U.S. DEPARTMENT OF THE INTERIOR

U.S. GEOLOGICAL SURVEY

Probability and Statistics for Petroleum Resource Assessment

By

Robert A. Crovelli[1]

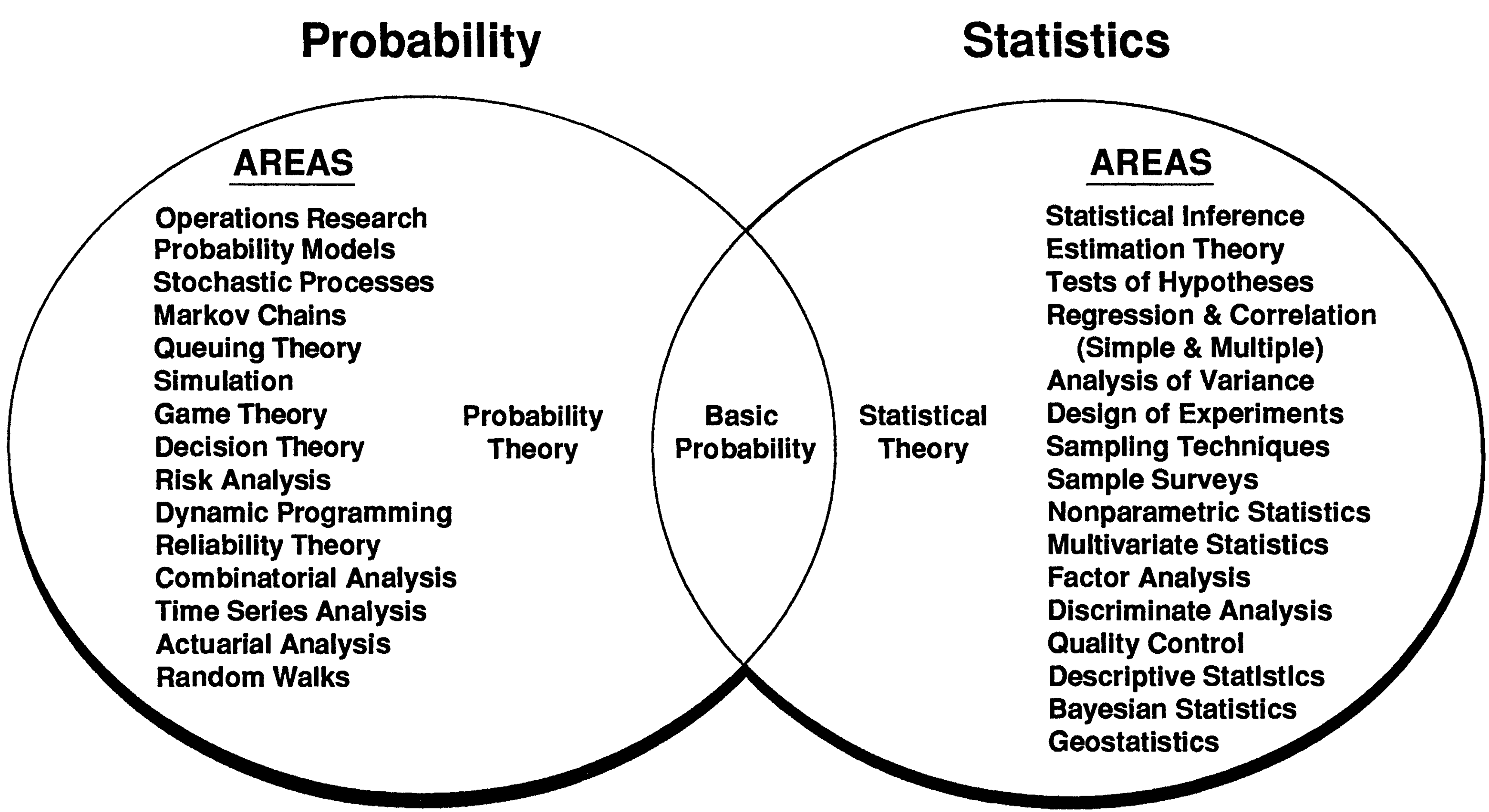
Open-File Report 93-582

[1]U.S. Geological Survey, Box 25046, MS 971, DFC, Denver, Colorado 80225

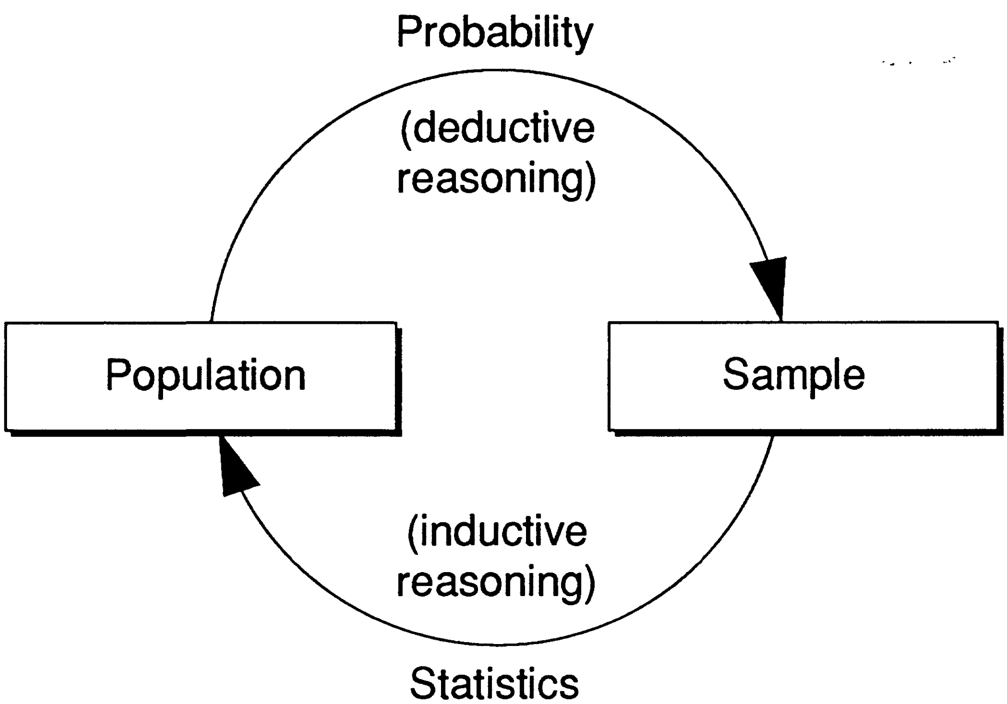
1993

TABLE OF CONTENTS

Venn Diagram for Describing the Fields of Probability and Statistics



The Relationship Between Probability and Inferential Statistics



PROBABILITY

I. Probability

A. Basic Concepts

1. Petroleum accumulation classification hierarchy

<u>Pool</u>: An individual accumulation or reservoir of oil or gas.

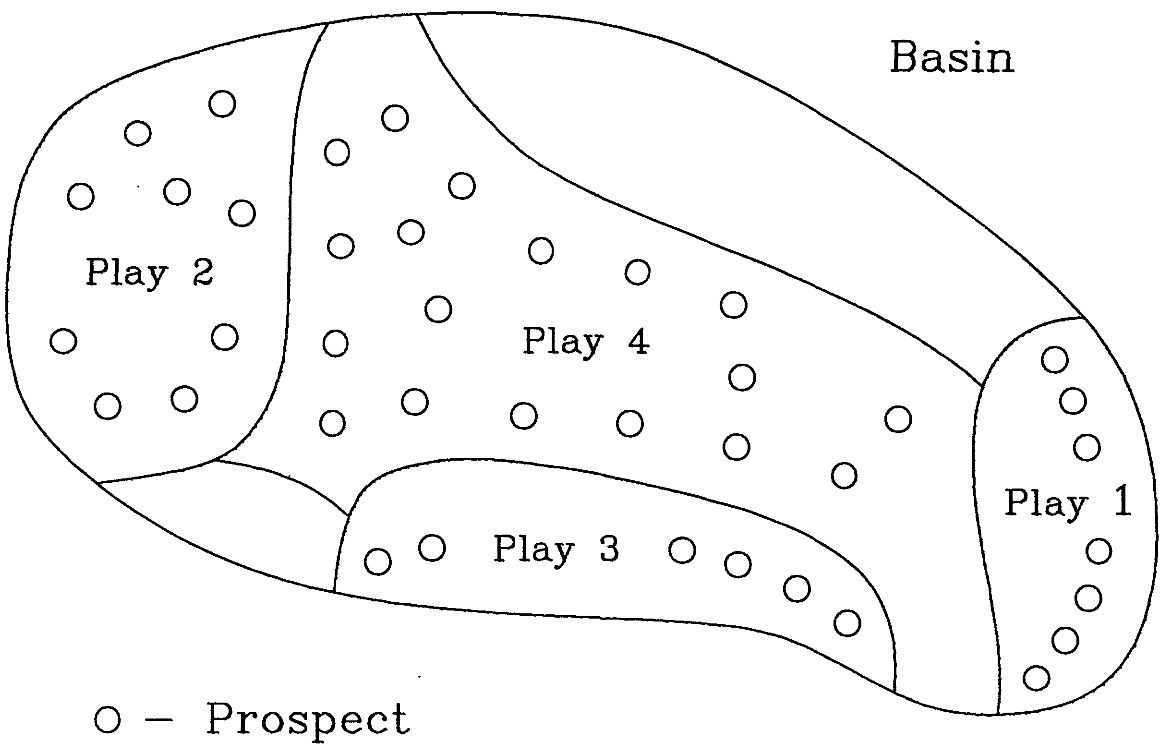
<u>Field</u>: A set of one or more pools of oil or gas that are related to a single structural or stratigraphic feature.

<u>Prospect</u>: A potential oil or gas field.

<u>Play</u>: A set of one or more prospects that are geologically related in their hydrocarbon sources, reservoirs, traps, and geologic histories.

<u>Province (or basin)</u>: A set of one or more plays that are hydrodynamically related.

<u>Region</u>: A set of one or more provinces that are geographically related.



2. Experiment, sample space, and event

Experiment: any process or action that generates observations.

Experiment: Three-prospect assessment

Suppose we are assessing three prospects in a new play. Each prospect results in one of two possible outcomes. Let "success" (S) denote having an oil or gas field and "failure" (F) denote being dry.

Sample space: a set of all possible outcomes (sample points) of an experiment.

Sample Space
SSS
SSF
SFS
SFF
FSS
FSF
FFS
FFF

Event: a subset of a sample space.

