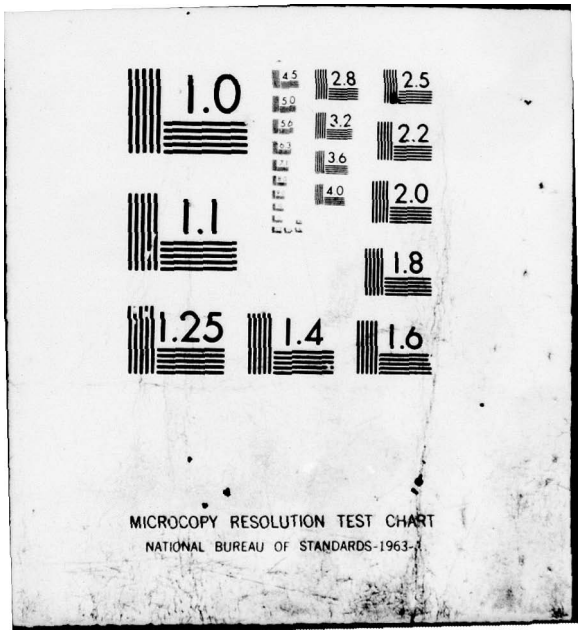
AD-A076 198 AIR FORCE MATERIALS LAB WRIGHT-PATTERSON AFB OH F/G 11/4
BASIC CONCEPTS OF STATISTICS AND THEIR APPLICATIONS IN COMPOSIT--ETC(U
JUN 79 W J PARK
UNCLASSIFIED AFML-TR-79-4070 NL

/OF/
AD
A076198

END
DATE
FILMED
11-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFML-TR-79-4070

BASIC CONCEPTS OF STATISTICS AND THEIR APPLICATIONS IN COMPOSITE MATERIALS

MECHANICS AND SURFACE INTERACTIONS BRANCH
NONMETALLIC MATERIALS DIVISION

JUNE 1979

TECHNICAL REPORT AFML-TR-79-4070
Final Report for Period 1 December 1978 – 31 January 1979

AIR FORCE MATERIALS LABORATORY
AIR FORCE WRIGHT AERONAUTICAL LABORATORIES
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This report has been reviewed by the Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

Won J. Park

WON J. PARK, Project Engineer
Mechanics & Surface Interactions Br.
Nonmetallic Materials Division

S. W. TSAI, Chief
Mechanics & Surface Interactions Br.
Nonmetallic Materials Division

FOR THE COMMANDER

J. M. KELBLE, Chief
Nonmetallic Materials Division

"If your address has changed, if you wish to be removed from our mailing list, or if the addressee is no longer employed by your organization please notify AFML/MBM *,W-PAFB, OH 45433 to help us maintain a current mailing list".*

Copies of this report should not be returned unless return is required by security considerations, contractual obligations, or notice on a specific document.

AIR FORCE/56780/28 September 1979 — 300

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFML-TR-79-4070	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER: Final rept.	
4. TITLE (and Subtitle): BASIC CONCEPTS OF STATISTICS AND THEIR APPLICATIONS IN COMPOSITE MATERIALS.	5. TYPE OF REPORT & PERIOD COVERED: 1 Dec. 1978 — 31 Jan. 1979,
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Won J. Park	8. CONTRACT OR GRANT NUMBER(s): In-House
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Air Force Materials Laboratory, Air Force Wright Aeronautical Laboratories, AFSC, Wright-Patterson AFB, Ohio 45433	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2303 / 2303 D4
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Materials Laboratory (AFML/MBM), Air Force Wright Aeronautical Laboratories, AFSC, Wright-Patterson AFB, Ohio 45433	12. REPORT DATE: JUNE 1979
	13. NUMBER OF PAGES: 70
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): 70	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Sample
Population
Probability distribution function
Estimation
Confidence Interval
Testing Hypothesis

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report is intended to present, in a simple and concise way, basic concepts of statistics and their applications in composite material engineering. The subjects are selected from the standpoint of key concepts and applications rather than theory and elegance. It is envisioned that this report is suitable for a short course (or refresher course) for composite material engineers, who have been working with data but have not done any formal course work in statistics.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

012320 LB

FOREWORD

This report was prepared in the Mechanics and Surface Interactions Branch (AFML/MBM), Nonmetallic Materials Division, Air Force Materials Laboratory, Air Force Wright Aeronautical Laboratories, Wright-Patterson AFB, Ohio under the Project No. 2303, Task No. 2303/D4. The time period covered by the effort was December 1, 1978 to January 31, 1979.

Dr. Won J. Park was the Project Engineer - a visiting scientist from Wright State University, Dayton, Ohio - under the university resident research program of Air Force Office of Scientific Research.

The author wishes to acknowledge Marvin Knight, AFML/MBM, for valuable suggestions on the report.

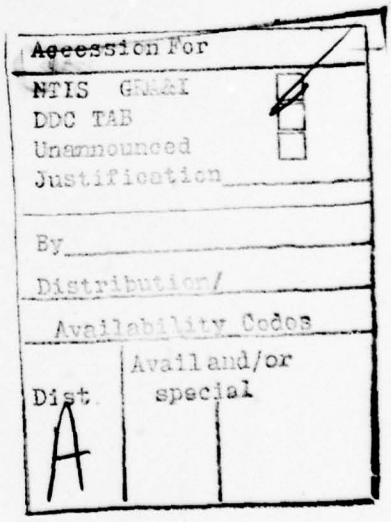


TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

Engineers often obtain data either to verify empirically a postulate of theory or to measure experimentally a relevant characteristic of a phenomenon.

Statistics is a methodology to interpret data and to make conclusions on a relevant characteristic from data.

There are basically two approaches in engineering analysis; one is deterministic and the other is probabilistic. This is illustrated by the following example.

Example 1.1. Suppose that we are going to hit a golf ball with an initial velocity v_o and angle α and to measure the horizontal distance x that it travels.

(a) Deterministic: Newton's gravitational law gives

$$x = \frac{v_o^2 \sin 2\alpha}{g}, \text{ where}$$

g is gravitational constant.

This result is applicable only under many ideal conditions.

(b) Probabilistic: Under the real condition, in which there may be presented irregular effects of air resistance, wind speed, moisture content of air, temperature change, etc., the deterministic modeling may be impossible. One of the probabilistic approaches is to conduct 100 identical experiments of hitting a golf ball and study their scattering and distribution of the distances that they traveled.

2. BASIC DATA ANALYSIS

We introduce in this section a basic technique of data handling.

Definition 2.1. The collection of all possible measurements of a characteristic is called a population. A collection of actual measurements on the characteristic (actual data) is called a sample. The number of measurements in the sample is called the sample size.

A population can be infinite or finite. A sample is always finite.

We want to draw statistical conclusions about the population characteristics on the basis of a sample of measurements on that characteristic.

Let us denote each measurement in a sample by $x_1, x_2, \ldots, x_n$, and let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be the arrangement of data in increasing order of magnitude, of the sample, i.e. $x_{(1)}$ is the smallest measurement and $x_{(n)}$ is the largest measurement in the sample.

Definition 2.2. The sample mean denoted by $\bar{x}$ is defined as the arithmetic mean of measurements;

$$\bar{x} = \frac{1}{n}(x_1+x_2+\ldots+x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The sample median denoted by $\tilde{x}$ is defined as the measurement of the middle magnitude in the sample;

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ 1/2\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}\right)+1}\right) & \text{if } n \text{ is even.} \end{cases}$$

The sample mean or sample median describes the "central tendency" of the sample (sample distribution).

Definition 2.3. The sample variance denoted by s^2 is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{and}$$

the sample standard deviation denoted by s is defined as

$$s = \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2} = [s^2]^{1/2}.$$

The sample standard deviation indicates the "dispersion" of the sample distribution.

Data A. Suppose that in an experiment of tossing a coin three times we have observed "the number of heads". The outcomes of the experiment in 10 trials are;

$$x_1=1, \quad x_2=0, \quad x_3=2, \quad x_4=2, \quad x_5=3,$$
$$x_6=0, \quad x_7=1, \quad x_8=1, \quad x_9=2, \quad x_{10}=2.$$

In the above experiment, our interest is "the number of heads".

The arrangement of the data in increasing order of magnitude is;

$$x_{(1)}=0, \quad x_{(2)}=0, \quad x_{(3)}=1, \quad x_{(4)}=1, \quad x_{(5)}=1,$$
$$x_{(6)}=2, \quad x_{(7)}=2, \quad x_{(8)}=2, \quad x_{(9)}=2, \quad x_{(10)}=3.$$

Example 2.1. The computations of $\bar{x}$, $\tilde{x}$ and s for Data A.

$$\bar{x} = \frac{1}{10} (1+0+\ldots+2) = \frac{14}{10} = 1.4$$

$$\tilde{x} = \frac{x_{(5)}+x_{(6)}}{2} = \frac{1+2}{2} = 1.5$$

$$s^2 = \frac{1}{10-1}\left[(1-1.4)^2+(0-1.4)^2+\ldots+(2-1.4)^2\right] = .9334$$

$$s = [s^2]^{1/2} = (.9334)^{1/2} = .9661.$$

When there is a large number of measurements in a sample, the information of the sample is usually summarized as a "frequency distribution". This is done by grouping the raw data into adjoining classes of measurements. The number of measurements in each class gives the "frequency" for that class of measurements. This grouping evidently suppresses the detail of individual measurements, but it provides a comprehensive "picture" of the sample. For example, it shows in which class the measurements occur most often and over what range of classes the measurements occur a large portion of the time. The "relative frequency" is defined as the frequency divided by the total number of samples.

Example 2.2. The frequency distribution for Data A.

Class (x_i)	Frequency (f_i)	Relative Frequency (p_i)
0	2	.2
1	3	.3
2	4	.4
3	1	.1
Total	n=10	1

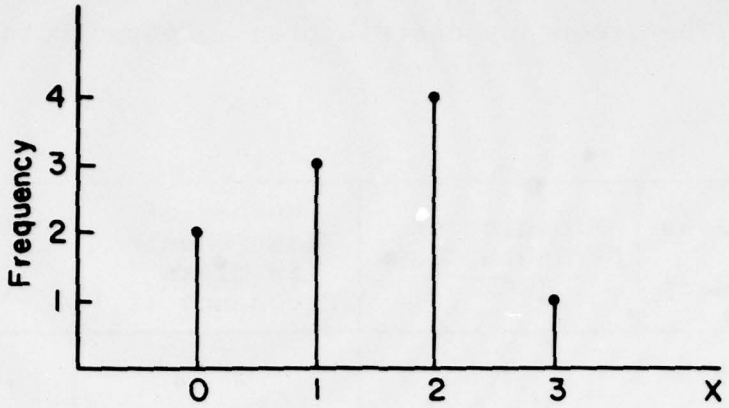


Figure 2.1. Histogram for Data A.

Data B. Tensile strength data for the composite AS/3501-5A/10 ply/0° in KSI, (ranked data).

	KSI	ranked		KSI	ranked
1	209	152	19	226	211
2	207	159	20	220	213
3	237	184	21	159	214
4	225	186	22	224	217
5	238	195	23	206	218
6	201	196	24	198	218
7	226	196	25	152	220
8	204	196	26	218	220
9	223	198	27	220	223
10	225	200	28	196	224
11	218	200	29	186	225
12	244	201	30	200	225
13	208	204	31	214	226
14	196	204	32	213	226
15	200	206	33	234	234
16	204	207	34	211	237
17	184	208	35	217	238
18	195	209	36	196	244

Example 2.3. The frequency distribution (grouped data) for Data B.

