n} e^{n\lambda}/M_n}{(n\lambda)^t e^{-n\lambda}/t!}$$

$$= \frac{t!}{n^t M_n}.$$

Because the result does not depend on λ, T is a sufficient statistic for λ.

5.2.6 Summary of Section 5.2

This formal summary of the desirable properties of point estimators discussed originally in Section 3.2 is intended to provide insight to the various methods of estimation. These were discussed initially in Chapter 3 and used in Chapter 4. They will be applied in one form or other throughout the book.

5.3 ESTIMATION OF CONFIDENCE INTERVALS

In Chapters 3 and 4 we discussed and applied methods of estimating the values of one or more parameters of a population; in Section 5.2 we examined more closely the properties of the resulting estimators. We have seen that point estimates can be erroneous; in fact the probability that an estimate is equal to an unknown parameter is zero. The resulting uncertainty can be quantified by the relative variances or mean square errors of the estimators. The next step of inference is interval estimation; here we determine two numbers, say, a and b, that are expected to include within their range an unknown parameter θ in a specified percentage of cases after repeated experimentation under identical conditions. That is, in place of one statistic that estimates θ, we find a range specified by two statistics, which includes it at a given level of probability. The end points a and b of this range are known as *confidence limits*, and the interval (a, b) is known as the *confidence interval*. We

(*b)

Supplementary Exercises (31–38)

31. An estimator $\hat{\theta}$ is said to be **consistent** if for any $\epsilon > 0$, $P(|\hat{\theta} - \theta| \geq \epsilon) \to 0$ as $n \to \infty$. That is, $\hat{\theta}$ is consistent if, as the sample size gets larger, it is less and less likely that $\hat{\theta}$ will be further than ϵ from the true value of θ. Show that $\bar{X}$ is a consistent estimator of μ when $\sigma^2 < \infty$ by using Chebyshev's inequality from Exercise 43 of Chapter 3. (*Hint:* The inequality can be rewritten in the form

$$P(|Y - \mu_Y| \geq \epsilon) \leq \sigma_Y^2/\epsilon^2$$

Now identify Y with $\bar{X}$.)

31. $$P(|\bar{X}-\mu|>\varepsilon)=P(\bar{X}-\mu>\varepsilon)+P(\bar{X}-\mu<-\varepsilon)=P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}>\frac{\varepsilon}{\sigma/\sqrt{n}}\right)+P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}<\frac{-\varepsilon}{\sigma/\sqrt{n}}\right)$$

$$=P\left(Z>\frac{\sqrt{n}\varepsilon}{\sigma}\right)+P\left(Z<\frac{-\sqrt{n}\varepsilon}{\sigma}\right)=\int_{\sqrt{n}\varepsilon/\sigma}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz+\int_{-\infty}^{\sqrt{n}\varepsilon/\sigma}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz$$

As $n\to\infty$, both integrals $\to 0$ since $\lim_{c\to\infty}\int_{c}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz=0$.

chebyshev's inequlity ((hann))p.63

$$\Pr(|Y-\mu_Y| \geq k\alpha_Y) \leq \frac{1}{k^2}$$

$$\rightarrow \Pr(|Y-\mu_Y| \geq \epsilon) \leq \frac{\sigma_Y^2}{\epsilon^2}$$

if $Y \rightarrow \bar{x}$, then $\mu_Y \rightarrow \mu_{\bar{x}} = \mu_x$ and $\sigma_Y^2 \rightarrow \sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n}$

$$\Pr(|\bar{x}-\mu_X| \geq \epsilon) \leq \frac{\sigma_{x^2}}{\epsilon n}$$

$n \rightarrow \infty$

$\Pr(\ ) = 0$

Lecture 4
Advanced Hydrolog
Dr. Kim, Young-Oh

• Method of Moments Estimation
  - estimates population parameters using the moments of samples
  - equates the first m parameters of the distribution to the first m sample moments
  - computationally simple

  ((example))

• Maximum likelihood Estimation
  - The best value of a parameter of a probability distribution should be that value which maximizes the likelihood or joint probability of occurrence of the observed sample
  - likelihood function

  - log likelihood function

  ((example))

value of the quantity to be estimated. The sample mean is an unbiased estimate of μ_X because

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu_X \tag{3.38}$$

The estimator v_X^2 is a biased estimator of σ_X^2 where [1]

$$E[v_X^2] = \frac{n-1}{n}\sigma_X^2 \tag{3.39}$$

Hence if many samples of size n are taken in a simulation experiment and the resulting values of v_X^2 are averaged, the average will approach $(n-1)\sigma_X^2/n$ and not σ_X^2 as might be desired. For this reason, the unbiased estimate of the variance

$$s_X^2 = \frac{nv_X^2}{n-1} \tag{3.40}$$

is often used instead of v_X^2. There is relatively little difference between the two estimators for moderate size n. Both, however, generally produce biased estimates of the standard deviation σ_X [35], although the bias decreases with increasing n.

The second important statistic often used to assess the accuracy of an estimator $\hat{\theta}$ is its variance $\mathrm{Var}(\hat{\theta})$, which equals $E\{(\hat{\theta} - E[\hat{\theta}])^2\}$. For the mean of a set of independent observations, the variance of the sample mean is given by

$$\mathrm{Var}(\bar{x}) = \frac{\sigma_X^2}{n} \tag{3.41}$$

It is common to call $\sigma_X/\sqrt{n}$ the *standard error* of $\bar{x}$ rather than its standard deviation.