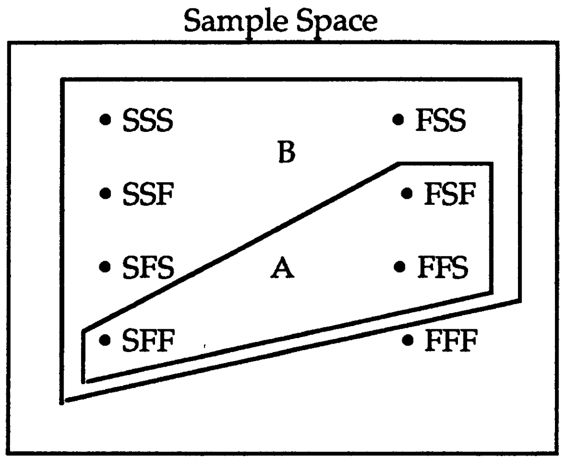
Let event A: Exactly one field

A = {SFF, FSF, FFS}

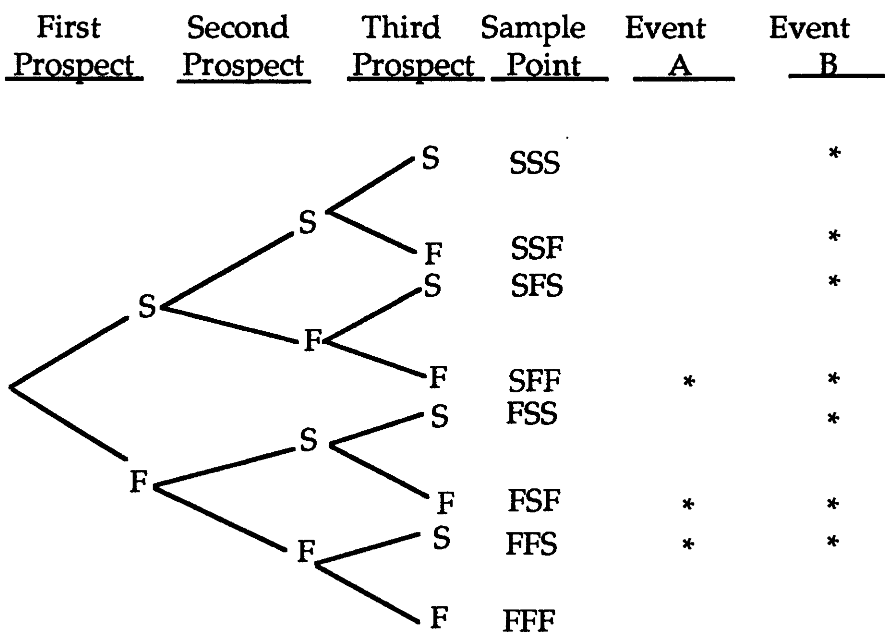
and event B: At least one field

B = {SSS, SSF, SFS, SFF, FSS, FSF, FFS}

3. Venn diagram



4. Tree diagram



5. Event relations

a. Union of events

The union of two events A and B, denoted by $A \cup B$ and read "A or B," is the event containing all outcomes in A or B or both.

Let A = {SFF, FSF, FFS} and B = {SSS, SSF, SFS, SFF, FSS, FSF, FFS},

then $A \cup B$ = {SSS, SSF, SFS, SFF, FSS, FSF, FFS}

b. Intersection of events

The intersection of two events A and B, denoted by $A \cap B$ and read "A and B," is the event containing all outcomes in both A and B.

Let A = {SFF, FSF, FFS} and B = {SSS, SSF, SFS, SFF, FSS, FSF, FFS},

then $A \cap B$ = {SFF, FSF, FFS}

c. Complement of event

The complement of an event A, denoted by A' and read "not A," is the event containing all outcomes of the sample space that are not in A.

Let A = {SFF, FSF, FFS},

then A' = {SSS, SSF, SFS, FSS, FFF}

d. Mutually exclusive events

Two events A and B are mutually exclusive or disjoint events if A and B have no outcomes in common, i.e., $A \cap B = \emptyset$

Let A = {SFF, FSF, FFS} and B = {SSF, SFS, FSS},

then $A \cap B = \emptyset$

Therefore, A and B are mutually exclusive events.

6. Combinatorial analysis (counting techniques)

   a. Fundamental principle of counting

      If an operation can be performed in n_1 ways, and if for each of these a second operation can be performed in n_2 ways, and for each of the first two a third operation can be performed in $n_3$ways, and so forth, then the sequence of k operations can be performed in $n_1n_2...n_k$ ways.

      Experiment: Three-prospect assessment

      Number of sample points in sample space = 2•2•2 = 8

   b. Combinations

      A <u>combination</u> is any <u>unordered</u> subset of r objects taken from a set of n distinct objects.

      The number of combinations of r objects taken from n distinct objects is

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

      where "n factorial" is n! = n(n-1)(n-2)•••(3)(2)(1) and 0! = 1.

      Number of sample points in event A = $\binom{3}{1} = \frac{3!}{1!2!} = \frac{3 \bullet 2 \bullet 1}{1 \bullet 2 \bullet 1} = 3$

      Number of sample points in event B = $\binom{3}{1} + \binom{3}{2} + \binom{3}{3} = 3 + 3 + 1 = 7$

   c. Permutations

      A <u>permutation</u> is any <u>ordered</u> subset of r objects taken from a set of n distinct objects.

      The number of permutations of r objects taken from n distinct objects is

$$P_{r,n} = \frac{n!}{(n-r)!}$$

      The number of ways of selecting with order (permutations) 3 prospects from a set of 5 prospects is

$$P_{3,5} = \frac{5!}{(5-3)!} = \frac{5!}{2!} = 5 \bullet 4 \bullet 3 = 60$$

7. Definitions of probability

Idea of probability:

The probability of an event is a numerical measure of the likelihood that the event will occur.

a. Classical definition of probability

If an experiment can result in any one of N different equally likely outcomes, and if exactly n of these outcomes correspond to event A, then the probability of event A is

$$P(A) = \frac{n}{N}$$

Example: Suppose four of ten prospects have petroleum fields. If three prospects are selected at random to be explored, what is the probability of getting exactly two fields?

Let A: Exactly two fields

$$P(A) = \frac{\binom{4}{2}\binom{6}{1}}{\binom{10}{3}} = \frac{\frac{4!}{2!2!} \bullet \frac{6!}{1!5!}}{\frac{10!}{3!7!}} = 0.3$$

b. Relative frequency definition of probability

Consider a sequence of repetitions of the same experiment under identical conditions. Let f_n denote the number of occurrences of the event A in the first n repetitions of the experiment. The ratio f_n/n then gives the relative frequency of occurrence of event A in the first n repetitions. The probability of event A is

$$P(A) = \lim_{n \to \infty} \frac{f_n}{n}$$

i.e., the limiting relative frequency of occurrence of event A in the long run.

The relative frequency definition is more general than the classical definition.

The relative frequency definition includes the classical definition.

Example: The probability of a dry hole in a particular explored basin is 0.8 from past statistical data.

c. Subjective definition of probability

A personal opinion (depending on the information held by a person at some time) of the likelihood that an event will occur. Subjective probability includes the case where past statistical data are not available and/or the information available is of an indirect nature.

The subjective definition is more general than the relative frequency definition.

The subjective definition includes the relative frequency definition.

In petroleum resource assessment the application, assignment and interpretation of probability is based on the subjective definition of probability.

Example: The probability of recoverable petroleum in an unexplored play is 0.3 without past statistical data.

d. Axiomatic definition of probability

The probability of an event A is the sum of the weights of all sample points in A. Therefore,

$$0 \leq P(A) \leq 1, P(\emptyset) = 0, \text{ and } P(S) = 1$$

where Ø denotes the empty set and S the sample space.

The theory of probability is based on the axiomatic definition of probability.

The classical and relative frequency definitions of probability can be derived as theorems from the axiomatic definition.

8. Probability of event relations

Suppose we assume equally likely sample points for the experiment: three-prospect assessment.

<u>Sample Space</u>

SSS
SSF
SFS
SFF
FSS
FSF
FFS
FFF

a. Given event A: Exactly one field

$$A = \{SFF, FSF, FFS\},$$

then $P(A) = 3/8$

b. Given event B: At least one field

$$B = \{SSS, SSF, SFS, SFF, FSS, FSF, FFS\},$$

then $P(B) = 7/8$

c. Given $A \cup B = \{SSS, SSF, SFS, SFF, FSS, FSF, FFS\}$,

then $P(A \cup B) = 7/8$

d. Given $A \cap B = \{SFF, FSF, FFS\}$,

then $P(A \cap B) = 3/8$

e. Given $A' = \{SSS, SSF, SFS, FSS, FFF\}$,

then $P(A') = 5/8$

f. Given $B' = \{FFF\}$,

then $P(B') = 1/8$

9. Conditional probability

Notation $P(A \mid B)$ denotes the conditional probability of event A given that the event B has occurred.

Given that B has occurred, event B becomes the new reduced sample space.

Conditional probability:

For any two events A and B with $P(B) > 0$, the conditional probability of A given that B has occurred is defined by

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

Consider the experiment: three-prospect assessment with equally likely sample points.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{3/8}{7/8} = 3/7$$

Independence:

Two events A and B are independent if $P(A \mid B) = P(A)$ and are dependent otherwise.

Consider the experiment: three-prospect assessment with equally likely sample points. Since

$$P(A \mid B) = 3/7 \text{ and } P(A) = 3/8 \Rightarrow P(A \mid B) \neq P(A),$$

events A and B are dependent.

Remark: If two events are mutually exclusive, then they are dependent.

10. Probability rules

a. Addition rule

If A and B are any two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

b. Special addition rule

If A and B are mutually exclusive events, then

$$P(A \cup B) = P(A) + P(B)$$

c. Complement rule

If A and A' are complementary events, then

$$P(A) + P(A') = 1$$

or

$$P(A) = 1 - P(A')$$

d. Multiplication rule

If A and B are any two events, then

$$P(A \cap B) = P(A \mid B)P(B)$$

and

$$P(A \cap B) = P(A)P(B \mid A)$$

e. Special multiplication rule

If A and B are independent events, then

$$P(A \cap B) = P(A)P(B)$$

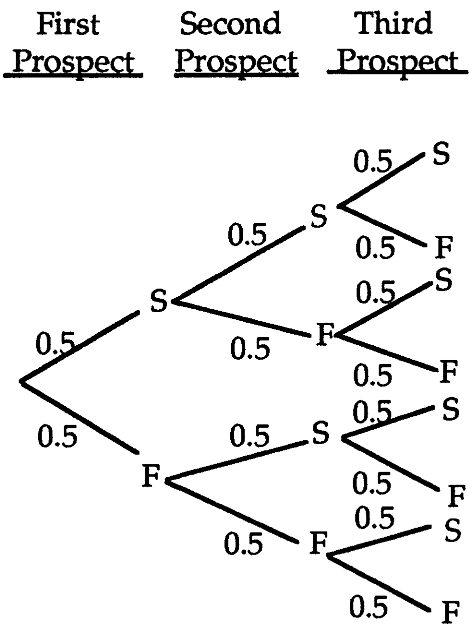
f. Another special multiplication rule

If A and B are mutually exclusive events, then

$$P(A \cap B) = 0$$

11. Applications of probability rules

a. Equally likely



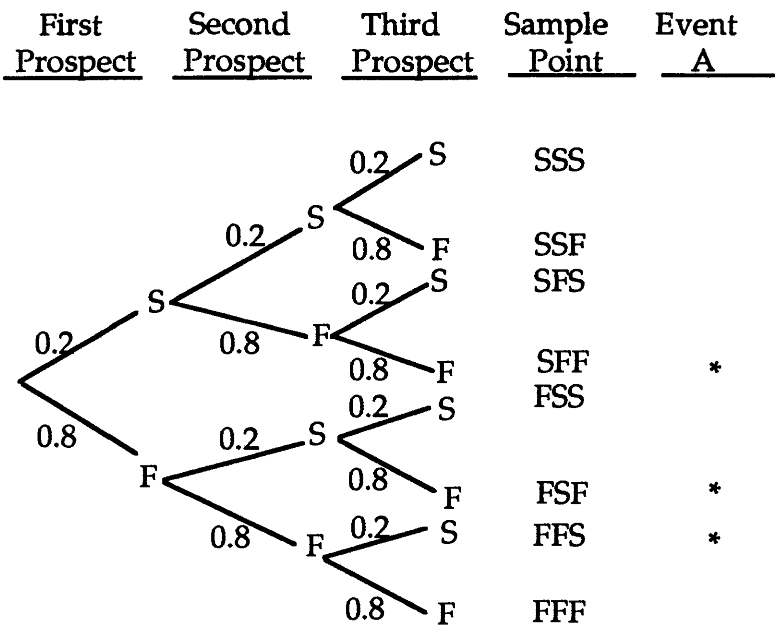
Sample Point	Event A	Probability
SSS		0.125
SSF		0.125
SFS		0.125
SFF	*	0.125
FSS		0.125
FSF	*	0.125
FFS	*	0.125
FFF		0.125
		1.000