Strength Class (KSI) Class Boundary	Midpoint of Strength Class (x_i^*)	Number of Measurements in Class Frequency (f_i)	Relative Frequency (p_i)
149.5 - 159.5	154.5	2	.0555
159.5 - 169.5	164.5	0	0
169.5 - 179.5	174.5	0	0
179.5 - 189.5	184.5	2	.0555
189.5 - 199.5	194.5	5	.139
199.5 - 209.5	204.5	9	.250
209.5 - 219.5	214.5	6	.167
219.5 - 229.5	224.5	8	.222
229.5 - 239.5	234.5	3	.083
239.5 - 249.5	244.5	1	.028
	Total	36	1.000

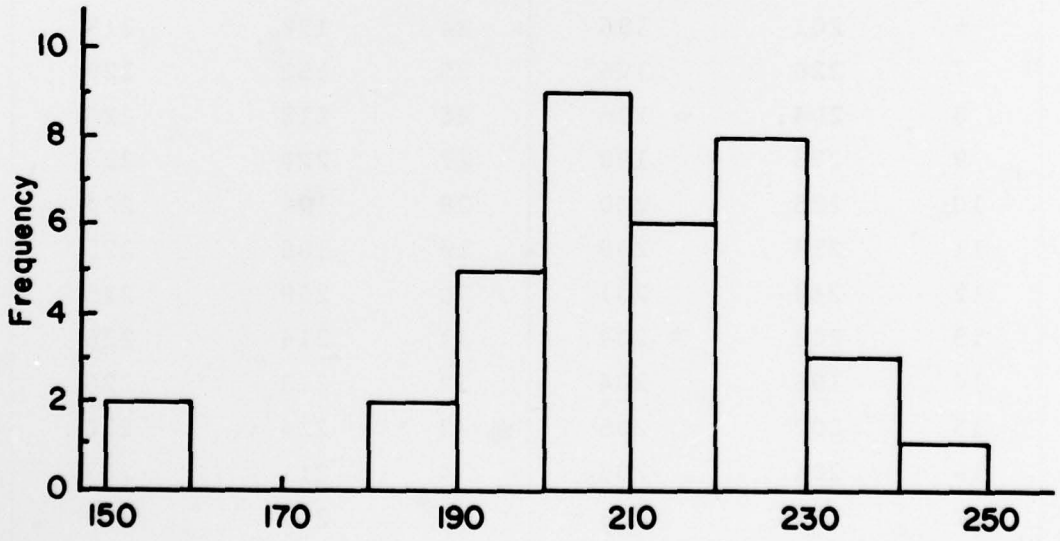


Figure 2.2. Histogram or Frequency Distribution for Data B.

The first column of Example 2.3 lists the "class boundaries", defining the class to which each measurement is assigned. These boundaries are usually given to a higher significant digit than the sample measurements, to ensure that measurements can be grouped unambiguously. Class boundaries are preferably chosen to give equal widths (e.g. 10 KSI in Example 2.3). The class width is selected to make tabulation convenient, and to result in 6 to 15 groups when covering the entire range of measurements.

Column 2 of Example 2.3 lists the "midpoint" of each class. This value is needed to represent the measurements in the class. Column 4 lists the "relative frequency" (relative to the sample size). Relative frequency tabulations are the usual form in which sample data are presented for statistical analysis. A "histogram" presents the number of occurrences in a class as the height of a rectangle over the corresponding measurement interval (see Figure 2.2).

The central tendency and dispersion of the sample distribution can be computed directly from the grouped data (as noted by the symbol *);

$$\bar{x}^* = \frac{1}{n} \sum_{i=1}^{k} x_i^* f_i \quad \text{and}$$

$$s^* = \left\{ \frac{1}{n-1} \sum_{i=1}^{k} (x_i^* - \bar{x}^*)^2 \, f_i \right\}^{1/2}$$

$$= \left\{ \frac{1}{n-1} \left[\sum_{i=1}^{k} (x_i^*)^2 \, f_i - \frac{1}{n} \left(\sum_{i=1}^{k} x_i^* \, f_i \right)^2 \right] \right\}^{1/2}$$

where

n = sample size,

k = number of classes,

x_i^* = midpoint of class,

f_i = frequency of class, and

$$\tilde{x}^* = L + \frac{n/2 - S_m}{f_m} w,$$

where

L = the lower boundary of the class containing the median,

S_m = number of measurements less than L

w = the width of the median class

f_m = the frequency of measurements in the median class.

Example 2.4. The computations for Data B.

(a) From the raw data;

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{36} (209+207+...+196)$$

$$= 209.28,$$

$$s = \left\{ \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\}^{1/2}$$

$$= \left\{ \frac{1}{36-1} \left[(209-209.28)^2 + (207-209.28)^2 \right. \right.$$

$$\left. \left. +...+(196-209.28)^2 \right] \right\}^{1/2}$$

$$= 19.65,$$

$$\tilde{x} = \frac{x_{(18)} + x_{(19)}}{2} = \frac{209 + 211}{2} = 210 .$$

(b) From the grouped data (Example 2.3);

x_i^*	f_i	$x_i^* f_i$	$(x_i^*)^2$	$(x_i^*)^2 f_i$
154.5	2	309.0	23,870.25	47,740.50
164.5	0	0	27,060.25	0
174.5	0	0	30,450.25	0
184.5	2	369.0	34,040.25	68,080.50
194.5	5	972.5	37,830.25	189,151.25
204.5	9	1,840.5	41,820.25	376,382.25
214.5	6	1,287.0	46,010.25	276,061.50
224.5	8	1,796.0	50,400.25	403,202.00
234.5	3	703.5	54,990.25	164,970.75
244.5	1	244.5	59,780.25	59,780.25

$$\sum_{i=1}^{10} x_i^* f_i = 7,522.0 \qquad \sum_{i=1}^{10} (x_i^*)^2 f_i = 1,585,369.00$$

$$\bar{x}^* = \frac{1}{n} \sum_{i=1}^{k} x_i^* f_i = \frac{1}{36}\, 7,522 = 208.94,$$

$$s^* = \left\{ \frac{1}{n-1} \left[\sum_{i=1}^{k} (x_i^*)^2 f_i - \frac{1}{n} \left(\sum_{i=1}^{k} x_i^* f_i \right)^2 \right] \right\}^{1/2}$$

$$= \left\{ \frac{1}{36-1} \left[1,585,369 - \frac{1}{36} (7,522)^2 \right] \right\}^{1/2}$$

$$= 19.82,$$

$$\tilde{x}^* = L + \frac{\frac{n}{2} - S_m}{f_m} w = 199.5 + \frac{18-9}{9} \cdot 10$$

$$= 209.5 .$$

When we compare $\bar{x}$ and $\bar{x}^*$, s and s* and $\tilde{x}$ and $\tilde{x}^*$, there are small discrepancies due to the fact that the midpoints of classes in the frequency distribution represent the measurements in the corresponding classes. These discrepancies are negligible for reasonably large sample size.

3. POPULATION DISTRIBUTIONS

The preceding tabulations, graphs, and measures refer to a sample; they are merely descriptions of the data comprising the sample. Statistical inferences are conclusions on a population characteristic, drawn from information contained in the sample.

The manifestation of a random phenomenon is a sample of observations. The measurement value of an observation cannot be predicted, but the relative frequency of occurrence of its value tends toward a stable value in a long sequence of observations. Probability theory, therefore, abstracts the random phenomenon by dealing only with the existence of a stable frequency pattern.

Definition 3.1. The outcome of an observation on a random phenomenon is called a random event, and the totality (population) of all possible distinct events, which are associated with a particular random phenomenon, is called the sample space.

The investigation of an engineering phenomenon involves measurements on a relevant characteristic of that phenomenon. Such a characteristic, therefore, can be represented by a "measurement variable". In the investigation of such a measurement variable, the occurrence of a particular measurement x is an "event" that is represented by a specific real number. The variable X, which gives rise to measurement x, is then termed a "random variable". The measurement value x is termed a realization of the random variable X.

The relative frequency of occurrence of a random event is represented by the probability of that event. It follows that each measurement x is associated with a probability value, $P_r\{X=x\}$, which represents the long-run relative frequency of the random variable X taking on the measured value x.

In the analysis of random phenomenon, two types of random variables are frequently encountered.

Definition 3.2. A random variable X is called discrete if it only assumes finite or countably infinite distinct measurement values. A random variable X is called continuous if X can be regarded as realizable over a continuous segment of the real line, i.e. a realization x may be any real number within some interval.

Example 3.1. Examples of random variables;

(a) Discrete;

X = the number of heads in a game of tossing a fair coin 3 times,

Y = the number of defects in a product,

Z = the number of arrivals of cargo vessels at a dock.

(b) Continuous;

T = the measurements on the life length of devices,

S = the strength of a composite material,

U = the flight range of a missile.

Example 3.2. Examples of events (continuation of Example 3.1)

{X=2} = the event that X is equal to 2, i.e. the event that the number of heads is 2, and

{S>50} = the event that the strength of a composite material is greater than 50.

(A) Discrete Random Variables

It is defined that a discrete random variable X can assume finite or countably infinite values, (say, a_1, a_2, ..., a_k) and the probability value $P_r\{X=a_i\}$ ($i=1, 2, \ldots, k$) represents the long-run relative frequency of the random variable X taking the value a_i. These long-run relative frequencies form a function, which is called the probability mass function of X.

Definition 3.3. $p(a_i) = P_r\{X=a_i\}$ as a function of a_i, $i = 1, 2, \ldots, k$, is called the probability mass function (sometimes abbreviated as p.m.f.) of the random variable X.

The graph of p.m.f. of X is;

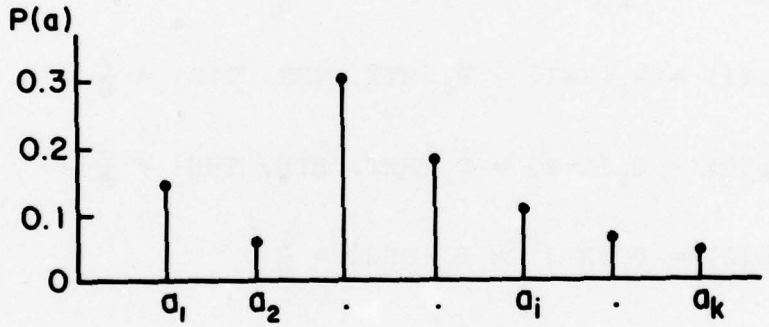


Figure 3.1. p.m.f. of X

The properties of $p(a_i)$ are,

1. $p(a_i) \geq 0$, all a_i

2. $\sum_{a_i} p(a_i) = 1$,

which follow from the properties of long-run relative frequency of random variable.

Example 3.3. The random variable X indicates the number of heads in an experiment of tossing a fair coin three times.

The sample space, S, is

$$\{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

and

$$P_r\{HHH\} = P_r\{\text{each outcome}\} = \frac{1}{8}$$

This random variable X takes values from

$$\{0, 1, 2, 3\} \text{ and}$$

$$p(0) = P_r\{X=0\} = P_r\{TTT\} = \frac{1}{8}$$

$$p(1) = P_r\{X=1\} = P_r\{HTT, THT, TTH\} = \frac{3}{8},$$

$$p(2) = P_r\{X=2\} = P_r\{HHT, HTH, THH\} = \frac{3}{8},$$

$$p(3) = P_r\{X=3\} = P_r\{HHH\} = \frac{1}{8}.$$

p.m.f. of X is;

k	0	1	2	3
p(k)	1/8	3/8	3/8	1/8

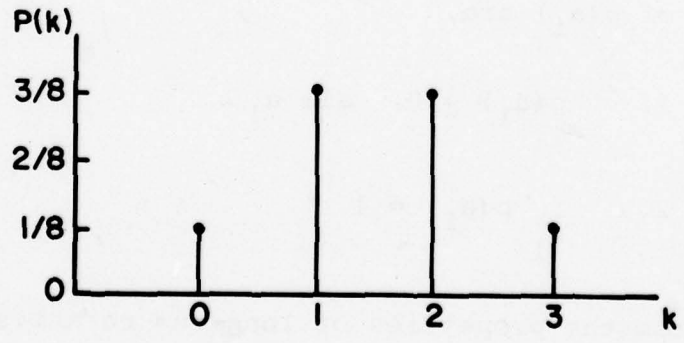


Figure 3.2. p.d.f. of X

One of the most important concepts in statistics is that of a mathematical expectation (or expected value). Let $p(a_i)$, $i = 1, 2, \ldots, k$ be the probability mass function of a random variable X.

Definition 3.4.