The bias measures the difference between the average value of an estimator and the quantity to be estimated. The variance measures the spread or width of the estimator's distribution. Both contribute to the amount by which an estimate deviates from the quantity to be estimated. These two errors are often combined into the *mean square error*, defined as

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] = \{E[\hat{\theta}] - \theta\}^2 + E\{(\hat{\theta} - E[\hat{\theta}])^2\} \\ &= [\mathrm{Bias}]^2 + \mathrm{Var}(\hat{\theta})\end{aligned} \tag{3.42}$$

where Bias is $E(\hat{\theta}) - \theta$. The MSE is the expected or average squared deviation of the estimator from the true value of the parameter. It is a convenient measure of how closely $\hat{\theta}$ approximates θ.

Estimation of the coefficient of skewness γ_X provides a good example of the use of the MSE for evaluating the total deviation of an estimate from the true population value. The sample estimate $\hat{\gamma}_X$ of γ_X is often biased, has a large variance, and was shown by Kirby [21] to be bounded so that

$$|\hat{\gamma}_X| \leq \frac{n-2}{(n-1)^{1/2}} \tag{3.43}$$

where n is the sample size. The bounds do not depend on the true skew γ_X. The bias and variance of $\hat{\gamma}_X$ depend on the sample size and the actual distribution of X. Table 3.2 contains the expected value and standard deviation of $\hat{\gamma}_X$ when X has either a normal distribution, for which $\gamma_X = 0$, or a gamma distribution with $\gamma_X = 0.25$, 0.50, 1.00, 2.00, or 3.00.

TABLE 3.2. Sampling Properties of Estimate of Coefficient of Skewness

Expected Value of $\hat{\gamma}_X$

Distribution of X	SAMPLE SIZE 10	20	50	80
Normal $\gamma_X = 0$	0.00	0.00	0.00	0.00
Gamma $\gamma_X = 0.25$	0.13	0.18	0.22	0.23
$\gamma_X = 0.50$	0.26	0.36	0.44	0.46
$\gamma_X = 1.00$	0.51	0.70	0.85	0.91
$\gamma_X = 2.00$	0.97	1.32	1.63	1.74
$\gamma_X = 3.00$	1.34	1.82	2.25	2.49
Upper bound on skew	2.67	4.13	6.86	8.78

Standard Deviation of $\hat{\gamma}_X$

Distribution of X	SAMPLE SIZE 10	20	50	80
Normal $\gamma_X = 0$	0.58	0.47	0.33	0.26
Gamma $\gamma_X = 0.25$	0.58	0.48	0.34	0.27
$\gamma_X = 0.50$	0.58	0.49	0.36	0.30
$\gamma_X = 1.00$	0.59	0.53	0.43	0.37
$\gamma_X = 2.00$	0.61	0.63	0.60	0.56
$\gamma_X = 3.00$	0.62	0.70	0.75	0.76

Source: J. R. Wallis, N. C. Matalas, and J. R. Slack, *Just a Moment! Appendix*, National Technical Information Service, PB-231 816, Springfield, Va., 1974.

To illustrate the magnitude of these errors, consider the mean square error of $\hat{\gamma}_X$ calculated from a sample of size 50 when X has a gamma distribution with $\gamma_X = 0.50$, a reasonable value for annual streamflows. The expected value of $\hat{\gamma}_X$ is 0.44; its variance equals $(0.36)^2$, its standard deviation squared. Using equation 3.42, the mean square error of $\hat{\gamma}_X$ is

$$\begin{aligned} \text{MSE}(\hat{\gamma}_X) &= (0.44 - 0.50)^2 + (0.36)^2 = 0.0036 + 0.1296 \\ &= 0.133 \end{aligned} \tag{3.44}$$

An unbiased estimate of γ_X is $(0.50/0.44)\hat{\gamma}_X$; the mean square error of this unbiased estimate of γ_X is

$$\text{MSE}\left(\frac{0.50\hat{\gamma}_X}{0.44}\right) = (0.50 - 0.50)^2 + \left[\left(\frac{0.50}{0.44}\right)(0.36)\right]^2 = 0.167 \quad (3.45)$$

The mean square error of the unbiased estimate of γ_X is larger than the mean square error of the biased estimate. Unbiasing $\hat{\gamma}_X$ results in a larger mean square error for all the cases listed in Table 3.2 except for the normal distribution and the gamma distribution with $\gamma_X = 3.00$.

As shown here for the skew coefficient, biased estimators often have smaller mean square errors than unbiased estimators; because the mean square error measures the total average deviation of an estimator from the quantity being estimated, this result demonstrates that the strict or unquestioning use of unbiased estimators is not advisable. Additional information on the sampling distribution of quantiles and moments is contained in Appendix 3A.

3.3 DISTRIBUTIONS OF RANDOM EVENTS

A frequent task in water resources planning is the development of a model of some probabilistic or stochastic phenomena such as streamflows, flood flows, rainfall, temperatures, or evaporation. This generally requires that one fit a probability distribution function to a set of observed values of the random variable. Sometimes, one's immediate objective is to estimate a particular quantile of the distribution, such as the 100-year flood or the 7-day, 10-year low flow. Then the fitted distribution will supply an estimate of this quantity. In a stochastic simulation model, fitted distributions are used to generate possible values of the random variable in question.

This section provides a brief introduction to the techniques useful for estimating the parameters of a probability distribution function and determining if the fitted distribution provides a reasonable or acceptable model of the data. Sections are also included on families of distributions based on the normal and gamma distributions. These two families have found frequent use in water resource planning.

3.3.1 Parameter Estimation and Model Adequacy

Given a set of observations to which a distribution is to be fit, one first selects a distribution function to serve as a model of the distribution of the data. The choice of distribution may be based on experience with data of that type, some understanding of the mechanisms giving rise to the data,

See Another File