$P(A) = P(SFF \cup FSF \cup FFS) = 3/8 = 0.375$

or $= P(SFF) + P(FSF) + P(FFS) = 0.125 + 0.125 + 0.125 = 0.375$

or $= P(S)P(F)P(F) + P(F)P(S)P(F) + P(F)P(F)P(S) = 3(0.5)^3 = 0.375$

b. Bernoulli process

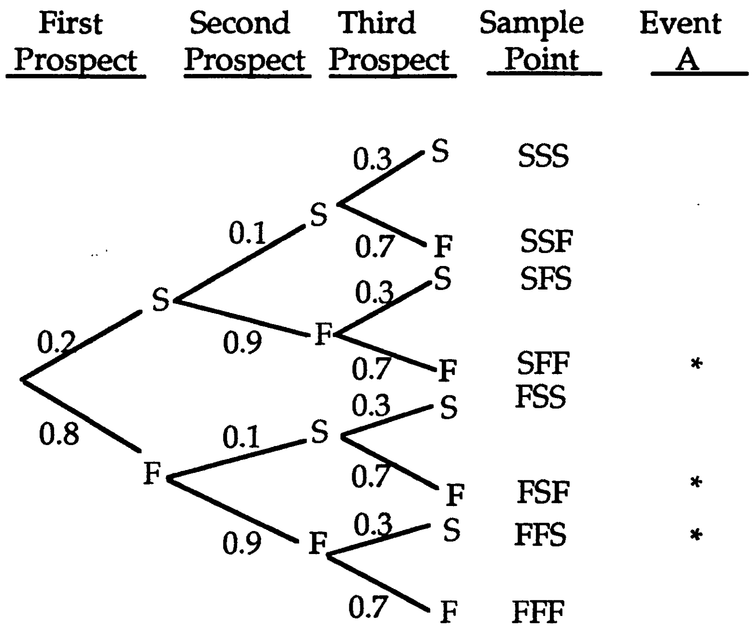


$$P(A) = P(SFF) + P(FSF) + P(FFS)$$

$$= P(S)P(F)P(F) + P(F)P(S)P(F) + P(F)P(F)P(S)$$

$$= 3P(S)P(F)P(F) = 3(0.2)(0.8)^2 = 0.384$$

c. Independence

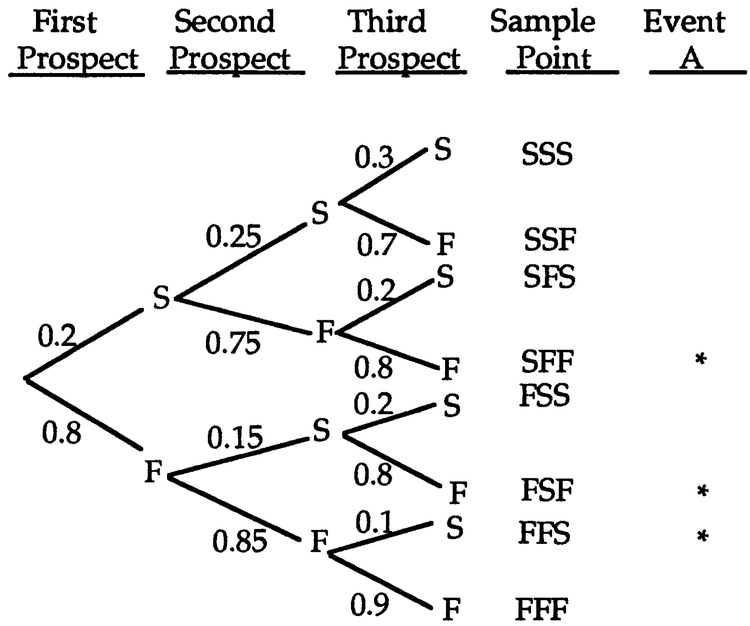


$$P(A) = P(S_1F_2F_3) + P(F_1S_2F_3) + P(F_1F_2S_3)$$

$$= P(S_1)P(F_2)P(F_3) + P(F_1)P(S_2)P(F_3) + P(F_1)P(F_2)P(S_3)$$

$$= (0.2)(0.9)(0.7) + (0.8)(0.1)(0.7) + (0.8)(0.9)(0.3) = 0.398$$

d. Dependence



$P(A) = P(S_1F_2F_3) + P(F_1S_2F_3) + P(F_1F_2S_3)$

$= P(S_1)P(F_2 \mid S_1)P(F_3 \mid S_1F_2) + P(F_1)P(S_2 \mid F_1)P(F_3 \mid F_1S_2) +$

$P(F_1)P(F_2 \mid F_1)P(S_3 \mid F_1F_2)$

$= (0.2)(0.75)(0.8) + (0.8)(0.15)(0.8) + (0.8)(0.85)(0.1) = 0.284$

12. Bayes' rule

Rule of total probability:

If the events $B_1, B_2, \ldots, B_k$ constitute a partition of the sample space such that $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i).$$

Example: Three-prospect assessment with two hypothesized possible states of nature

B_1: Exactly one field; estimated prior probability $P(B_1) = 3/4$

B_2: Exactly two fields; estimated prior probability $P(B_2) = 1/4$

Let event A: First wildcat well drilled results in a dry hole

Conditional probabilities: $P(A \mid B_1) = 2/3$ and $P(A \mid B_2) = 1/3$

Therefore,
$$P(A) = P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2)$$
$$= (3/4)(2/3) + (1/4)(1/3) = 7/12$$

Bayes' rule:

If the events $B_1, B_2, \ldots, B_k$ constitute a partition of the sample space, where $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A such that $P(A) \neq 0$,

$$P(B_r \mid A) = \frac{P(B_r \cap A)}{P(A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}$$

for $r = 1, 2, \ldots, k$.

Example: Posterior probabilities are

$$P(B_1 \mid A) = \frac{P(B_1)P(A|B_1)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2)}$$

$$= \frac{(3/4)(2/3)}{(3/4)(2/3) + (1/4)(1/3)} = \frac{6/12}{7/12} = 6/7$$

$$P(B_2 \mid A) = \frac{P(B_2)P(A|B_2)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2)}$$

$$= \frac{(1/4)(1/3)}{(3/4)(2/3) + (1/4)(1/3)} = \frac{1/12}{7/12} = 1/7$$

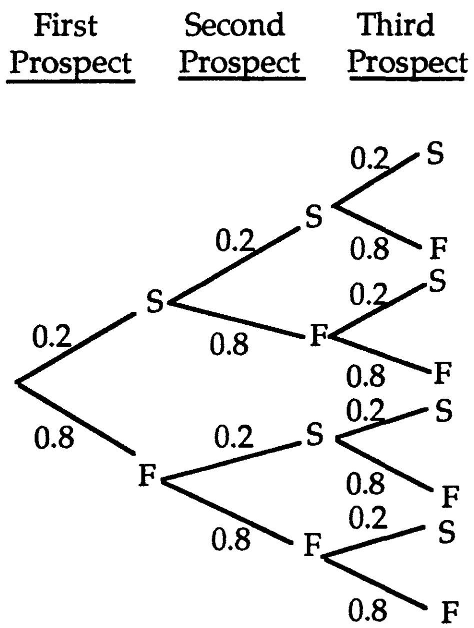
B. Random Variables and Probability Distributions

A random variable X is a function that associates a real number with each element in the sample space.

1. Discrete random variables

a. Binomial random variable

Bernoulli process



Sample Point	Probability	x
SSS	0.008	3
SSF	0.032	2
SFS	0.032	2
SFF	0.128	1
FSS	0.032	2
FSF	0.128	1
FFS	0.128	1
FFF	0.512	0
	1.000	

Let random variable X: Number of fields (successes)

Possible distinct values x = 0, 1, 2, 3

Note that P(X = 1) = P(A) = 0.384

A discrete random variable X can take on a countable number of values.

b. Examples of discrete random variables

X: Number of discoveries

X: Number of dry holes

X: Number of prospects

X: Number of petroleum accumulations

X: Number of oil fields

X: Number of gas fields

X: Number of exploratory wells

2. Discrete probability distributions

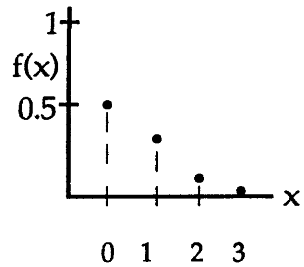
Probability distributions can be expressed in the form of tables, graphs, and formulas.

Binomial distribution

a. Probability mass function (pmf)

$f(x) = P(X = x)$

x	f(x)
0	0.512
1	0.384
2	0.096
3	0.008



b. Cumulative (less than) distribution function (cdf)

$F(x) = P(X \le x)$

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 0.512 & \text{for } 0 \le x < 1 \\ 0.896 & \text{for } 1 \le x < 2 \\ 0.992 & \text{for } 2 \le x < 3 \\ 1 & \text{for } x \ge 3 \end{cases}$$

c. Complementary cumulative (more than) distribution function (ccdf)

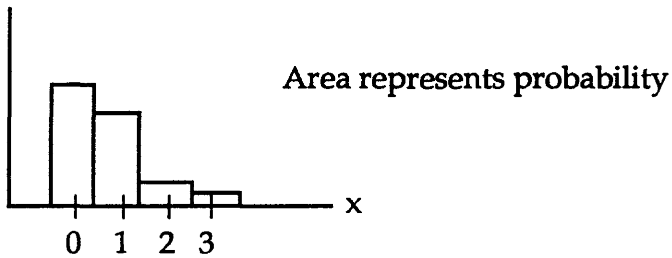
$R(x) = P(X > x) = 1 - F(x)$

$$R(x) = \begin{cases} 1 & \text{for } x < 0 \\ 0.488 & \text{for } 0 \le x < 1 \\ 0.104 & \text{for } 1 \le x < 2 \\ 0.008 & \text{for } 2 \le x < 3 \\ 0 & \text{for } x \ge 3 \end{cases}$$

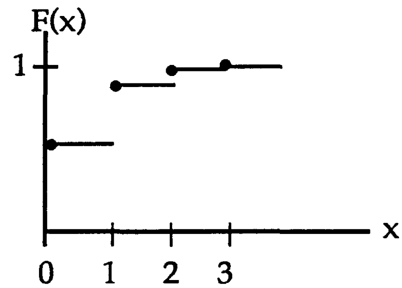
3. Graphs of discrete probability distributions

Binomial distribution

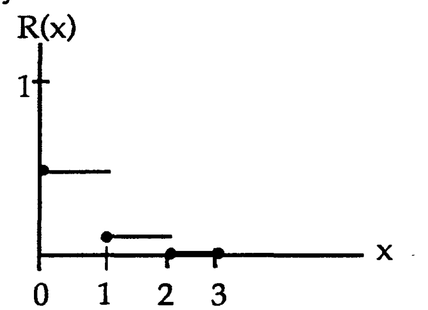
a. Probability histogram



b. Cumulative (less than) distribution function (cdf)



c. Complementary cumulative (more than) distribution function (ccdf)

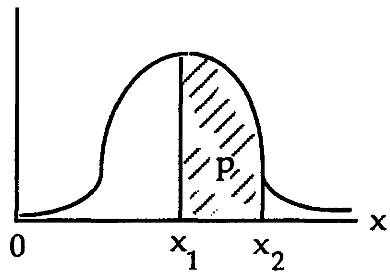


4. Continuous random variables

a. Concept of a continuous random variable

A continuous random variable X can take on a continuum of values.