(a) The expected value (mean) of the random variable X is defined by

$$\mu = E(X) = \sum_{i=1}^{k} a_i p(a_i),$$

(b) The variance of the random variable of X is defined by

$$\sigma^2 = Var(X) = E(X-\mu)^2 = \sum_{i=1}^{k} (a_i-\mu)^2 p(a_i),$$

$$= \sum_{i=1}^{k} a_i^2 p(a_i) - \mu^2$$

(c) The standard deviation of the random variable X is defined by

$$\sigma = s.d.(X) = [Var(X)]^{1/2},$$

(d) In general, for a function, g(x), the expectation of g(X) is defined by

$$E[g(X)] = \sum_{i=1}^{k} g(a_i) p(a_i).$$

Example 3.4. Let p.m.f. of a random variable X be

$$p(0) = 1/8,\ p(1) = 3/8,\ p(2) = 3/8 \text{ and } p(3) = 1/8$$

(See Example 3.3), then

$$\mu = \sum_{k=0}^{3} kp(k) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8}$$

$$= \frac{12}{8} = 1.5$$

$$\sum_{k=0}^{3} k^2 p(k) = 0^2 \times \frac{1}{8} + 1^2 \times \frac{3}{8} + 2^2 \times \frac{3}{8} + 3^2 \times \frac{1}{8},$$

$$\sigma^2 = \sum_{k=1}^{3} k^2 p(k) - \mu^2 = 3 - (1.5)^2 = .75$$

$$\sigma = [\text{Var}(X)]^{1/2} = (.75)^{1/2} = .866.$$

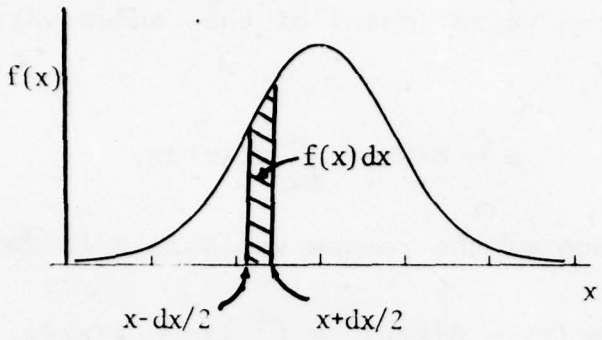
It is noted that the random variable X, "the number of heads in tossing a fair coin three times", given in Example 3.3, has such a simple structure that its probability distribution (population distribution) has been easily figured out. It is not always easy to characterize or find the probability distribution for most engineering measurement random variables. The distinction between sample relative frequency distribution (mean or standard deviation) and population probability distribution (mean or standard deviation) can be clearly observed from Example 2.2 (Example 2.1) and Example 3.3 (Example 3.4). The role of statistics is to draw some conclusions on the population (random variable X) from the information contained in a sample of measurements on X.

Now we are going to consider continuous random variables.

(B) Continuous Random Variables

Definition 3.4. The probability of a continuous random variable X taking a measurement value in the interval $(x-\frac{dx}{2}, x+\frac{dx}{2})$, as a function of x, is written in terms of the probability density function, (p.d.f.), f(x) as

$$P_r\{x - \frac{dx}{2} \leq X \leq x + \frac{dx}{2}\} = f(x)dx.$$



The sample relative frequency, when the class interval reduces toward zero and the sample size increased sufficiently large, approaches a smooth curve, which will be p.d.f. f(x).

The properties of p.d.f. f(x) are,

1. $f(x) \geq 0$ for all real number x

2. $\int_{-\infty}^{\infty} f(x)dx = 1.$

Definition 3.5. The function $F(x) = P_r\{X \leq x\} = \int_{-\infty}^{x} f(u)du$ is called a cumulative distribution function (c.d.f.) of X.

We note that $P_r\{a < X \leq b\} = \int_a^b f(x)dx = F(b) - F(a)$ for any real numbers a and b with a<b.

The relations between p.d.f. f(x) and c.d.f. F(x) are

$$F(x) = \int_{-\infty}^{x} f(u)du \quad \text{and} \quad \frac{dF(x)}{dx} = f(x).$$

For a continuous random variable X, $P_r\{X=a\} = 0$ for any real number a.

Just like discrete random variables, the expected values for continuous random variable are also important concepts.

Definition 3.6. Let a random variable X have p.d.f. f(x).

(a) The expected value (mean) of the random variable X is defined by

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx,$$

(b) The variance of the random variable X is defined by

$$\sigma^2 = Var(X) = E(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx, \text{ and}$$

(c) In general, for a function g(x), the exception of g(X) is defined by

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

It is noted that mean $\mu=E(X)$ describes the "central tendency" of the (population) probability distribution f(x) and standard deviation σ indicates the "dispersion" of the probability distribution f(x).

Example 3.5. Let random variable X have

$$\text{p.d.f.} \quad f(x) = e^{-x}, \text{ for } x \geq 0.$$

This is a probability density function since

$$f(x) = e^{-x} \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} f(x)dx = \int_0^{\infty} e^{-x}dx = 1.$$

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^{\infty} xe^{-x}dx = 1,$$

$$Var(X) = E(X^2) - \mu^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - 1^2$$

$$= \int_0^{\infty} x^2 e^{-x}dx - 1 = 2 - 1 = 1,$$

$$F(x) = \int_{-\infty}^{x} f(u)du = \int_{0}^{x} e^{-u}du = 1 - e^{-x}, \quad x \geq 0.$$

Comparison of Discrete and Continuous Random Variables

	Discrete	Continuous
range	$x_1, x_2, \ldots, x_n, \ldots$	real line $(-\infty, \infty)$
probability distribution	p.m.f. p_i	p.d.f. $f(x)$
mean μ	$\Sigma\, x_i p_i$	$\int xf(x)dx$
variance σ^2	$\Sigma\, x_i^2 p_i - \mu^2$	$\int x^2 f(x)dx - \mu^2$

4. VARIOUS CONTINUOUS DISTRIBUTIONS

The statistical model of a random phenomenon is formulated in terms of either p.d.f. $f(x)$ or c.d.f. $F(x)$.

Random phenomena exhibit a great variety of forms of specific relative frequency patterns. A statistical model of such a pattern should be flexible enough to accommodate a variety of relative frequency patterns. To achieve such flexibility, a specific mathematic function $f(x)$ or $F(x)$ is generalized by introducing into it a suitable "distribution parameter", denoted by $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$. The resulting generalized statistical model is called a family of density functions or distribution functions.

Thus, a general statistical model is $f(x;\underset{\sim}{\theta})$ or $F(x,\underset{\sim}{\theta})$, where the symbol $\underset{\sim}{\theta}$ represents the parameters which index that model.

The central aim of statistical analysis can now be restated to construct a statistical model $f(x,\underset{\sim}{\theta})$ or $F(x,\underset{\sim}{\theta})$ of the random variable X on the basis of the incomplete information contained in a sample of measurements on X.

We introduce here the families of distribution functions

(A) Exponential Distributions
(B) Normal Distributions
(C) Weibull Distributions
(D) Chi-Square Distributions,

which may be suitable for use in composite engineering applications.

(A) Exponential Distributions

A random variable X is said to have an <u>exponential distribution</u> if its p.d.f. is of the form

$$f(x;\theta) = \theta e^{-\theta x}, \quad x \geq 0. \quad (\theta \geq 0),$$

where θ is a parameter.

The c.d.f. of X is,

$$F(x) = \int_{-\infty}^{x} f(u)du = \int_{0}^{x} \theta e^{-\theta u}du = 1 - e^{-\theta x}, \quad x \geq 0.$$

The expected mean value and variance of X are;

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{\infty} x\theta e^{-\theta x}dx = \frac{1}{\theta} ,$$

$$E(X^2) = \int_{0}^{\infty} x^2\theta e^{-\theta x}dx = \frac{2}{\theta^2} , \text{ and}$$

$$\sigma^2 = Var(X) = E(X^2) - \mu^2 = \frac{2}{\theta^2} - \left(\frac{1}{\theta}\right)^2 = \frac{1}{\theta^2} .$$

Most applications of the exponential distribution are based on its "memoryless property", when the measurement random variable X has a time dimension. This property refers to the feature of a phenomenon in which the history of past events does not influence the probability of occurrence of present or future events.

(B) Normal Distributions

The normal family of probability density functions is defined as

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 ,$$

$$-\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$$

where μ is a location parameter and σ is a scale parameter.

We can denote this distribution by $X \overset{d}{\sim} N(\mu,\sigma)$ if a random variable has the above p.d.f..

Theorem 4.1. If $X \overset{d}{\sim} N(\mu,\sigma)$, then $E(X) = \mu$ and $s.d.(X) = \sigma$.

This theorem implies that the parameters μ and σ are precisely the mean and standard deviation of the random variable X respectively.

If a random variable Z has the normal distribution with $\mu=0$ and $\sigma=1$, i.e. $Z \overset{d}{\sim} N(0,1)$, then Z is said to have the standard normal distribution. The p.d.f. of Z is

$$f(z;0,1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) , \quad -\infty < z < \infty$$

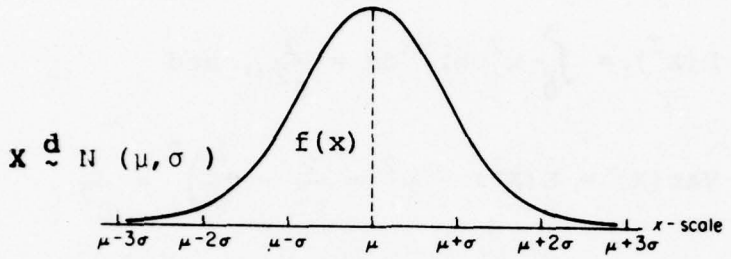


Figure 4.1. Normal Density Function

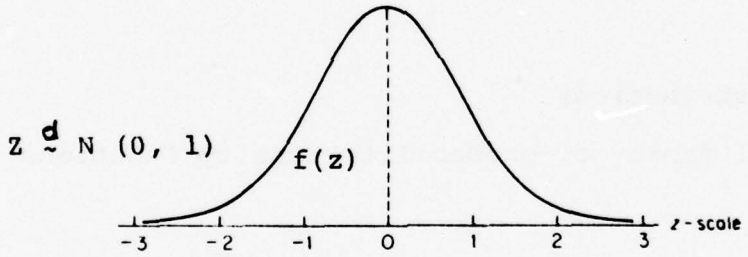


Figure 4.2. Standard Normal Density Function

Both pictures of the probability densities of X and Z are identical except for the scales on the horizontal lines. They are bell-shaped curves and symmetric about the means. The following theorem is useful in computing the probabilities of random variables $X \overset{d}{\sim} N(\mu,\sigma)$.

Theorem 4.2. If $X \overset{d}{\sim} N(\mu,\sigma)$, then

$$Z = \frac{X-\mu}{\sigma} \overset{d}{\sim} N(0,1) .$$

The values for the c.d.f. of the standard normal distribution is given in Table I.

TABLE I. NORMAL DISTRIBUTION

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp - \frac{t^2}{2}\, dt$$

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

x	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.891	4.417
F(x)	.90	.95	.975	.99	.995	.999	.9995	.99995	.999995
2[1 - F(x)]	.20	.10	.05	.02	.01	.002	.001	.0001	.00001

from "Introduction to the Theory of Statistics" by A.M. Mood, McGraw-Hill, New York, 1950.

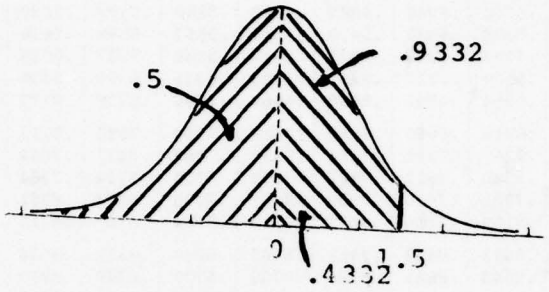
Example 4.1. Let $Z \overset{d}{\sim} N(0,1)$, then

(a) $P_r\{0<Z\leq 1.5\} = P_r\{Z\leq 1.5\} - P_r\{Z\leq 0\}$

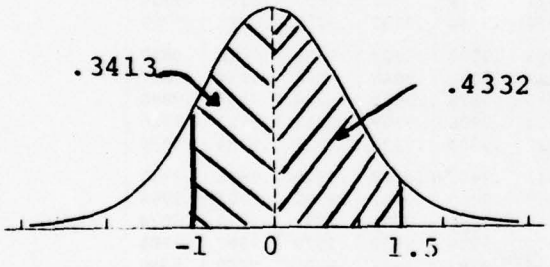
$= F(1.5) - F(0)$, where $F(z)$ is c.d.f. of Z

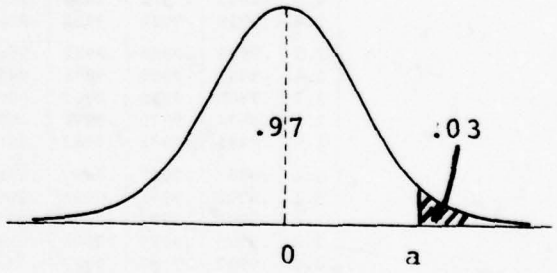
$= .9332 - .5$ (from Table I)

$= .4332$



(b) $P_r\{-1\leq Z\leq 1.5\} = .4332 + .3413 = .7745$

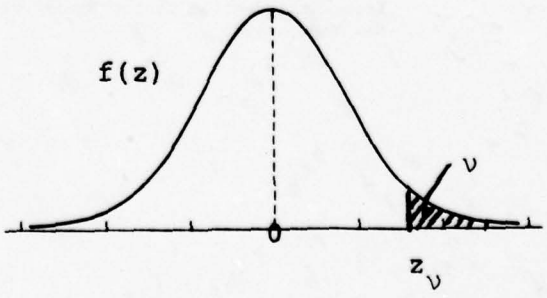




(c) $P_r\{Z>a\} = .03$ $\quad a = ??$ $\quad (a = 1.88)$

Definition 4.1. A symbol z_ν is defined to designate a number satisfying $P_r\{Z\geq z_\nu\} = \nu$ $(0<\nu<1)$,

(see the following example).



Example 4.2. (from Table I)

ν	z_ν
.05	1.645
.025	1.960
.01	2.326

Example 4.3. If $X \overset{d}{\sim} N(70,5)$, then

$$P_r\{X \geq 60\} = P_r\left\{\frac{X-70}{5} \geq \frac{60-70}{5}\right\}$$

$$= P_r\{Z \geq -2\} = P_r\{Z \leq 2\} = .9772.$$

The normal distribution is undoubtedly the best-known statistical measurement model. This distribution has wide applicability because of the central limit theorem, which will be presented in a later section.

(C) Weibull Distributions

The family of Weibull probability density functions (two parameter family) is defined as

$$f(x;\beta,\alpha) = \left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left\{-\left(\frac{x}{\beta}\right)^{\alpha}\right\},$$

$$x,\beta,\alpha > 0,$$

where β is a scale parameter and α is a shape parameter. The Weibull c.d.f. is found directly as

$$F(x;\beta,\alpha) = \int_0^x f(x';\beta,\alpha)dx' = 1-\exp\left\{-\left(\frac{x}{\beta}\right)^{\alpha}\right\},$$

$$x > 0.$$

If a random variable X has the above Weibull distribution, then we will write $X \overset{d}{\sim} W(\beta,\alpha)$.

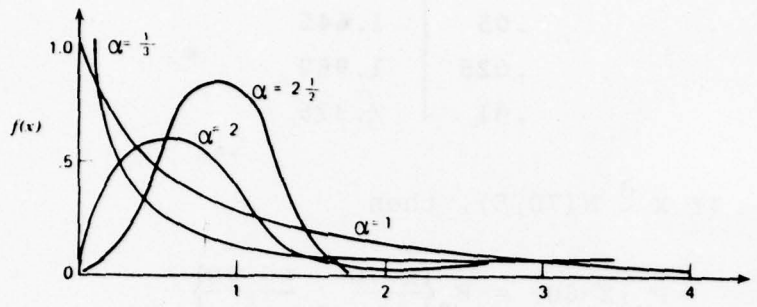


Figure 4.3. Weibull Density Function

Theorem 4.3. If $X \overset{d}{\sim} W(\beta,\alpha)$, then

$$E(X) = \mu = \beta\Gamma\left(1+\frac{1}{\alpha}\right),$$

$$\text{Var}(X) = \sigma^2 = \beta^2\left\{\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right\}, \text{ and}$$

$$\frac{\sigma}{\mu} = \left\{\frac{\Gamma\left(1+\frac{2}{\alpha}\right)}{\Gamma^2\left(1+\frac{1}{\alpha}\right)} - 1\right\}^{1/2},$$

where Γ denotes the Gamma function $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ ($\frac{\sigma}{\mu}$ is called the coefficient of dispersion of X).

Example 4.4. Let $X \overset{d}{\sim} W(\beta,\alpha)$ and

$$Y = aX + b$$

where a>0, b are constants, then the c.d.f. of Y can be computed;

$$P_r\{Y \leq y\} = P_r\{aX+b \leq y\}$$

$$= P_r\left\{X \leq \frac{y-b}{a}\right\}$$

$$= 1 - \exp\left\{-\left[\frac{\frac{y-b}{a}}{\beta}\right]^{\alpha}\right\}$$

$$= 1 - \exp\left\{-\left[\frac{y-b}{a\beta}\right]^{\alpha}\right\}.$$

The Weibull model can be used to represent a wide variety of engineering random phenomena and also the Weibull distribution is used to model the breaking strength and fatigue life of metals and composite materials.

(D) Chi-Square Distributions

A random variable X has a Chi-square distribution with n degree of freedoms (we write $X \overset{d}{\sim} \chi^2(n)$) if its p.d.f. is

$$f(x;n) = \frac{1}{2\Gamma\left(\frac{n}{2}\right)} \left(\frac{x}{2}\right)^{\frac{n}{2}-1} e^{-x/2}, \quad x > 0.$$

(here n is a parameter).

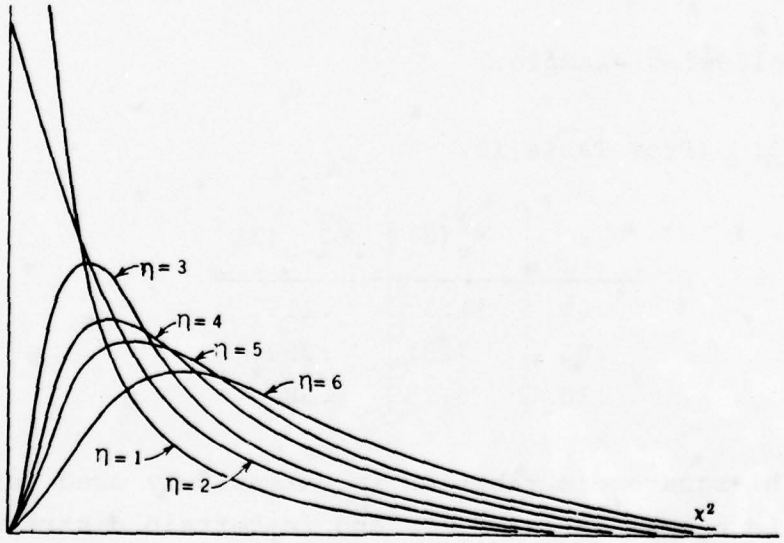


Figure 4.4. Chi-Square Density Function

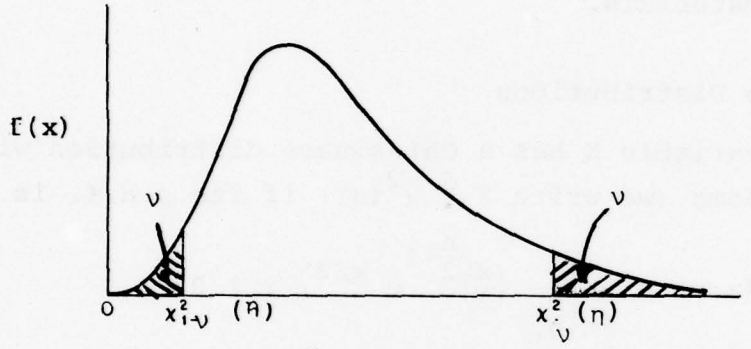
Theorem 4.4. If $X \overset{d}{\sim} \chi^2(n)$, then

$$E(X) = \mu = n \quad \text{and} \quad Var(X) = \sigma^2 = 2n.$$

Definition 4.2. Symbols $\chi^2_{\nu}(n)$ and $\chi^2_{1-\nu}(n)$ are defined to designate the numbers satisfying,

$$P_r\{X \geq \chi^2_{\nu}(n)\} = \nu \quad \text{and} \quad P_r\{X \geq \chi^2_{1-\nu}(n)\} = 1-\nu,$$

where $X \overset{d}{\sim} \chi^2(n)$.



χ^2 - density function

(See the following example.)

Example 4.5. (from Table II)

ν	$\chi^2_{\nu}(3)$	$\chi^2_{1-\nu}(3)$
.01	11.3	.115
.05	7.81	.352
.10	6.25	.584

The Chi-square distribution is extensively used for inferences on the normal variance σ^2, and in certain distribution tests.

The values for the Chi-square distribution are given in Table II.

TABLE II. CHI-SQUARE DISTRIBUTION

$$F(u) = \int_0^u \frac{x^{(n-2)/2} e^{-x/2}\, dx}{2^{n/2}\, \Gamma(n/2)}$$

n \ F	.005	.010	.025	.050	.100	.250	.500	.750	.900	.950	.975	.990	.995
1	.0393	.0157	.0982	.0393	.0158	.102	.455	1.32	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.103	.211	.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	.0717	.115	.216	.352	.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	.207	.297	.484	.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	.412	.554	.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	.676	.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7

from "Biometrika", Vol. 32 (1941).

Example 4.6. (Hahn-Kim and Yang's residual strength degradation models for the composite material fatigue failures).

Let S = applied load at each fatigue cycle,

R(n) = residual strength at nth fatigue cycle (random variables)

R(0) = the static strength (random variable)

N = the fatigue life strength in cycle (random variable).

Assumptions;

(a) $R(0) \overset{d}{\sim} W(\beta,\alpha)$

(α and β can be estimated from the static strength data),

(b) The degradation rule;

$$R^c(n) = R(0)^c - \beta^c K S^b n$$

(c, K and b can be estimated from the residual strength data).

The c.d.f. of the residual strength R(n) is;

$$F_{R(n)}(x) = P_r\{R(n) \leq x\} = P_r\{R^c(n) \leq x^c\}$$

$$= P_r\{R^c(0) - \beta^c K S^b n \leq x^c\} \text{ (from the assumption (b))}$$

$$= P_r\{R^c(0) \leq x^c + \beta^c K S^b n\}$$

$$= P_r\{R(0) \leq (x^c + \beta^c K S^b n)^{1/c}\}$$

$$= 1 - \exp - \left[\frac{(x^c + \beta^c K S^b n)^{1/c}}{\beta}\right]^{\alpha}$$

(from the assumption (a) that $R(0) \overset{d}{\sim} W(\beta,\alpha)$)

$$(T) = 1 - \exp\left\{-\left[\frac{x^c+\beta^c KS^b n}{\beta^c}\right]^{\alpha/c}\right\}.$$

The c.d.f. of the fatigue life N is;

(note that $N \geq n$ if and only if $R(n) \geq \sigma_{max} = S$)

$$F_N(n) = P_r\{N \leq n\} = P_r\{R(n) \leq S\}$$

$$= 1 - \exp\left\{-\left[\frac{(S^c+\beta^c KS^b n)}{\beta^c}\right]^{\alpha/c}\right\} \quad \text{(from (T))}$$

$$= 1 - \exp\left\{-\left[\frac{n+(S^c/K\beta^c S^b)}{1/KS^b}\right]^{\alpha/c}\right\}.$$

Note that the fatigue life strength N has a three parameter Weibull distribution.

5. JOINT DISTRIBUTIONS (CONTINUOUS RANDOM VARIABLE)

The statistical model $F(x,y;\theta)$, which describes the joint behavior of two random variables X and Y, is called joint cumulative distribution function or joint c.d.f.. Its probability interpretation is

$$F(x,y;\theta) = P_r\{X \leq x \text{ and } Y \leq y | \theta\}.$$

For two continuous random variables X and Y, the joint p.d.f. is

$$f(x,y;\theta) = \frac{\partial^2 F(x,y;\theta)}{\partial x \partial y}$$

so that

$$F(x,y;\theta) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v;\theta)\,dudv.$$

The joint p.d.f. $f(x,y;\theta)$ satisfies

(a) $f(x,y;\theta) \geq 0$ for any real x and y

(b) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y;\theta)\,dxdy = 1.$

From a joint p.d.f. $f(x,y)$ of random variable X and Y, one can obtain p.d.f. of X (or Y) alone (call marginal p.d.f. of X (or Y)) by integrating out as follows;

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)\,dx.$$

Example 5.1. Let the joint p.d.f. of two exponential variables be

$$f(x,y;\theta_1,\theta_2) = \theta_1\theta_2 \exp\{-\theta_1 x - \theta_2 y\}.$$

The marginal p.d.f. of X is obtained by integrating out variable y;

$$f_1(x;\theta_1) = \int_0^\infty f(x,y;\theta_1,\theta_2)dy = \theta_1 \exp\{-\theta_1 x\}.$$

Definition 5.1. The conditional probability density function of X given Y=y is defined as

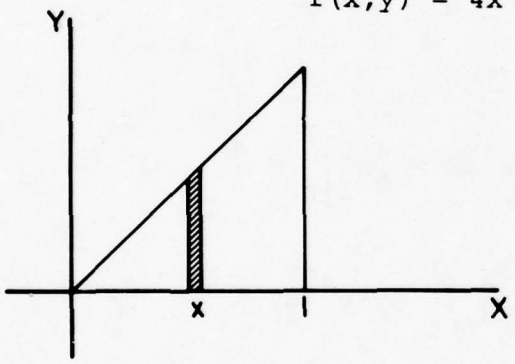
$$f(x|y) = \frac{f(x,y)}{f_2(y)}$$

where $f(x,y)$ is the joint p.d.f. of X and Y and $f_2(y)$ is the marginal p.d.f. of Y, provided that $f(y)>0$.