X = x where x is a real number in an interval, e.g., $0 < x < \infty$ or any positive number.



p is area between x_1 and x_2.

Total area under curve equals 1.

$$P(x_1 < X < x_2) = p$$

b. Examples of continuous random variables

X: Oil field size

X: Gas field size

X: Area of closure

X: Reservoir thickness

X: Reservoir depth

X: Effective porosity

X: Hydrocarbon saturation

X: Reservoir pressure

X: Reservoir temperature

5. Continuous probability distributions

Uniform or rectangular distribution

a. Probability density function (pdf)

$$f(x) = \begin{cases} \dfrac{1}{b - a} & \text{for } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

Parameters: a and b real numbers with a < b

b. Cumulative (less than) distribution function (cdf)

$F(x) = P(X \le x)$

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \dfrac{x - a}{b - a} & \text{for } a \le x \le b \\ 1 & \text{for } x > b \end{cases}$$

c. Complementary cumulative (more than) distribution function (ccdf)

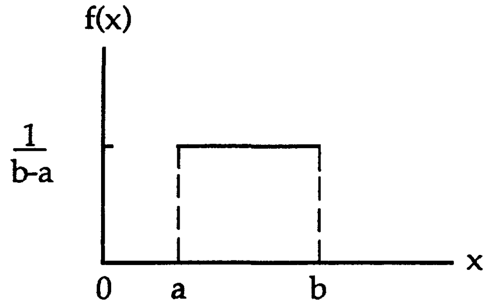
$R(x) = P(X > x) = 1 - F(x)$

$$R(x) = \begin{cases} 1 & \text{for } x < a \\ \dfrac{b - x}{b - a} & \text{for } a \le x \le b \\ 0 & \text{for } x > b \end{cases}$$

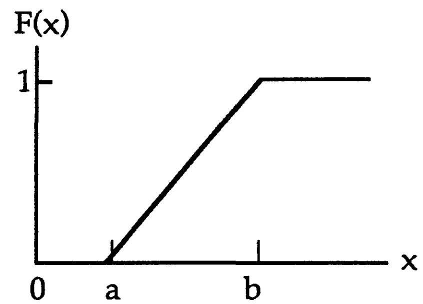
6. Graphs of continuous probability distributions

Uniform or rectangular distribution

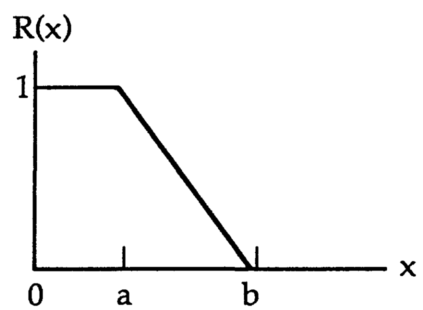
a. Probability density function (pdf)



b. Cumulative (less than) distribution function (cdf)

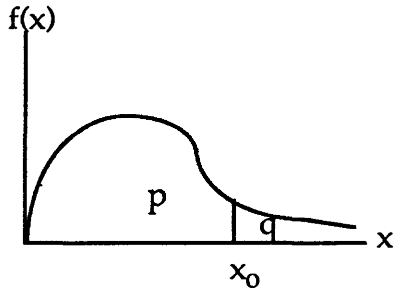


c. Complementary cumulative (more than) distribution function (ccdf)



7. General graphs of continuous probability distributions

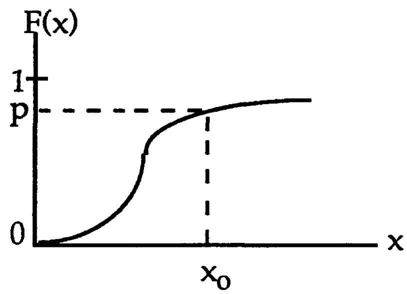
a. Probability density function (pdf)



$p + q = 1$

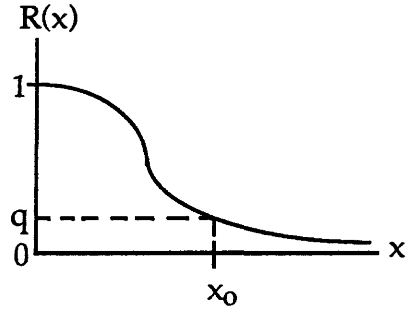
b. Cumulative (less than) distribution function (cdf)

$F(x) = P(X \le x)$



c. Complementary cumulative (more than) distribution function (ccdf)

$R(x) = P(X > x) = 1 - F(x)$



8. Monte Carlo simulation

a. Binomial distribution

X: Number of fields (successes) in 3 prospects (x = 0, 1, 2, 3).

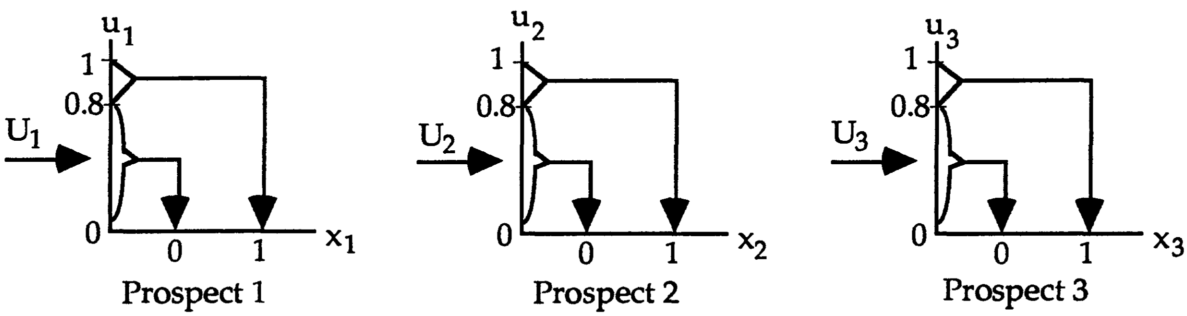
P(field) = 0.2 and P(dry) = 0.8

$X = X_1 + X_2 + X_3$ where

X_1: Number of fields in prospect 1 (x_1 = 0,1),

X_2: Number of fields in prospect 2 (x_2 = 0, 1),

X_3: Number of fields in prospect 3 (x_3 = 0, 1).



Make 5000 simulation passes and compute X each pass.

On each pass, select 3 random numbers (U_1, U_2, U_3) between 0 and 1.

U_i is uniformly distributed over the interval [0,1].

If $0 \le U_i < 0.8$ assign 0, and if $0.8 \le U_i \le 1$ assign 1.

Generate a relative frequency distribution from the 5000 values of X.

For example,

Pass No.	U_1	U_2	U_3	X_1	X_2	X_3	X
1	0.56	0.82	0.12	0	1	0	1
2	0.71	0.63	0.29	0	0	0	0
3	0.89	0.95	0.38	1	1	0	2
•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•
5000	0.47	0.08	0.69	0	0	0	0

A relative frequency distribution of X from an actual simulation, compared to the exact probability distribution of X from the analytic method:

x	freq.	rel. freq.
0	2590	0.518
1	1900	0.380
2	466	0.093
3	44	0.009

x	f(x)
0	0.512
1	0.384
2	0.096
3	0.008

b. Probability distribution for field size

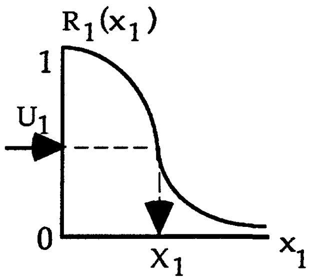
X: Oil field size (barrels),

$X = X_1 \bullet X_2 \bullet X_3$ where

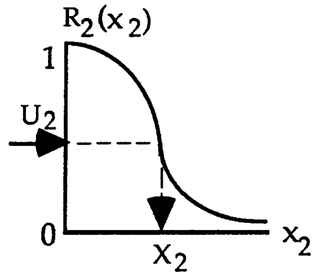
X_1: Area of closure (acres),

X_2: Reservoir thickness (feet),

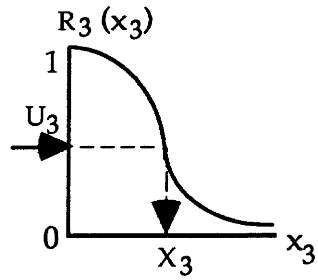
X_3: Oil yield factor (barrels/acre-foot).



Closure



Thickness



Oil Yield Factor

Make 5000 simulation passes and compute X each pass.

On each pass, select 3 random numbers (U_1, U_2, U_3) between 0 and 1.

U_i is uniformly distributed over the interval [0, 1].

From U_1 determine X_1, from U_2 determine X_2, and from U_3 determine X_3, using their ccdf curves as inverse functions.

Generate a relative frequency distribution from the 5000 values of X

c. Comparison of the analytic probability method and the Monte Carlo simulation method

Difficulty of Problem

	Tractable	Partly Tractable	Totally Untractable
Analytic Method	Exact solution	Part exact Part approximate	No solution
Monte Carlo Method	Approximate solution	Approximate solution	Approximate solution

Advantages of the analytic probability method over the Monte Carlo simulation method

1. Exact or part exact solution
2. Much faster procedure on computer (possibly thousands of times faster)
3. More flexible (separate system into modules)
4. Unique solution (whereas Monte Carlo method gives different solution each time applied)
5. Dependency capability (whereas Monte Carlo method generally assumes independence)
6. Mathematical equations describe probabilistic relationships among the random variables.