The concept of conditional p.d.f. is introduced to predict the probabilistic behavior of the first random variable in terms of a specific value of the second random variable.

Example 5.2. Let the joint p.d.f. of (X,Y) be

$$f(x,y) = 4x^2, \quad 0<x<1, \quad 0<y<x.$$



$$f_1(x) = \int_0^x 4x^2 \cdot dy$$

$$= 4x^2 \cdot y\Big|_0^x = 4x^3, \quad 0<x<1$$

$$f_2(y) = \int_y^1 4x^2 dx = \frac{4}{3}x^3\Big|_y^1 = \frac{4}{3}(1-y^3), \quad 0<y<1.$$

$$f(x|y) = \frac{f(x,y)}{f_2(y)} = \frac{4x^2}{\frac{4}{3}(1-y^3)}, \quad 0<y<x<1,$$

$$f(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{4x^2}{4x^3} = \frac{1}{x}, \quad 0<y<x<1.$$

The conditional p.d.f. $f(x|y)$ contains information on the dependence of X on Y. The covariance is a measure of the dispersion of both X and Y, just as the variance measures the dispersion of a single random variable.

Definition 5.2. Let random variable (X,Y) have joint p.d.f. $f(x,y)$. The covariance of X and Y, denoted by $\text{cov}(X,Y)$, is defined by,

$$\text{cov}(X,Y) = E[(X-EX)(Y-EY)] = E(XY) - (EX)(EY)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y)\, f(x,y)\,dxdy \ .$$

The correlation coefficient of X and Y, denoted by $\rho(X,Y)$, is defined by

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\text{s.d.}(X)\ \text{s.d.}(Y)} \ .$$

Example 5.3. $f(x,y) = 4x^2$, $0<x<1$, $0<y<x$.

$$\mu_X = E(X) = \int_0^1 x\cdot 4x^3 dx = \int_0^1 4x^4 dx = \frac{4}{5}x^5\Big|_0^1 = \frac{4}{5}$$

$$\mu_Y = E(Y) = \int_0^1 y\frac{4}{3}(1-y^3)dy = \frac{4}{3}\int_0^1 (y-y^4)dy$$

$$= \frac{4}{3}\left(\frac{y^2}{2} - \frac{y^5}{5}\right)\Big|_0^1 = \frac{4}{3}\cdot\left(\frac{1}{2} - \frac{1}{5}\right) = \frac{2}{5}$$

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy = \int_0^1\int_0^x x\cdot y4x^2 dy\cdot dx$$

$$= \int_0^1 4x^3\cdot\frac{y^2}{2}\Big|_0^x dx = \int_0^1 2x^5 dx = \frac{2}{6}\,x^6\Big|_0^1 = \frac{1}{3}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_1(x)dx = \int_0^1 x^2 4x^3 \cdot dx = \int_0^1 4x^5 dx = \frac{4}{6} = \frac{2}{3}$$

$$E(Y^2) = \int_{-\infty}^{\infty} y^2 f_2(y)dy = \int_0^1 y^2 \frac{4}{3}(1-y^3)dy$$

$$= \frac{4}{3}\int_0^1 (y^2-y^5)dy = \left(\frac{4}{3}\right)\left(\frac{1}{3} - \frac{1}{6}\right) = \frac{2}{9}$$

$$Var(X) = E(X^2) - (EX)^2 = \frac{2}{3} - \left(\frac{4}{5}\right)^2 = \frac{2}{75}$$

$$Var(Y) = E(Y^2) - (EY)^2 = \frac{2}{9} - \left(\frac{2}{5}\right)^2 = \frac{14}{225}$$

$$cov(X,Y) = E(XY) - (EX)(EY) = \frac{1}{3} - \frac{4}{5} \cdot \frac{2}{5} = \frac{1}{75} .$$

Suppose that simultaneous measurements on two random variables are taken. These random variables are independent (intuitively) if any information on one random variable provides no information on the other random variable. In this case, the conditional p.d.f. of one random variable, given any value of an independent random variable, equals the marginal p.d.f. of the first random variable. That is,

$$f(x|y;\theta) = f(x;\theta).$$

But, from the definition of the conditional probability (Definition 5.1),

$$f(x|y;\theta) = f(x,y;\theta)/f(y;\theta),$$

thus the independence of X and Y means that

$$f(x,y;\theta) = f(x;\theta)\ f(y;\theta),$$

for any real x and y.

Definition 5.3. Two random variables X and Y are said to be "independent" if their joint p.d.f. is equal to the product of the p.d.f.'s of X and Y, that is,

$$f(x,y;\theta) = f(x;\theta) \cdot f(y;\theta).$$

Which of the following pairs of measurements (events) are independent?

1. number of years of education vs. amount of yearly income
2. having green eyes vs. being a banker
3. breaking strength of composite vs. its fatigue life
4. number appeared in 1st toss of a die vs. number appeared in 2nd toss of a die.

(Answer; (2) and (4)).

Example 5.4. Let random variables X and Y be

X = number appeared in 1st toss of a die
Y = number appeared in 2nd toss of a die

$$P_r\{X=3,\ Y=4\} = \frac{1}{36}$$

$$P_r\{X=3\} = \frac{1}{6},\ P_r\{Y=4\} = \frac{1}{6}$$

$$\therefore P_r\{X=3,\ Y=4\} = \frac{1}{36} = P_r\{X=3\}\cdot P_r\{Y=4\}.$$

Also $$P_r\{X=i,\ Y=j\} = P_r\{X=i\}\cdot P_r\{Y=j\}$$

for i and j = 1,2,...,6.

Therefore, X and Y are independent.

We note that if X and Y are independent with p.d.f. $f_X(x)$ and $f_Y(y)$, then their joint p.d.f. $f(x,y)$ can be obtained by

$$f(x,y) = f_X(x) \cdot f_Y(y).$$

Example 5.5. We introduce here a concept of failure rate, which is one of the basic concepts in reliability, in terms of definition of the conditional probability.

Let T be a failure time, random variable, of a device with c.d.f. $F(t)$ and p.d.f. $f(t)$.

The failure rate $h(t)$ at time t is;

$$h(t)dt = P_r\{\text{instantaneous failure of the device during the time interval } (t,t+dt) \mid \text{it survived up to } t\}$$

$$= \frac{f(t)dt}{1-F(t)}$$

$$\therefore \quad h(t) = \frac{f(t)}{1-F(t)}$$

We note that if a failure time T has a Weibull $W(\beta,\alpha)$ distribution, then its c.d.f. $F(t)$ is $F(t) = P_r\{T \le t\} = 1 - \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\}$ and its p.d.f. ($f(t)$ is $f(t) = \frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\}$, hence its failure rate $h(t)$ is given by

$$h(t) = \frac{f(t)}{1-F(t)} = \frac{\alpha}{\beta} \cdot \left(\frac{t}{\beta}\right)^{\alpha-1} = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-1}$$

6. SAMPLING AND ESTIMATIONS

In the preceding discussions a random variable X represents a measurable characteristic of some underlying phenomenon and its p.d.f. $f(x;\theta)$ is the probability distribution of X. Realizations of that random variable X gives rise to quantitative information on that characteristic.

Suppose that we have taken, from an experiment, a sample of n measurements $x_1, x_2, \ldots, x_n$ on a random variable X with the p.d.f. $f(x;\theta)$. These measurements $x_1, x_2, \ldots, x_n$ are considered as a realization of random variables $X_1, X_2, \ldots, X_n$ (random sample), which we are going to define.

Definition 6.1. A collection of random variable $X_1, X_2, \ldots, X_n$ is called a random sample (of size n) from a random variable X (or from a p.d.f. $f(x;\theta)$) if

(a) each random variable X_i has p.d.f. $f(x;\theta)$

(b) $X_1, X_2, \ldots, X_n$ are independent random variables.

We note that each sample X_i is a random variable until its measurement x_i is taken and its p.d.f. is $f(x_i;\theta)$ and the word "random" in the random sample means each sample value X_i is not influenced in any way by any other sample value X_j $(j \neq i)$.

It follows clearly that the joint p.d.f. of a random sample $X_1, X_2, \ldots, X_n$ is

$$f(x_1,x_2,\ldots,x_n:\theta) = f(x_1;\theta)\, f(x_2;\theta) \ldots f(x_n;\theta)$$

$$= \prod_{i=1}^{n} f(x_i;\theta).$$

If we let $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ and consider $L(\theta)$ as a function of the parameter θ, then it is called the likelihood function of a sample $X_1, X_2, \ldots, X_n$.

Suppose that random measurement characteristic X has a p.d.f. $f(x;\theta)$ and a family of distributions that $f(x;\theta)$ belongs to is a

known family (say Normal or Weibull family). Only unknown criteria about $f(x;\theta)$ is the parameter θ, whose value is required to be estimated in a statistical inference problem.

A function $\hat{\theta}_n = \hat{\theta}(X_1, X_2, \ldots, X_n)$ of a random sample $X_1, X_2, \ldots, X_n$, which approximates the unknown value of the parameter θ in a certain sense, is called an "estimator" of θ. As a function of random variables, $\hat{\theta}_n$ is a random variable and hence it has a distribution function. A realization $\hat{\theta}_n = \hat{\theta}(x_1, x_2, \ldots, x_n)$ of $\hat{\theta}_n = \hat{\theta}(X_1, X_2, \ldots, X_n)$ is called an "estimate" of θ (an estimate of θ is a number).

The following criterias can be applied to characterize good estimators;

1. Unbiased if $E[\hat{\theta}_n] = \theta$ for each n,
2. Consistent if $P_r\{|\hat{\theta}_n - \theta| \geq \varepsilon\} \leq \varepsilon$ for any real number ε and sufficiently large n,
3. Minimum variance if $E(\hat{\theta}_n - \theta)^2 \leq E(\theta^* - \theta)^2$ for any other estimator θ^* of θ with $E(\theta^*) = \theta$,
4. Maximum likelihood (M.L.E.) if $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ achieves maximum at $\theta = \hat{\theta}_n$.

The usual procedure to obtain M.L.E. of θ is to take the derivative $\frac{d}{d\theta} L(\theta)$ and solve for $\frac{d}{d\theta} L(\theta) = 0$.

For various families of distributions, the sample mean, $\bar{X}$, and sample standard deviation, S, are good estimates for the population mean μ and population standard deviation σ respectively.

Example 6.1. Let a random variable X have the exponential distribution with parameter θ.

$$f(x;\theta) = \theta e^{-\theta x}, \quad x \geq 0.$$

The likelihood function $L(\theta)$ of $X_1, X_2, \ldots, X_n$ is

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}.$$

$$\frac{dL(\theta)}{d\theta} = n\theta^{n-1} e^{-\theta\Sigma x_i} - \theta^n \Sigma x_i e^{-\theta\Sigma x_i}$$

$$= \theta^{n-1} e^{-\theta\Sigma x_i} (n-\theta\Sigma x_i) = 0$$

if and only if $n - \theta \sum_{i=1}^{n} x_i = 0$.

i.e. $\theta = \dfrac{n}{\sum_{i=1}^{n} X_i}$.

Therefore, M.L.E. of θ is

$$\hat{\theta}_n = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}_n}, \text{ where } \bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Example 6.2. Let $X \overset{d}{\sim} N(\mu,\sigma)$, where the value of σ is known.

$$L(\mu) = \prod_{i=1}^{n} f(x_i;\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{- \frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i-\mu}{\sigma}\right)^2\right\}.$$

$$\frac{dL(\mu)}{d\mu} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{- \frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i-\mu}{\sigma}\right)^2\right\} \cdot \frac{2}{\sigma} \sum_{i=1}^{n} (x_i-\mu).$$

Hence $\dfrac{dL(\mu)}{d\mu} = 0$ if and only if $\sum_{i=1}^{n} (x_i-\mu) = 0$ if and only if

$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n$. Therefore, $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n$ is M.L.E. of μ

It can be easily shown that $\hat{\mu}_n$ is also unbiased, consistent and minimum variance estimator of μ. The fact that the distribution of the estimator $\hat{\mu}_n = \bar{X}_n$ is $N(\mu,\sigma/\sqrt{n})$ follows from the following theorem.

Theorem 6.1. If $X \overset{d}{\sim} N(\mu,\sigma)$, then $\bar{X}_n \overset{d}{\sim} N(\mu,\sigma/\sqrt{n})$.

In fact, when sample size n is sufficiently large ($n \geq 30$), the sample mean $\bar{X}_n$ has always an approximately normal distribution regardless of the shape of the population distribution.

Theorem 6.2. (Central limit theorem) Let a random variable X have a finite mean μ and standard deviation σ. Then the sample mean $\bar{X}_n$ has an approximately normal distribution with the mean μ and the standard deviation $\sigma/\sqrt{n}$ for sufficiently large n regardless of the p.d.f. of X.

The following Figure 6.1 illustrates the fact that the distribution of $\bar{X}_n$ approaches the normal probability density function as n increases when the population distribution are either an exponential or uniform distribution.

The central limit theorem is very useful in statistical inferences because the unknown population mean μ is usually estimated by the sample mean $\bar{X}_n$ and the distribution of $\bar{X}_n$ is approximated by a normal distribution (for sufficiently n) according to the central limit theorem.

Linear Regression - Data Analysis

For two random variables X and Y, a linear relationship $Y = aX + b$, where a and b are some real constants, can be investigated by linear regression analysis.

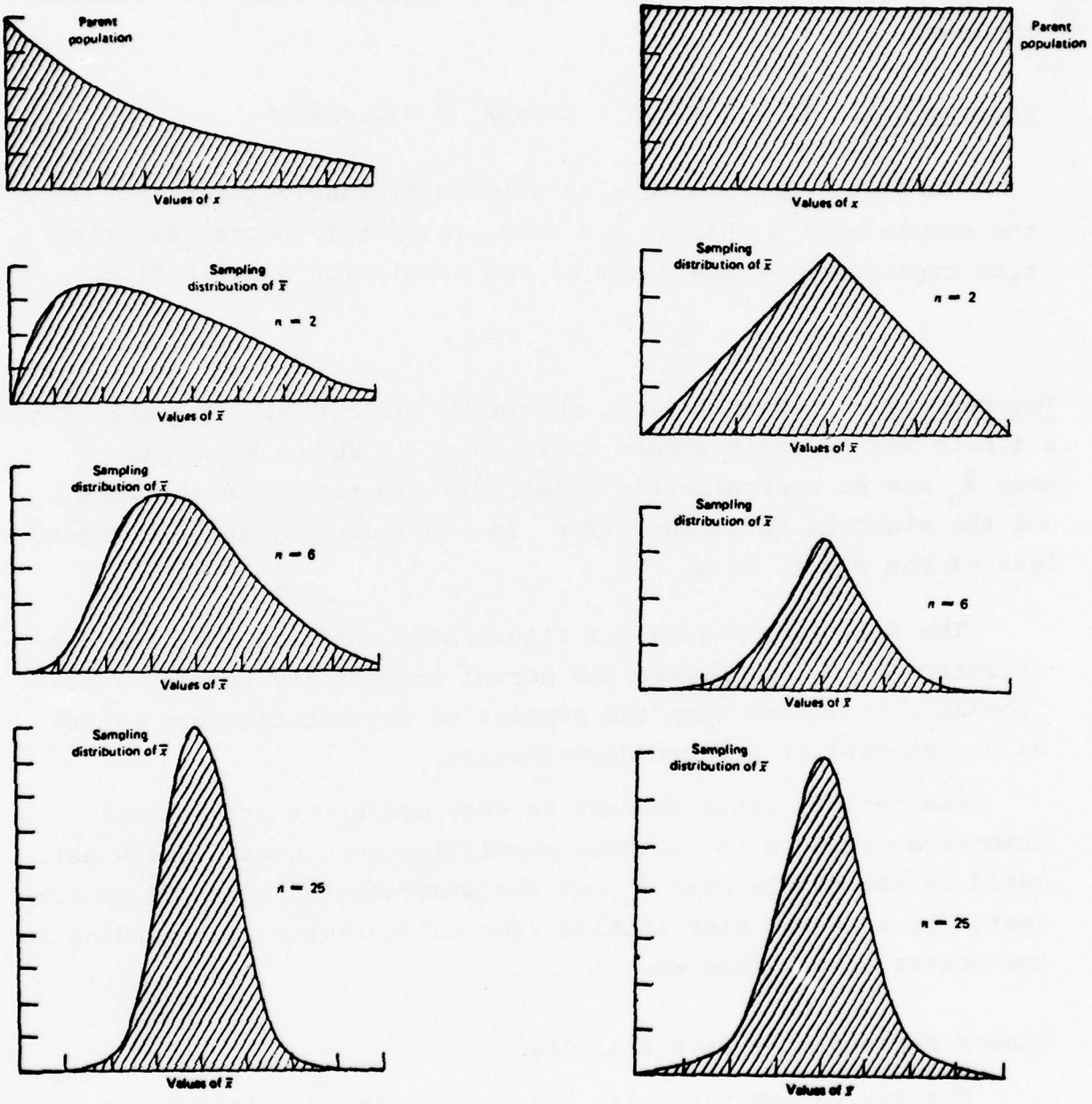
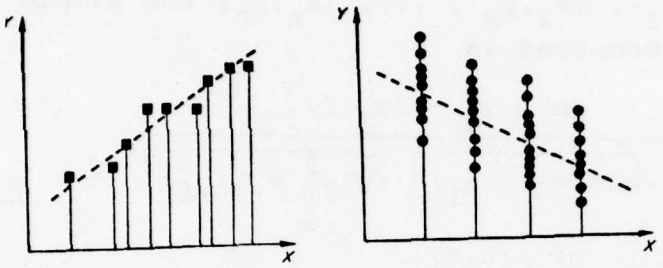


Figure 6.1. Distributions of $\bar{X}_n$.

Suppose that we have a random sample of size n, (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) taken from the pair (X,Y). It is desired to obtain a straight line $y = ax+b$ which fits the data best. In other words, the constants a and b in $y = ax+b$ should be estimated so that the mean square error $Q = \sum_{i=1}^{n} [y_i - (ax_i-b)]^2$ is a minimum.



Theorem 6.3. Based on the data (x_1,y_1), (x_2,y_2), ..., (x_n,y_n), the solution for the linear regression is

$$a = \frac{n\Sigma x_i y_i - \Sigma x_i \cdot \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}, \text{ and}$$

$$b = \frac{\Sigma y_i - a\Sigma x_i}{n}.$$

Proof: Let $Q = \sum_{i=1}^{n} [y_i - (ax_i+b)]^2$.

$\frac{\partial Q}{\partial a} = 0$ and $\frac{\partial Q}{\partial b} = 0$ reduces to two equations with two unknowns;

$$\begin{cases} \Sigma x_i y_i = a\Sigma x_i^2 + b\Sigma x_i \\ \Sigma y_i = a\Sigma x_i + nb \end{cases}$$

and a and b are the solution for the above equations.

The applications of this linear regression technique in composite material engineering are

(1) Estimation on Weibull parameters (presented in next section)

(2) S-N curve fitting (see Example 6.4.)

The (population) correlation coefficient $\rho(X,Y)$ between two random variables X and Y has been defined in Section 5. Based on the data (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) the sample correlation coefficient r is computed as

$$r = \frac{n\Sigma x_i y_i - \Sigma x_i \cdot \Sigma y_i}{\sqrt{n\Sigma x_i^2 - (\Sigma x_i)^2}\ \sqrt{n\Sigma y_i^2 - (\Sigma y_i)^2}},$$

which is an estimate of $\rho(X,Y)$.

Example 6.3. Let X and Y denote one's age and blood pressure respectively. Age and blood pressure data for five persons is

X	15	27	20	45	40
Y	110	115	113	132	135

	x	y	x^2	xy
	15	110	225	1650
	27	115	729	3105
	20	113	400	2265
	45	132	2025	5940
	40	135	1600	5400
Total	147	605	4975	18355
n=5	Σx_i	Σy_i	Σx_i^2	$\Sigma x_i y_i$

$$a = \frac{n\Sigma x_i y_i - \Sigma x_i \cdot \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{5\text{x}18355 - 147\text{x}605}{5\text{x}4975 - (147)^2}$$

$$= \frac{2840}{3286} = .86$$

$$b = \frac{\Sigma y_i - \Sigma a x_i}{n} = \frac{605 - .86 \times 147}{5} = 95.72$$

$$\therefore \quad y = ax + b = .86x + 95.72$$

Example 6.4. The fatigue life N of a composite material is observed under applied stress levels S = 45, 55 and 65 KSI and the data obtained is as follows.

Data C.

S	45	55	65
N in Cycle	11,220,000	70,800	2,630
	15,136,000	112,200	6,610
		182,000	7,410
		338,000	9,330

Since our form of fatigue model is $N = BS^a$, by taking the LOG of both sides, we obtain

$$\text{LOG } N = a \text{ LOG } S + b,$$

where b = LOG B.

	x = LOG S	y = LOG N	x^2	xy
n=10	1.65	7.05	2.7225	11.6325
	1.65	7.18	2.7225	11.8470
	1.74	4.85	3.0276	8.4390
	1.74	5.05	3.0276	8.7870
	1.74	5.26	3.0276	9.1524
	1.74	5.53	3.0276	9.6222
	1.81	3.42	3.2761	6.1902
	1.81	3.82	3.2761	6.9142
	1.81	3.87	3.2761	7.0047
	1.81	3.98	3.2761	7.2038
Total	17.50	50.01	30.6598	86.7930

$$a = \frac{10 \times 86.793 - 17.5 \times 50.01}{10 \times 30.66 - (17.5)^2} = \frac{-7.245}{0.35}$$

$$= -20.7$$

$$b = \frac{50.01 - (-20.7) \times 17.5}{10} = 41.23$$

$$\therefore \quad \text{LOG } N = -20.7 \text{ LOG } S + 41.23.$$

7. ESTIMATIONS ON THE WEIBULL DISTRIBUTION

For static strength tests or fatigue life tests of various composite materials, the two-parameter Weibull distribution $W(\beta,\alpha)$ is widely used as their failure distributions.

We note again that if $X \overset{d}{\sim} W(\beta,\alpha)$, then its p.d.f. is expressed by

$$f(x;\beta,\alpha) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp - \left[\frac{x}{\beta}\right]^{\alpha}, \quad x \geq 0.$$

(A) M.L.E. (Maximum Likelihood Estimate)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the $W(\beta,\alpha)$ Weibull distribution. The likelihood function $L(\beta,\alpha)$ is

$$L(\beta,\alpha) = \prod_{i=1}^{n} f(x_i;\beta,\alpha)$$

$$= \left(\frac{\alpha}{\beta}\right)^{n} \prod_{i=1}^{n} \left(\frac{x_i}{\beta}\right)^{\alpha-1} \exp\left[-\frac{\Sigma(x_i{}^{\alpha})}{\beta^{\alpha}}\right].$$

The M.L.E.'s $\hat{\beta}$ and $\hat{\alpha}$ of β and α should be the solutions of the equations

$$\frac{\partial L(\beta,\alpha)}{\partial \beta} = 0 \quad \text{and} \quad \frac{\partial L(\beta,\alpha)}{\partial \alpha} = 0, \text{ i.e.}$$

$$\beta = \left[\frac{1}{n}\sum_{i=1}^{n} x_i^{\alpha}\right]^{1/\alpha}$$

$$\frac{1}{\alpha} - \frac{\sum_{i=1}^{n} x_i^{\alpha} \ln x_i}{\sum_{i=1}^{n} x_i^{\alpha}} + \frac{1}{n}\sum_{i=1}^{n} \ln x_i = 0 .$$

The M.L.E., $\hat{\alpha}$, of α can be obtained by an iterative procedure. If the value of α is known (or estimated), then the M.L.E., $\hat{\beta}_n$, of β is

$$\hat{\beta}_n = \left[\frac{1}{n}\sum_{i=1}^{n} X_i^{\alpha}\right]^{1/\alpha}.$$

The probability distribution of $\hat{\beta}_n$ is

$$2n\left[\frac{\hat{\beta}_n}{\beta}\right]^{\alpha} \overset{d}{\sim} \chi^2(2n).$$

(B) Graphical Plotting (Linear Regression)

Let a breaking strength (or fatigue life) characteristic X have a Weibull distribution with the parameters β and α. Its c.d.f. is

$$F(x) = 1 - \exp - \left(\frac{x}{\beta}\right)^{\alpha}, \quad x \geq 0.$$

The linear regression method can be applied to estimate the parameters β and α.