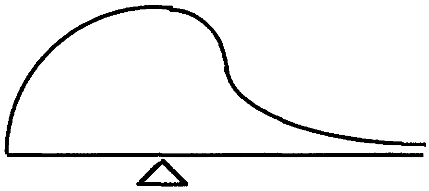
C. Descriptive Parameters

1. Measures of central location

Also called measures of central tendency

a. Mean

i. Probability density function (pdf)



ii. Experiment: Three-prospect assessment

Binomial random variable X: Number of fields

Binomial distribution:

x	f(x)
0	0.512
1	0.384
2	0.096
3	0.008

The mean or expected value of X is

$$\mu = E(X) = \sum_{x} x f(x)$$

$$= (0)(0.512) + (1)(0.384) + (2)(0.096) + (3)(0.008)$$

$$= 0.6$$

iii. Uniform or rectangular distribution

The mean or expected value of X is

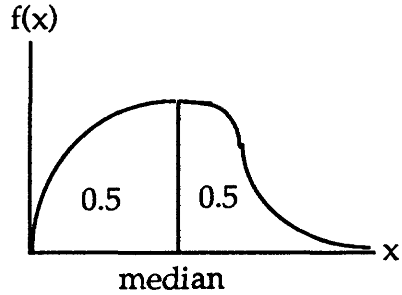
$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

$$= \int_{a}^{b} x\left(\frac{1}{b-a}\right) dx$$

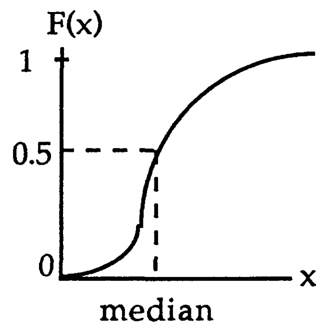
$$= \frac{a+b}{2}$$

b. Median

i. Probability density function (pdf)

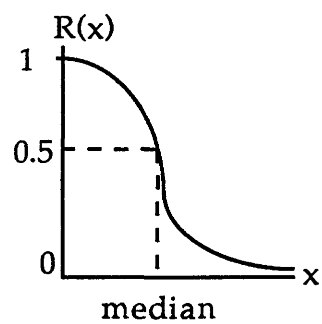


ii. Cumulative (less than) distribution function (cdf)



$$F\ (\text{median}) = P(X \leq \text{median}) = 0.5$$

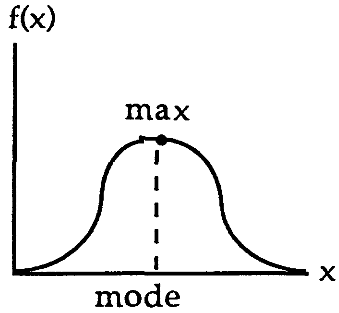
iii. Complementary cumulative (more than) distribution function (ccdf)



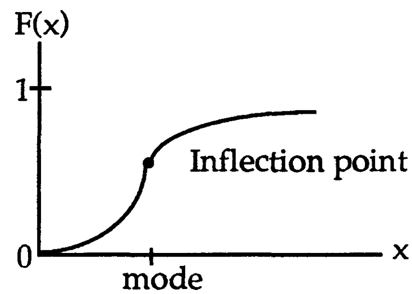
$$R(\text{median}) = P(X > \text{median}) = 0.5$$

c. Mode

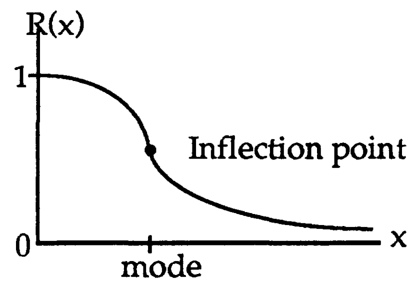
i. Probability density function (pdf)



ii. Cumulative (less than) distribution function (cdf)

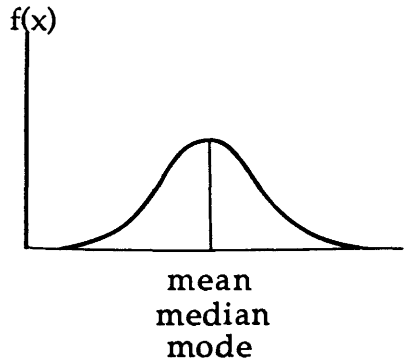


iii. Complementary cumulative (more than) distribution function (ccdf)

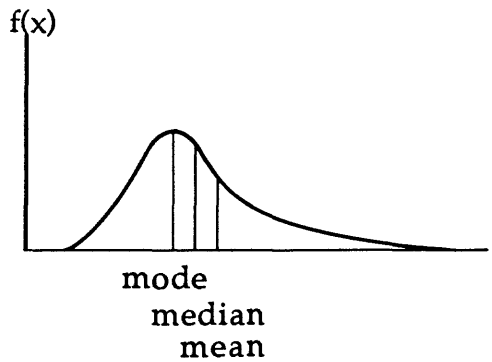


2. Mean, median, and mode related to skewness

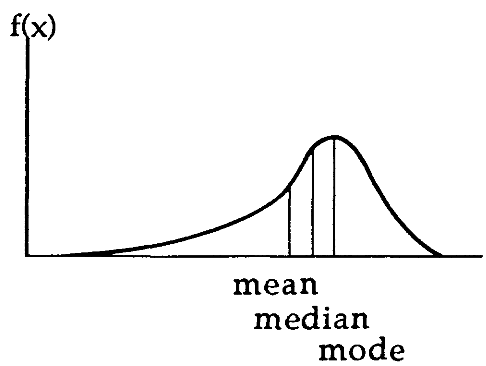
a. Symmetric probability density function (pdf)



b. Positively skewed probability density function (pdf)



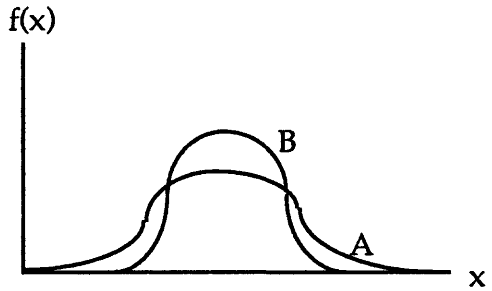
c. Negatively skewed probability density function (pdf)



3. Measures of variation

a. Variance

i. Two probability density functions with different variations



Pdf A has more variation or dispersion or spread than pdf B.

ii. Experiment: Three-prospect assessment

Binomial random variable X: Number of fields

Binomial distribution:

x	f(x)
0	0.512
1	0.384
2	0.096
3	0.008

The mean of X is $\mu = 0.6$

The variance of X is

$$\sigma^2 = E[(X - \mu)^2] = \sum_x (x - \mu)^2 f(x)$$

$$= (0 - 0.6)^2(0.512) + (1 - 0.6)^2(0.384) + (2 - 0.6)^2(0.096)$$

$$+ (3 - 0.6)^2(0.008)$$

$$= 0.48$$

Theorem: $\sigma^2 = E(X^2) - \mu^2$

$$E(X^2) = (0^2)(0.512) + (1^2)(0.384) + (2^2)(0.096) + (3^2)(0.008)$$

$$= 0.84$$

$$\sigma^2 = 0.84 - (0.6)^2 = 0.48$$

b. Standard deviation

i. The standard deviation of X is the positive square root of the variance of X, i.e.,

$$\sigma = \sqrt{\sigma^2}$$

ii. Experiment: Three-prospect assessment

Binomial random variable X: Number of fields

The standard deviation of X is

$$\sigma = \sqrt{0.48} = 0.69$$

iii. Chebyshev's Theorem:

Given <u>any</u> random variable X with mean μ and standard deviation σ, then

for any k > 0, $P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$.

For k = 2, $P(\mu - 2\sigma < X < \mu + 2\sigma) \geq \frac{3}{4}$

For k = 3, $P(\mu - 3\sigma < X < \mu + 3\sigma) \geq \frac{8}{9}$

iv. Experiment: Three-prospect assessment

Binomial random variable X: Number of fields

The mean of X is $\mu = 0.6$

The standard deviation of X is $\sigma = 0.69$

For k = 2, $P[0.6 - 2(0.69) < X < 0.6 + 2(0.69)]$

$$= P(-0.78 < X < 1.98) = 0.896 > 0.75$$

4. Fractiles

a. Fractiles are values of a random variable that correspond to "more than" or excedence probabilities.

The p100th fractile ($0 \leq p \leq 1$), denoted by F_{p100}, is the value of a random variable X such that $P(X > F_{p100}) = p$.

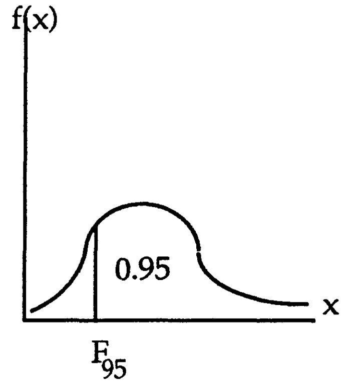
<u>For example</u>

The 95th fractile, F_{95}, is the value of X such that $P(X > F_{95}) = 0.95$.

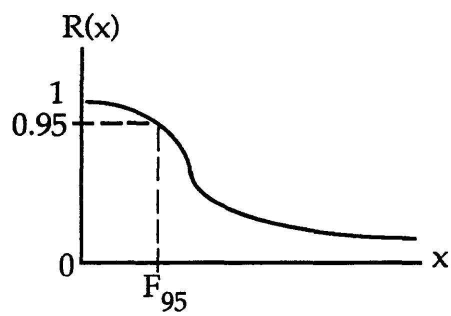
The 5th fractile, F_5, is the value of X such that $P(X > F_5) = 0.05$.

The 50th fractile, F_{50}, is the median.

b. Probability density function (pdf)



c. Complementary cumulative (more than) distribution function (ccdf)



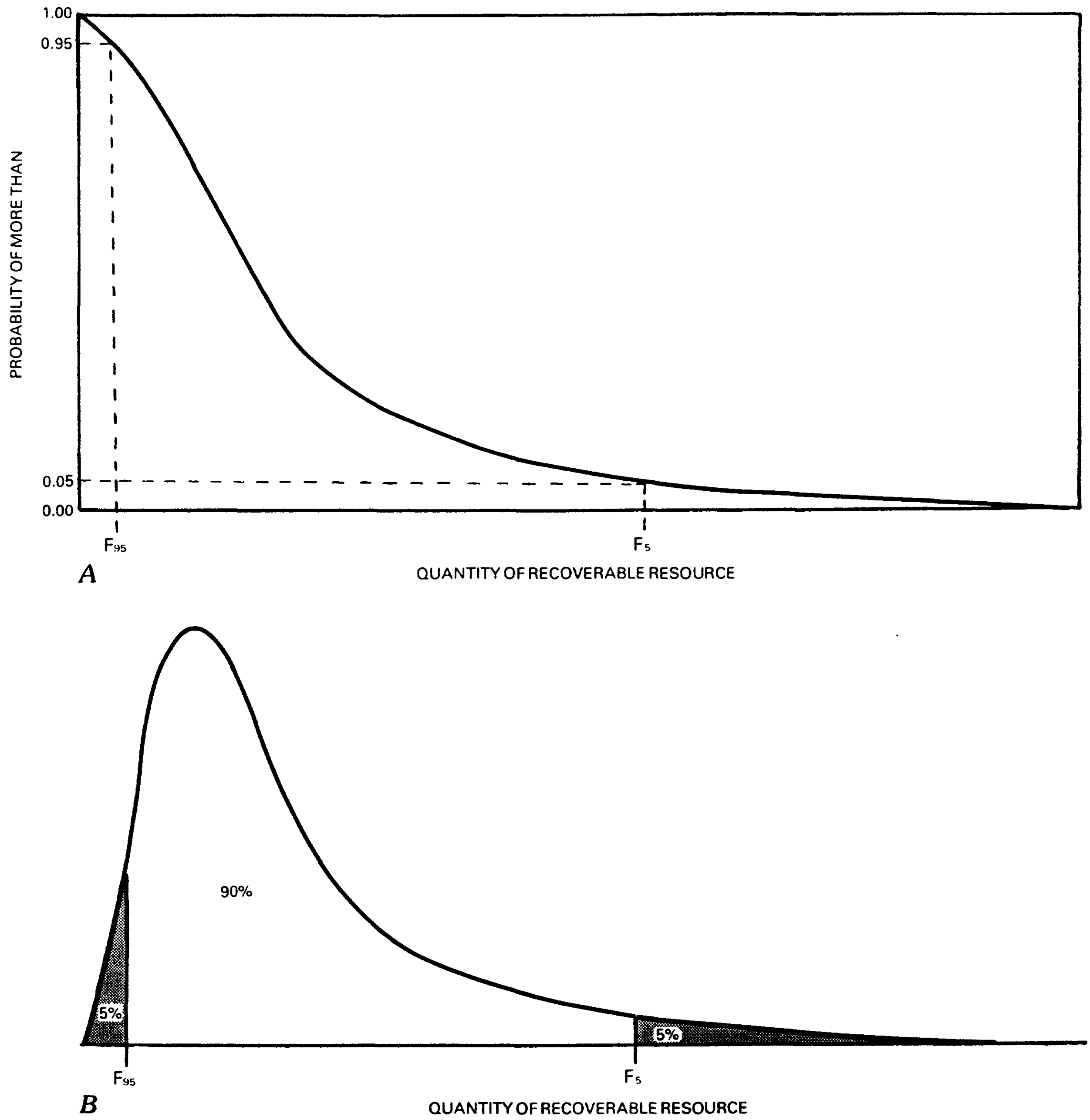
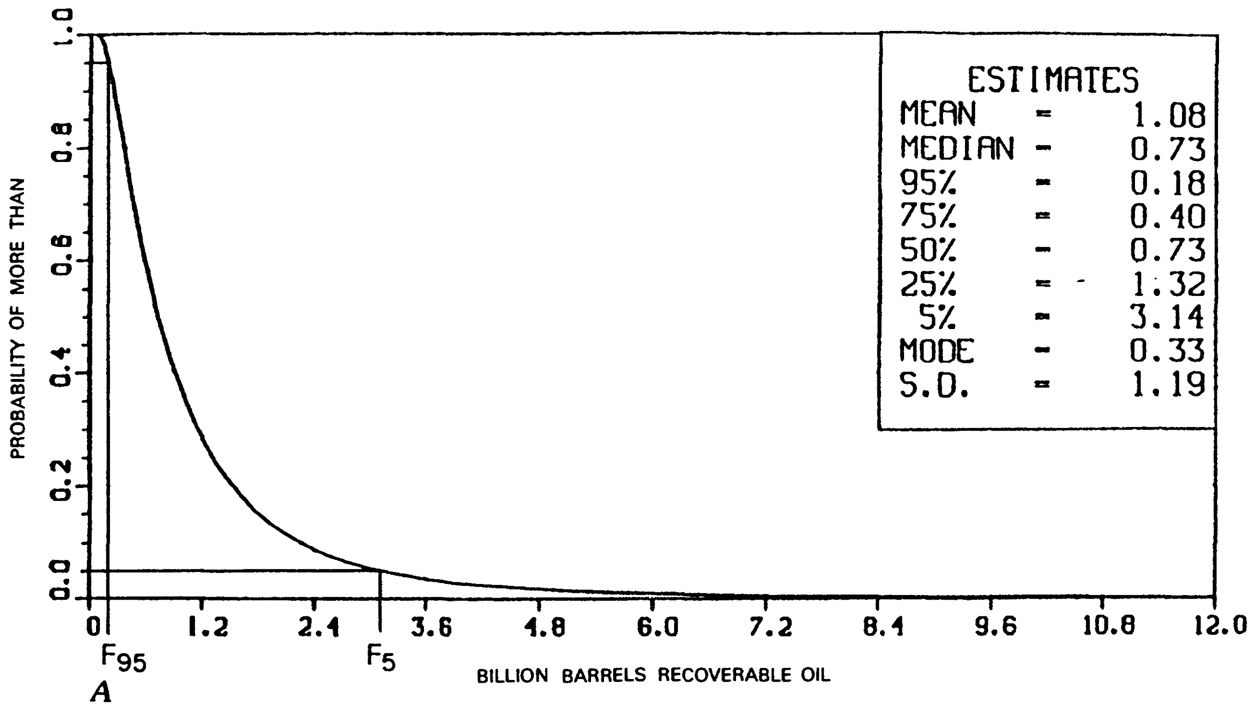


Figure 8.--Typical conditional probability distribution of an undiscovered recoverable resource shown as *A*, conditional *more-than* cumulative distribution function, and *B*, conditional probability density function. F_{95} denotes the 95th fractile; the probability of *more than* the amount is 95 percent. F_5 denotes the 5th fractile; the probability of *more than* the amount is 5 percent.

Source: Dolton and others, 1981



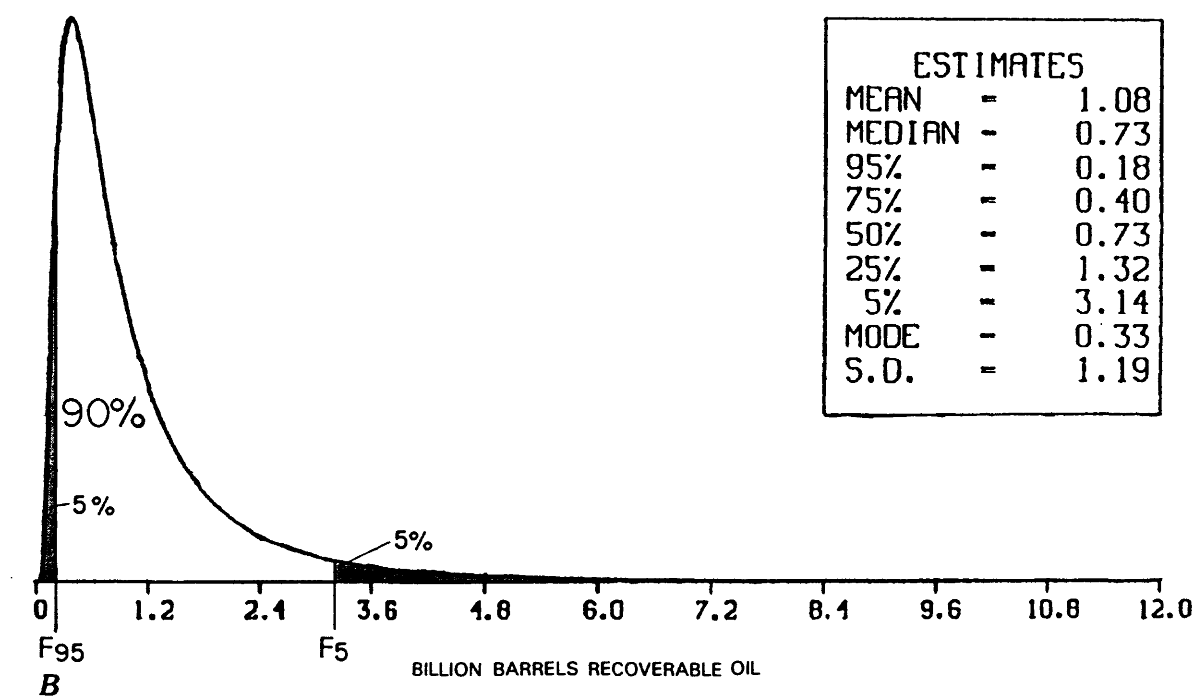
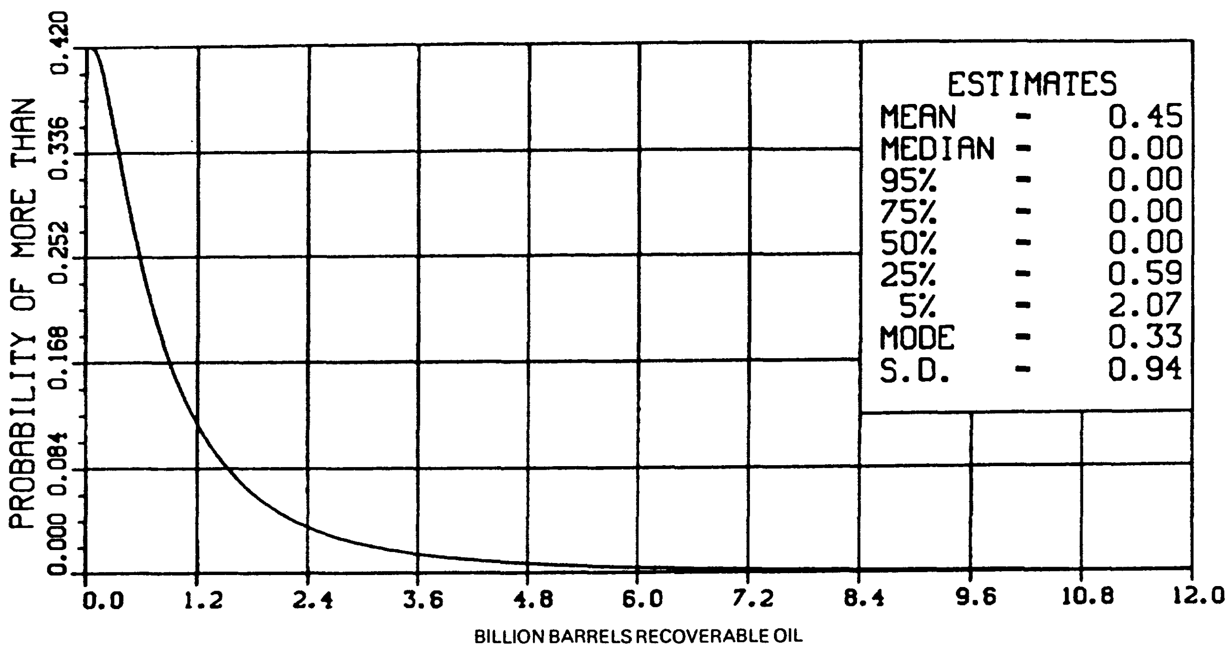


Figure 9.--Conditional probability distribution of the undiscovered recoverable oil for the North Atlantic Shelf province expressed as *A*, conditional *more-than* cumulative distribution function, and *B*, conditional probability density function. Estimates are mean, median, mode, standard deviation (S.D.), and fractiles that correspond to the percentages listed.