Taking the $\ell n \ell n$ transform of F(x), we have

$$\ell n \ell n \frac{1}{1-F(x)} = \alpha \ell n x - \alpha \ell n \beta.$$

Suppose that we have tested n specimens and obtained data x_1, x_2, ..., x_n from the $W(\beta,\alpha)$ distribution.

Let

$$y = \ell n \ell n \frac{1}{1-F(x)},$$

$$z = \ell n x,$$

$$b = -\alpha \ell n \beta ,$$

$$p_i = \frac{i}{1+n} , \quad i=1,2,\ldots,n,$$

$$y_i = \ell n \ell n \frac{1}{1-p_i}, \quad i=1,2,\ldots,n,$$

$$z_i = \ell n \, x_{(i)}, \quad i=1,2,\ldots,n,$$

where $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are the order statistics (the arrangement of $x_1, x_2, \ldots, x_n$ in increasing order) of $x_1, x_2, \ldots, x_n$.

Based on the observations $(z_1,y_1), (z_2,y_2), \ldots, (z_n,y_n)$, a linear regression line $y = \hat{\alpha}z+\hat{b}$ can be estimated and $\hat{\alpha}$ and $\hat{\beta} = \exp -[\hat{b}/\hat{\alpha}]$ are the estimates for α and β.

Data D. Tensile strength of AS/3501-5A/16 ply/0° in KSI (n=8)

152	196	200	207
214	220	226	244

Example 7.1. Assume that Data D came from a Weibull distribution $W(\beta,\alpha)$.

(a) Estimation of β and α by Graphical Plotting;

rank (i)	$x_{(i)}$	p_i	$\frac{1}{1-p_i}$	$z_i = \ell n \, x_{(i)}$	$y_i = \ell n \ell n \frac{1}{1-p_i}$	z_i^2	$z_i y_i$
1	152	1/9	9/8	5.02	-2.139	25.20	-10.783
2	196	2/9	9/7	5.28	-1.381	27.88	- 7.292
3	200	3/9	9/6	5.30	-1.903	28.09	- 4.786
4	207	4/9	9/5	5.33	- .531	28.41	- 2.830
5	214	5/9	9/4	5.37	- .210	28.84	- 1.128
6	220	6/9	9/3	5.39	.094	29.05	.507
7	226	7/9	9/2	5.42	.408	29.38	2.211
8	244	8/9	9/1	5.50	.787	30.25	4.329
		Total		42.61	-3.875	227.10	-19.727

$$\hat{\alpha} = \frac{8x(-19.727) - 42.61x(-3.875)}{8x227.10 - (42.61)^2} = \frac{7.30}{1.188} = 6.14$$

$$\hat{b} = \frac{-3.875 - 6.14x42.61}{8} = -33.19$$

$$\hat{\beta} = \exp -\left[\frac{\hat{b}}{\hat{\alpha}}\right] = 222.64$$

(b) Estimation of β by M.L.E.

Suppose that the value of β is known to be 6.

$$\hat{\beta} = \left[\frac{1}{n}\sum_{i=1}^{n} x_i^{\alpha}\right]^{1/\alpha}$$

$$= \left[\frac{1}{8}(152^6 + 196^6 + \ldots + 244^6)^{1/6}\right]$$

$$= 213.86$$

8. CONFIDENCE INTERVALS

(A) For Normal Distributions

We start this section by giving an example to motivate and understand the concept of confidence intervals.

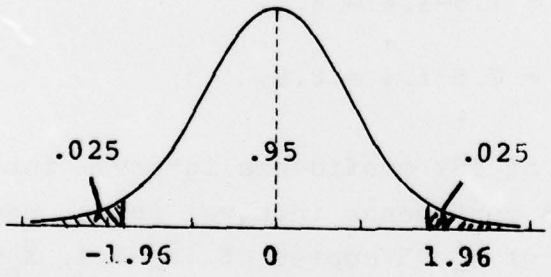
Example 8.1. Suppose that a measurement characteristic X has $N(\mu,\sigma)$ distribution, where μ is unknown and σ is a known value. On the basis of a random sample $X_1, X_2, \ldots, X_n$ from X, we know that the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimator of the unknown parameter μ and

$$\bar{X}_n \overset{d}{\sim} N(\mu,\sigma/\sqrt{n}), \quad \text{i.e.}$$

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{d}{\sim} N(0,1).$$

Hence $$P_r\left\{-1.96 \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq 1.96\right\} = .95 .$$

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



(Note that $z_{.025} = 1.960$ from Example 4.2.)

Now $$-1.96 \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq 1.96,$$

$$-1.96\,\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n - \mu \leq 1.96\,\frac{\sigma}{\sqrt{n}},$$

$$-\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}},$$

$$\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}},$$

$$\therefore P_r\left\{\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right\} = .95$$

It shows that, with the probability .95, the population mean μ is between $\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}$ and $\bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}$.

Suppose that $\sigma=5$, then

$$P_r\{\bar{X}_n - 1.4 \leq \mu \leq \bar{X}_n + 1.4\} = .95.$$

The interval $(\bar{X}_n-1.4, \bar{X}_n+1.4)$ is a 95% confidence interval for μ.

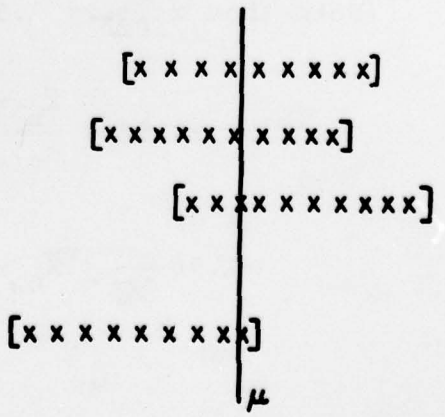
If we have taken a random sample $X_1, X_2, \ldots, X_n$ and obtained $\bar{X}_n=7.5$, for n=49, then

$$\bar{X}_n-1.4 = 7.5-1.4 = 6.1,$$

$$\bar{X}_n+1.4 = 7.5+1.4 = 8.9,$$

(6.1,8.9) is one estimate of 95% confidence interval for μ. "$(\bar{X}_n-1.4, \bar{X}_n+1.4)$ is a 95% confidence interval for μ" means that if we obtained 100 copies of $\bar{X}_n$, 95 copies of $(\bar{X}_n-1.4, \bar{X}_n+1.4)$ contain the population mean μ.

$\bar{X}_n$	$(\bar{X}_n-1.4,\ \bar{X}_n+1.4)$
7.5	(6.1, 8.9)
7.2	(5.8, 8.6)
8.1	(6.7, 9.5)
⋮	⋮
6.3	(4.9, 7.7)



Definition 8.1. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population distribution $f(x;\theta)$, where θ is unknown parameter. $[\underline{\theta}(X_1,X_2,\ldots,X_n), \overline{\theta}(X_1,X_2,\ldots,X_n)]$ is called a $(1-\nu)$ 100% confidence interval for θ if

$$P_r\{\underline{\theta}(X_1,X_2,\ldots,X_n) \leq \theta \leq \overline{\theta}(X_1,X_2,\ldots,X_n)\}$$

$$= 1-\nu.$$

We note that

(1) the lower limit $\underline{\theta}(X_1,X_2,\ldots,X_n)$ and the upper limit $\overline{\theta}(X_1,X_2,\ldots,X_n)$ are both functions of the random sample $X_1,X_2,\ldots,X_n$.

(2) In Example 8.1,

$$\underline{\theta}(X_1,X_2,\ldots,X_n) = \overline{X}_n - 1.96\,\frac{\sigma}{\sqrt{n}},$$

$$\overline{\theta}(X_1,X_2,\ldots,X_n) = \overline{X}_n + 1.96\,\frac{\sigma}{\sqrt{n}}.$$

(3) Possible values for "ν" are usually

ν	$1-\nu$	$(1-\nu)$ 100%
.01	.99	99%
.02	.98	98%
.05	.95	95%
.10	.90	90%

Theorem 8.1. If $X \overset{d}{\sim} N(\mu,\sigma)$ with σ known, then $(1-\nu)$ 100% confidence interval for μ is

$$\left[\bar{X}_n - Z_{\frac{\nu}{2}} \frac{\sigma}{\sqrt{n}} , \quad \bar{X}_n + Z_{\frac{\nu}{2}} \frac{\sigma}{\sqrt{n}}\right].$$

Example 8.2. Suppose that $X \overset{d}{\sim} N(\mu,\sigma=5)$ and $\bar{X}_n = 70$ computed from 25 samples. To find 90% confidence interval for μ;

$$\nu = .10 \ , \ \frac{\nu}{2} = .05, \ Z_{\frac{\nu}{2}} = Z_{.05} = 1.645$$

$$\underline{\mu} = \bar{X}_n - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 70 - 1.645 \times \frac{5}{\sqrt{25}} = 68.355$$

$$\bar{\mu} = \bar{X}_n + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 70 + 1.645 \times \frac{5}{\sqrt{25}} = 71.645$$

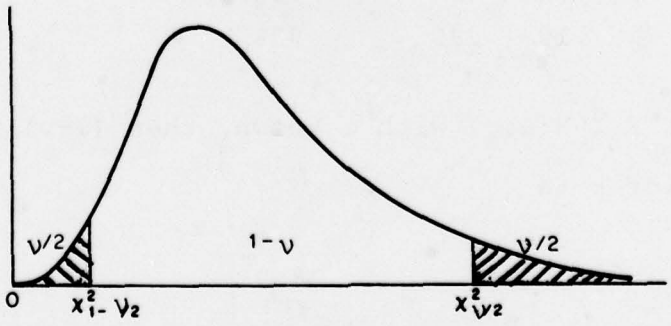
$\therefore$ (68.355, 71.645) is a 90% confidence interval for μ.

(B) For Weibull Distributions

Suppose that $X \overset{d}{\sim} W(\beta,\alpha)$ with α known and β unknown parameters. It has been shown that

$$\hat{\beta}_n = \left[\frac{1}{n} \sum_{i=1}^{n} X_i^{\alpha}\right]^{1/\alpha} \text{ is M.L.E. of } \beta \text{ and}$$

$$(*) \quad 2n\left(\frac{\hat{\beta}_n}{\beta}\right)^{\alpha} \overset{d}{\sim} \chi^2(2n).$$



Theorem 8.2 (Two-sided)

$$\left[\left(\frac{2n}{\chi^2_{\frac{\nu}{2}}(2n)}\right)^{1/\alpha}\hat{\beta}_n \ , \ \left(\frac{2n}{\chi^2_{1-\frac{\nu}{2}}(2n)}\right)^{1/\alpha}\hat{\beta}_n\right]$$

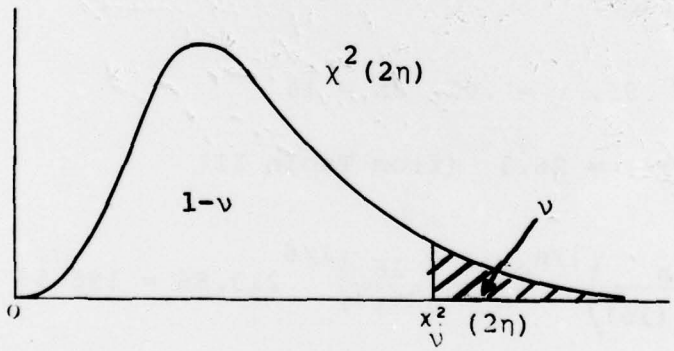
is $(1-\nu)$ 100% confidence interval for β.