Source: Dolton and others, 1981



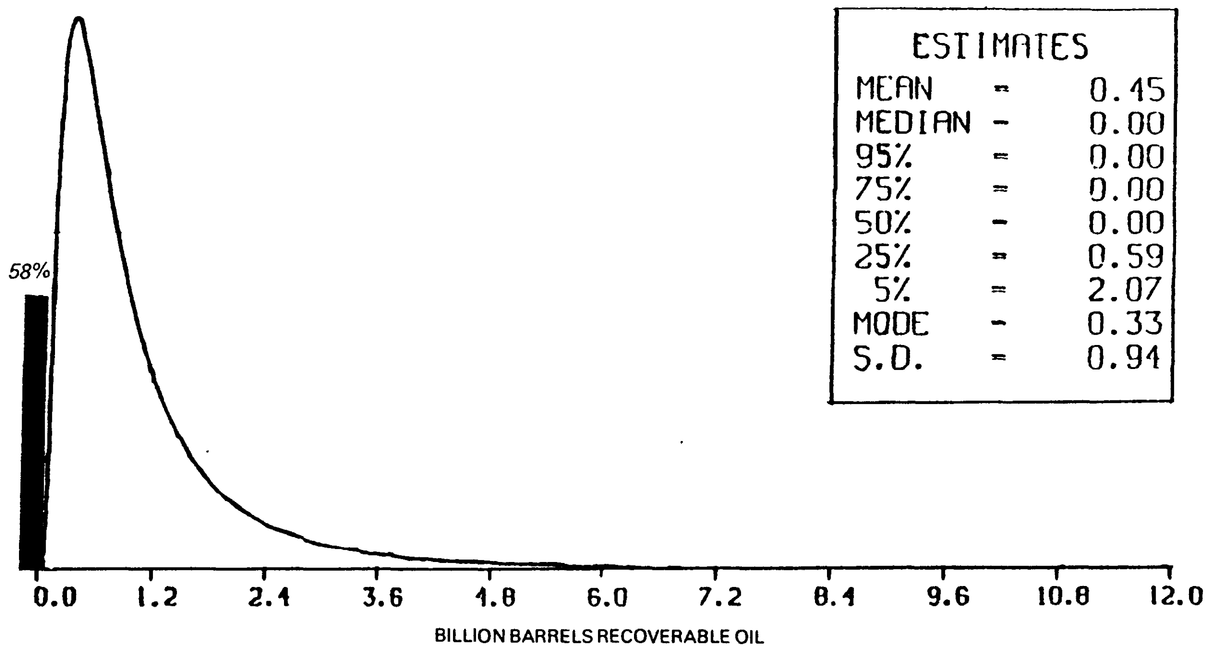


Figure 10.--Probability distribution of the undiscovered recoverable oil for the North Atlantic Shelf province expressed as *A*, *more-than* cumulative distribution function, and *B*, probability density function. *A* has the value of the marginal probability (0.42) at zero resource. *B* has a spike at zero resource of probability weight 1-0.42=0.58 which represents the chance of no recoverable oil being present. Estimates are mean, median, mode, standard deviation (S.D.), and fractiles that correspond to percentages listed.

Source: Dolton and others, 1981

EXAMPLES

Examples of two probability curves on the same graph are (1) conditional and unconditional resource potential, and (2) recoverable and economically recoverable resource potential.

Example 1

LOGRAF is used to duplicate the probability graphs that were originally generated by EXACTDIS for the national assessment of undiscovered conventional oil and gas resources by the U.S. Geological Survey (Mast and others, 1989).

Figure 1a consists of cumulative probability distributions for undiscovered recoverable and undiscovered economically recoverable conventional **crude oil** resources of the United States. Figure 1a' is a summary of the input and output of the assessment, including the lognormal parameters and the conditional and unconditional estimates for each probability curve. The input into LOGRAF are estimates of the following parameters for each distribution:

Recoverable resources--$p = 1$, $\theta = 0$, $F^c_{95} = 33.2$, $F^c_5 = 69.9$

Economically recoverable--$p = 1$, $\theta = 0$, $F^c_{95} = 20.7$, $F^c_5 = 53.8$

and with units of billions of barrels.

Figure 1b consists of cumulative probability distributions for undiscovered recoverable and undiscovered economically recoverable conventional **natural gas** resources of the United States. Figure 1b' is a summary of the input and output of the assessment, including the lognormal parameters and the conditional and unconditional estimates for each probability curve. The input into LOGRAF are estimates of the following parameters for each distribution:

Recoverable resources--$p = 1$, $\theta = 0$, $F^c_{95} = 306.8$, $F^c_5 = 507.2$

Economically recoverable--$p = 1$, $\theta = 0$, $F^c_{95} = 208.2$, $F^c_5 = 325.5$

and with units of trillions of cubic feet.

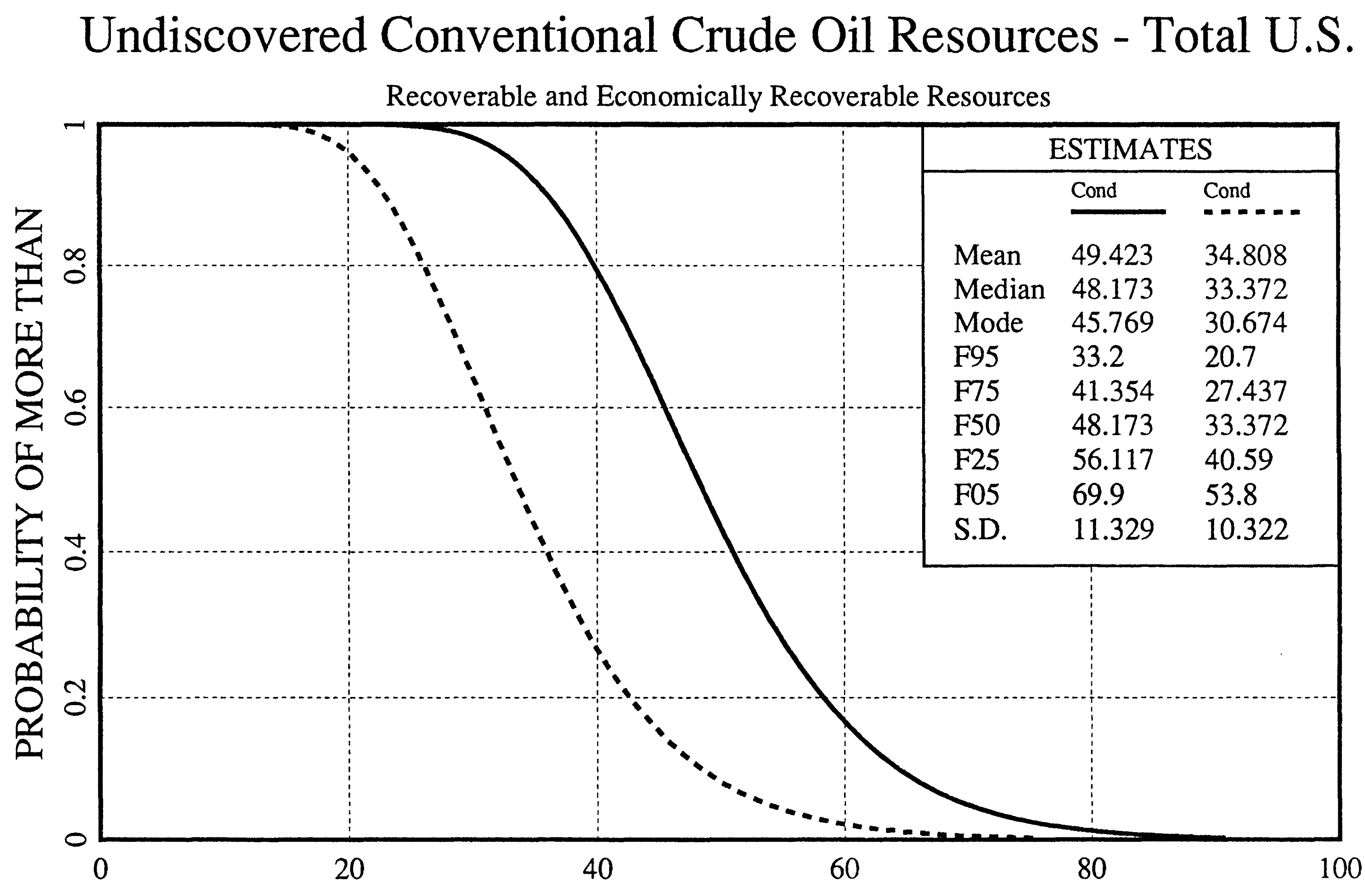


Figure 1a.--LOGRAF output of cumulative probability distributions for undiscovered recoverable (solid curve) and undiscovered economically recoverable (dashed curve) conventional **crude oil** resources of the United States. Both curves are conditional (cond) probability distributions. (Curves duplicated from Figure 10 in Mast and others, 1989.)

LOGRAF 92.7 15-Dec-1992 17:09:09 C:\LOGRAF\GEOTECH\USOIL.DAT

Title : Undiscovered Conventional Crude Oil Resources - Total U.S.
Subtitle : Recoverable and Economically Recoverable Resources
Units : BILLIONS OF BARRELS

INPUT: Probability curve #1

Marginal probability	1
Shift parameter	0
Conditional F95	33.2
Conditional F05	69.9

OUTPUT:

Lognormal parameters

Mu	3.8748
Sigma	0.2263

Conditional estimates

Mean	49.423
Median	48.173
Mode	45.769
F95	33.2
F90	36.045
F75	41.354
F50	48.173
F25	56.117
F10	64.383
F05	69.9
S.D.	11.329

Unconditional estimates *

Mean	49.423
Median	48.173
Mode	45.769
F95	33.2
F90	36.045
F75	41.354
F50	48.173
F25	56.117
F10	64.383
F05	69.9
S.D.	11.329

INPUT: Probability curve #2

Marginal probability	1
Shift parameter	0
Conditional F95	20.7
Conditional F05	53.8

OUTPUT:

Lognormal parameters

Mu	3.5077
Sigma	0.2903

Conditional estimates

Mean	34.808
Median	33.372
Mode	30.674
F95	20.7
F90	23.003
F75	27.437
F50	33.372
F25	40.59
F10	48.415
F05	53.8
S.D.	10.322

Unconditional estimates *

Mean	34.808
Median	33.372
Mode	30.674
F95	20.7
F90	23.003
F75	27.437
F50	33.372
F25	40.59
F10	48.415
F05	53.8
S.D.	10.322

* Because the marginal probability is equal to 1, the unconditional and conditional estimates are equal.

Figure 1a'.--LOGRAF summary of input and output estimates for undiscovered recoverable (curve #1) and undiscovered economically recoverable (curve #2) conventional **crude oil** resources of the United States. For additional output see figure 1a.