Proof: $P_r\left\{\chi^2_{1-\frac{\nu}{2}}(2n) \le 2n\left(\frac{\hat{\beta}_n}{\beta}\right)^{\alpha} \le \chi^2_{\frac{\nu}{2}}(2n)\right\} = 1-\nu$ from (*) and simplify the equation.

Theorem 8.3. (one-sided)

$$\left[\left(\frac{2n}{\chi^2_{\nu}(2n)}\right)^{1/\alpha}\hat{\beta} \ , \ \infty\right]$$

is $(1-\nu)$ 100% confidence interval for β.



Proof: $P_r\left\{2n\left(\frac{\beta_n}{\beta}\right)^{\alpha} \le \chi^2_{\nu}(2n)\right\} = 1-\nu$ and simplify the equation.

Example 8.3. Continuation of Example 7.1.

From Data D, $\hat{\beta}_n = 213.86$, $\alpha = 6$, and $n = 8$. 95% confidence interval for β is,

(a) Two-sided;

$1-\nu = .95,\ \nu = .05,\ \frac{\nu}{2} = .025,\ 1-\frac{\nu}{2} = .975$

$2n = 16$

$\chi^2_{.025}(16) = 28.8, \quad \chi^2_{.975}(16) = 6.91$

(from Table II)

$$\left(\frac{2n}{\chi^2_{.975}(16)}\right)^{1/\alpha}\hat{\beta}_n = \left(\frac{16}{28.8}\right)^{1/6} 213.86 = 193.91$$

$$\left(\frac{2n}{\chi^2_{.025}(16)}\right)^{1/\alpha}\hat{\beta}_n = \left(\frac{16}{6.91}\right)^{1/6} 213.86 = 245.98$$

$\therefore$ (193.91, 245.98) is 95% confidence interval (two-sided) for β.

(b) One-sided;

$1-\nu = .95,\ \nu = .05,\ 2n = 16$

$\chi^2_{.05}(16) = 26.3$ (from Table II)

$$\left(\frac{2n}{\chi^2_{.05}(16)}\right)^{1/\alpha}\hat{\beta}_n = \left(\frac{16}{26.3}\right)^{1/6} 213.86 = 196.86$$

$\therefore$ (196.86, ∞) is 95% confidence interval (one-sided) for β.

(C) Applications in Probabilistic Design

Suppose that the breaking strength (or fatigue life) X of a composite material has a Weibull $W(\beta,\alpha)$ distribution with α

value known. On the basis of test data $x_1, x_2, \ldots, x_n$ from the population X, the unknown parameter β is estimated by the M.L.E. $\hat{\beta}_n = \left[\frac{1}{n}\sum_{i=1}^{n} x_i^{\alpha}\right]^{1/\alpha}$. The lower limit of one-sided 95% confidence interval for β is computed as $\hat{\beta}_L = \left(\frac{2n}{\chi^2_{.05}(2n)}\right)^{1/\alpha}\hat{\beta}_n$ so that $P_r\{\beta \geq \hat{\beta}_L\} = .95$.

Note that $\hat{\beta}_L$ is a very "conservative" estimate of β and can be used for computing design strength (or life). If $X \overset{d}{\sim} W(\hat{\beta}_L, \alpha)$, then the design strength (or life) x can be related to the reliability as

$$\text{Reliability } R = P_r\{X \geq x\} = \exp - \left(\frac{x}{\hat{\beta}_L}\right)^{\alpha}$$

if and only if $x = \hat{\beta}_L\left(\ell n \frac{1}{R}\right)^{1/\alpha}$. The following design values are defined for the reliability R = .99 and R = .90.

Definition 8.2.

$$x_A \text{ (A-allowable)} = \hat{\beta}_L\left(\ell n \frac{1}{.99}\right)^{1/\alpha}$$

$$x_B \text{ (B-allowable)} = \hat{\beta}_L\left(\ell n \frac{1}{.90}\right)^{1/\alpha}.$$

Example 8.4. Continuation of Example 8.3.

(b) $\hat{\beta}_L = 196.86$ and $\alpha = 6$.

A-allowable; $x_A = 196.86 \times \left(\ell n \frac{1}{.99}\right)^{1/6} = 91.45$

B-allowable; $x_B = 196.86 \times \left(\ell n \frac{1}{.90}\right)^{1/6} = 135.29$.

9. TESTING HYPOTHESES

It is often the purpose of a statistical inference to answer a "yes" or "no" question about some characteristics of a population.

Suppose that a measurement random variable X has a p.d.f. $f(x;\theta)$ with unknown θ. On the basis of a random sample $x_1, x_2, \ldots, x_n$ from $f(x;\theta)$, it is desired to decide whether a hypothesized proposition on θ is true or false. Suppose that two contradictory states of nature are represented by two hypotheses,

$$\text{null hypothesis } H_0;\ \theta = \theta_0$$
$$\text{alternative hypothesis } H_1;\ \theta = \theta_1$$

and a decision should be made to reject H_0 (accept H_1) or to accept H_0 (reject H_1).

Two possible errors exist in this decision making process;

$$\text{Type I; reject } H_0 \text{ when } H_0 \text{ is true}$$
$$\text{Type II; accept } H_0 \text{ when } H_1 \text{ is true .}$$

The null hypothesis H_0 should be the one which corresponds to the state of nature that one would not eliminate unless there is a strong evidence to do so. It follows then that the Type I error, "reject H_0 when H_0 is true" is the more serious error.

The testing procedure is then in such a way that the probability of Type I error (the worst error) should be less than a small prescribed probability, say 0.1, 0.05, 0.02 or 0.01, and the probability of Type II error should be minimized.

The level of significance ν (ν = .1, 0.05, 0.02, or 0.01) is defined to be the maximum probability of the Type I error, i.e.

$$P_r\{\text{Reject } H_0 | H_0 \text{ true}\} \leq \nu.$$

The testing of the hypotheses is then to find a region for "rejecting H_0", which is called a "critical region", so that

$$P_r\{\text{Reject } H_0 | H_0 \text{ true}\} = P_r\{\text{critical region} | H_0 \text{ true}\} \leq \nu.$$

The critical region can be expressed in terms of a statistic $\hat{\theta} = \theta(X_1, X_2, \ldots, X_n)$ a function of a random sample $X_1, X_2, \ldots, X_n$ from $f(x;\theta)$, which is related to the parameter θ that the hypotheses are stated on.

Example 9.1. Suppose that a composite material strength X has $N(\mu, 10)$ distribution and it is desired to test,

$$H_0;\ \mu = 90$$

$$H_1;\ \mu > 90 \quad \text{with } \nu = .05.$$

On the basis of nine samples of tests, the sample mean $\bar{X}_n = 93.2$ is obtained.

Note that we reject H_0 (accept H_1) if and only if we have $\bar{X}_n > a$ for some real number a.

In order to find a critical region, i.e. to find the value of a,

$$P_r\{\text{Reject } H_0 \mid H_0 \text{ true}\} = P_r\{\bar{X}_n > a \mid \mu = 90\}$$

$$= P_r\left\{\frac{\bar{X}_n - 90}{10/\sqrt{9}} > \frac{a-90}{10/\sqrt{9}}\right\} \left(\text{since } \bar{X}_n \overset{d}{\sim} N\left(\mu, \frac{10}{\sqrt{9}}\right) \right.$$

$$\left. \text{according to Theorem 6.1}\right)$$

$$= P_r\left\{Z \geq \frac{a-90}{10/\sqrt{9}}\right\} = .05 = \nu, \quad \text{i.e.}$$

$$\frac{a-90}{10/3} = 1.645 \text{ (from Table I) and}$$

$$a = 90 + 1.645 \times \frac{10}{3} = 95.483,$$

hence we have a critical region;

Reject H_0 if and only if $\bar{X}_n > 95.483$.

Therefore our conclusion on the basis of $\bar{X}_n = 93.2$ is then to "accept H_0".

Example 9.2. Suppose that $X \overset{d}{\sim} W(\beta;\alpha=10)$ and it is desired to test

$$H_0;\ \beta = 100$$

$$H_1;\ \beta > 100 \quad \text{with } \nu = 0.05.$$

On the basis of nine data points, we obtained

$$\hat{\beta}_9 = \left[\frac{1}{n}\sum_{i=1}^{n} x_i^{\alpha}\right]^{1/\alpha} = \left[\frac{1}{9}\sum_{i=1}^{n} x_i^{10}\right]^{1/10}$$

$$= 106.3$$

Under H_0, $2n\left(\frac{\hat{\beta}_n}{\beta}\right)^{\alpha} = 2 \times 9\left(\frac{\hat{\beta}_9}{100}\right)^{10} \overset{d}{\sim} \chi^2(18)$ so $P_r\left\{18\left(\frac{\hat{\beta}}{100}\right)^{10} \geq \chi^2_{.05}(18)\right\}$ $= .05$, i.e. $\chi^2_{.05}(18) = 28.87$ (from Table II) and $P_r\{\hat{\beta}_9 \geq 104.84\}$ $= .05$. Hence, reject H_0 if and only if $\hat{\beta}_9 \geq 104.84$. Therefore, our decision is to "reject H_0".

Chi-Square Goodness of Fit Tests

On the basis of a random sample $x_1, x_2, \ldots, x_n$, it is often desired to know "what (family) is the population distribution $f(x;\theta)$". Is it Weibull, Normal or something else?

To test hypotheses,

$$H_0;\ f(x;\theta) \text{ is Weibull (Normal, etc.)}$$

$$H_1;\ f(x;\theta) \text{ is not Weibull}$$

with the level of significance ν, we use the Chi-square goodness of fit test;

Classes	Observed Frequency	Expected Frequency
a_0-a_1	f_1	e_1
a_1-a_2	f_2	e_2
.	.	.
.	.	.
.	.	.
a_{m-1}-a_m	f_m	e_m

n = sample size

m = number of classes

f_i = number of samples between a_{i-1} and a_i

$e_i = n \times p_i$, where $p_i = P_r\{a_{i-1} \leq X \leq a_i\}$ and the random variable X has a distribution given under H_0.

The statistic $\chi^2 = \sum_{i=1}^{m} \frac{(f_i - e_i)^2}{e_i}$ measures the difference between observed and expected frequencies and under null hypothesis H_0,

$$\chi^2 \overset{d}{\sim} \chi^2(m-r),$$

where $r = 1 +$ {number of parameters estimated to compute the expected frequency e_i}.

Now reject H_0 if $\chi^2 \geq \chi^2_{\nu}(m-r)$

Example 9.3.

H_0: Data B came from a Weibull distribution $\nu = .05$

$n = 36$, $\alpha = 13$ and $\beta = 213$

classes	frequency f_i	p_i	expected frequency $e_i = n \times p_i$
less than 159.5	2	.023	.83
159.5 - 169.5	0	.027	.97
169.5 - 179.5	0	.053	1.89
179.5 - 189.5	2	.094	3.38
189.5 - 199.5	5	.151	5.44
199.5 - 209.5	9	.206	7.42
209.5 - 219.5	6	.219	7.85
219.5 - 229.5	8	.157	5.64
229.5 - 239.5	3	.061	2.21
more than 239.5	1	.010	.37

$$m = 10, \quad r = 3$$

$$\chi^2 = \sum_{i=1}^{10} \frac{(f_i - e_i)^2}{e_i} = 4.695$$

$$\chi^2_{.05}(7) = 14.1 \text{ (from Table II).}$$

Reject H_0 if $\chi^2 \geq 14.1$.

Our decision is to accept H_0.

10. SUMMARY

(a) A two parameter Weibull distribution $W(\beta,\alpha)$ is used for representing the static strength or fatigue life distributions of composite materials.

(b) Hahn-Kim and Yang's residual strength degradation model was presented as a fatigue model of composite materials.

(c) A graphical plotting and a maximum likelihood estimation method was introduced for estimating the parameters of the Weibull distributions. The 95% lower confidence bound of β was obtained for design purposes (A-allowable and B-allowable).

(d) The problem of testing hypotheses on the Weibull scale parameter β was solved (χ^2-test) when the shape parameter α is known.

(e) Some additional results on the estimations and testing hypotheses involving the Weibull distributions are needed to be investigated and summarized for the applications in composite materials engineering.

☆U.S.Government Printing Office: 1979 — 657-002/161