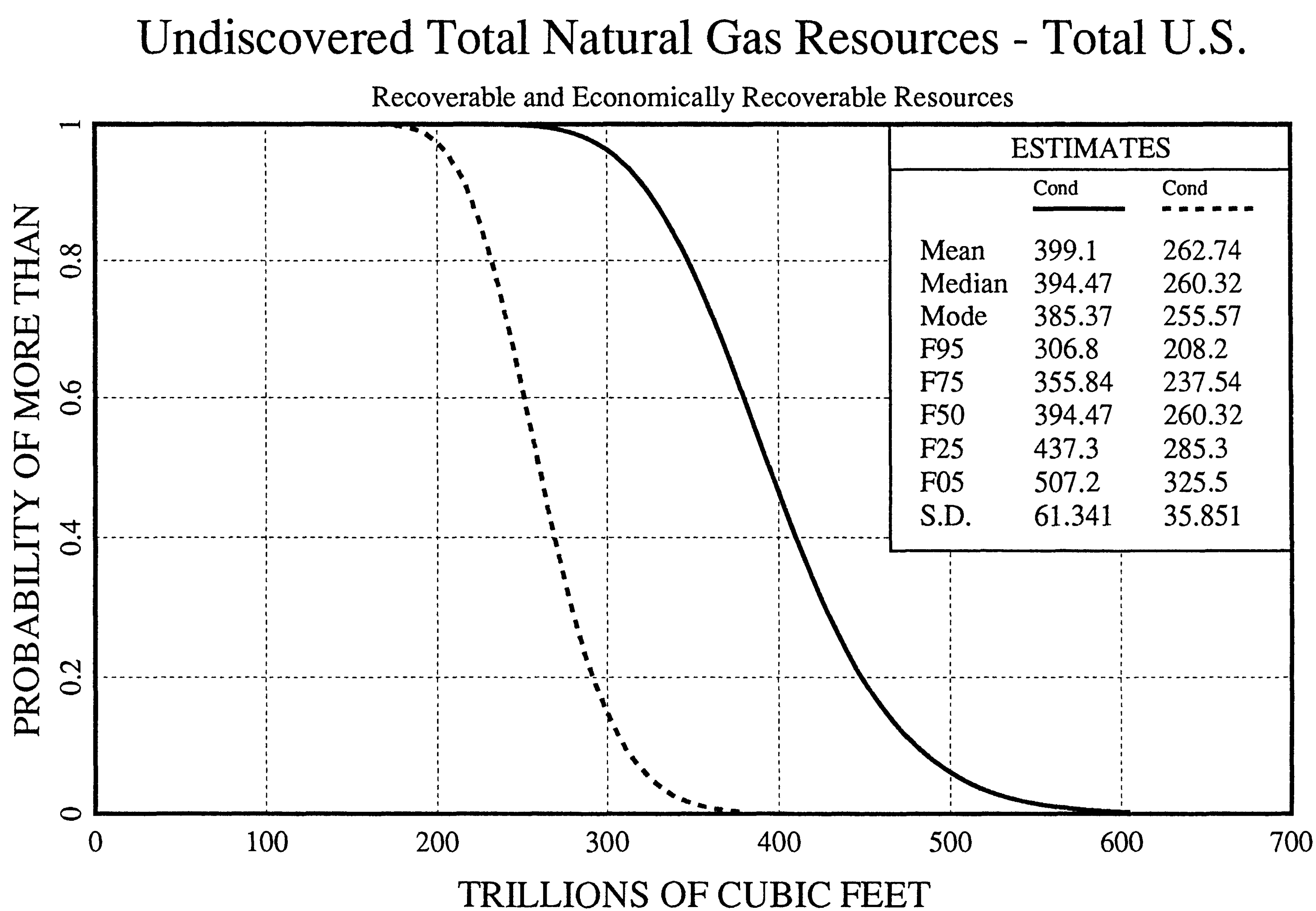


Figure 1b.--LOGRAF output of cumulative probability distributions for undiscovered recoverable (solid curve) and undiscovered economically recoverable (dashed curve) conventional **natural gas** resources of the United States. Both curves are conditional (cond) probability distributions. (Curves duplicated from Figure 11 in Mast and others, 1989.)

LOGRAF 92.7 15-Dec-1992 17:09:16 C:\LOGRAF\GEOTECH\USGAS.DAT

Title : Undiscovered Total Natural Gas Resources - Total U.S.
Subtitle : Recoverable and Economically Recoverable Resources
Units : TRILLIONS OF CUBIC FEET

INPUT: Probability curve #1

Marginal probability	1
Shift parameter	0
Conditional F95	306.8
Conditional F05	507.2

OUTPUT:

Lognormal parameters

Mu	5.9776
Sigma	0.1528

Conditional estimates

Mean	399.1
Median	394.47
Mode	385.37
F95	306.8
F90	324.31
F75	355.84
F50	394.47
F25	437.3
F10	479.81
F05	507.2
S.D.	61.341

Unconditional estimates *

Mean	399.1
Median	394.47
Mode	385.37
F95	306.8
F90	324.31
F75	355.84
F50	394.47
F25	437.3
F10	479.81
F05	507.2
S.D.	61.341

INPUT: Probability curve #2

Marginal probability	1
Shift parameter	0
Conditional F95	208.2
Conditional F05	325.5

OUTPUT:

Lognormal parameters

Mu	5.5619
Sigma	0.1358

Conditional estimates

Mean	262.74
Median	260.32
Mode	255.57
F95	208.2
F90	218.73
F75	237.54
F50	260.32
F25	285.3
F10	309.83
F05	325.5
S.D.	35.851

Unconditional estimates *

Mean	262.74
Median	260.32
Mode	255.57
F95	208.2
F90	218.73
F75	237.54
F50	260.32
F25	285.3
F10	309.83
F05	325.5
S.D.	35.851

* Because the marginal probability is equal to 1, the unconditional and conditional estimates are equal.

Figure 1b'.--LOGRAF summary of input and output estimates for undiscovered recoverable (curve #1) and undiscovered economically recoverable (curve #2) conventional **natural gas** resources of the United States. For additional output see figure 1b.

D. Some Continuous Probability Distributions

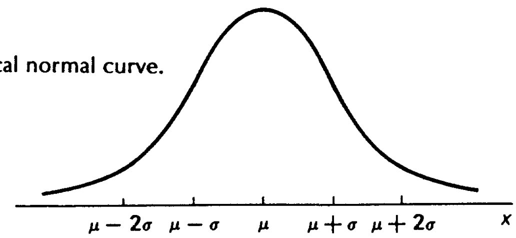
1. Normal distribution

a. Probability density function (pdf)

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(1/2)[(x-\mu)/\sigma]^2} \qquad -\infty < x < \infty$$

Parameters are the mean μ ($-\infty < \mu < \infty$) and standard deviation $\sigma > 0$.

Typical normal curve.



68.26% within $\mu \pm \sigma$

95.44% within $\mu \pm 2\sigma$

99.74% within $\mu \pm 3\sigma$

b. Areas under the normal curve

Theorem: If X has a normal distribution with mean μ and standard deviation σ, then

$$P(X < x) = P(Z < \frac{x-\mu}{\sigma} = z)$$

where Z has the standard normal distribution ($\mu = 0, \sigma = 1$).

Example: Porosity

Suppose the distribution of porosity (%) from a well in the Denver-Julesburg basin of southwestern Nebraska can be modeled as a normal distribution with $\mu = 18\%$ and $\sigma = 3\%$.

X: Porosity (%)

$P(X < 20) = P(Z < \frac{20 - 18}{3} = 0.67) = 0.7486$ from Table A.1

$P(15 < X < 20) = P(X < 20) - P(X < 15) = P(Z < 0.67) - P(Z < -1)$

$= 0.7486 - 0.1587$ from Table A.1

$= 0.5899$

$P(X > 20) = 1 - P(X < 20) = 1 - 0.7486 = 0.2514$

2. PROBDIST model selection menu

Select probability distribution model for TRIAL.DAT

	Model	Min			Ave			Max	Shape or Scale
1	Probability Histogram	F100	F95	F75	F50	F25	F5	F0	
2	Probability Histogram	F100			F50			F0	
3	Normal	F100						F0	
4	Normal				Mean				σ
5	Truncated Normal	F100			Mean			F0	σ
6	Lognormal	F100			F50			F0	
7	Truncated Lognormal	F100			Mean (normal)			F0	σ
8	Exponential	F100						F0	
9	Truncated Exponential	F100						F0	β
10	Pareto	F100						F0	
11	Truncated Pareto	F100						F0	d
12	Uniform	F100						F0	
13	Triangular	F100			Mode			F0	

NOTE: F50 = Median and P (X > F50) = 0.50

MOVE video bar to desired model. RETURN to select, CTRL-G to see sample graph.

PROBDIST 89.11 FIG15.PDL 14:38:33 3-May-1990

Project name : Open File Report
Estimation name : Test data
Units : none
Model : 7-fractile Probability Histogram

INPUT: PARAMETERS

VARIABLE NAME	Min F100	F95	F75	Median F50	F25	F5	Max F0
Sample data	0.00000	1.00000	3.00000	6.00000	14.0000	19.0000	28.0000

OUTPUT: ESTIMATES

VARIABLE NAME	MEAN	S. D.	F100	F95	F75	F50	F25	F5	F0
Sample data	8.52500	6.52745	0.00000	1.00000	3.00000	6.00000	14.0000	19.0000	28.0000

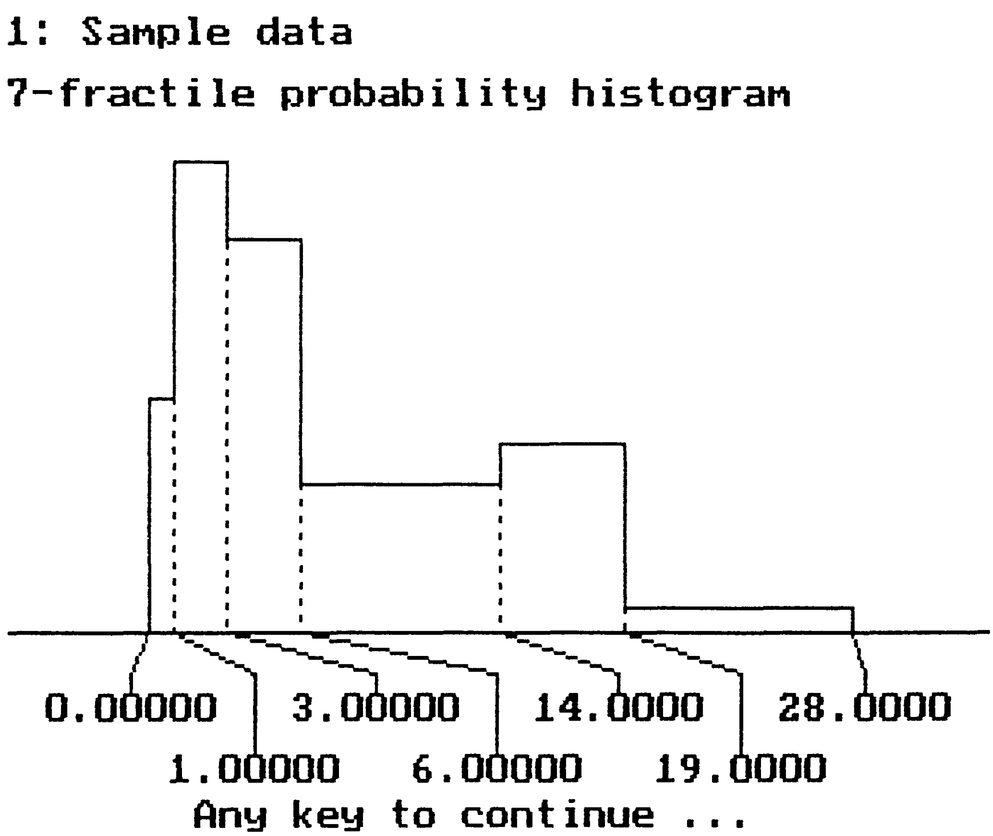


Figure 15. Output of PROBDIST for the 7-fractile histogram model.

PROBDIST 89.11 FIG16.PDL 14:39:16 3-May-1990

Project name : Open File Report
Estimation name : Test data
Units : none
Model : 3-fractile Probability Histogram

INPUT: PARAMETERS

VARIABLE NAME	Min F100	Median F50	Max F0
Sample data	2.00000	8.00000	10.0000

OUTPUT: ESTIMATES

VARIABLE NAME	MEAN	S. D.	F100	F95	F75	F50	F25	F5	F0
Sample data	7.37916	1.75230	2.00000	4.00000	6.50000	8.00000	8.66666	9.33333	10.0000

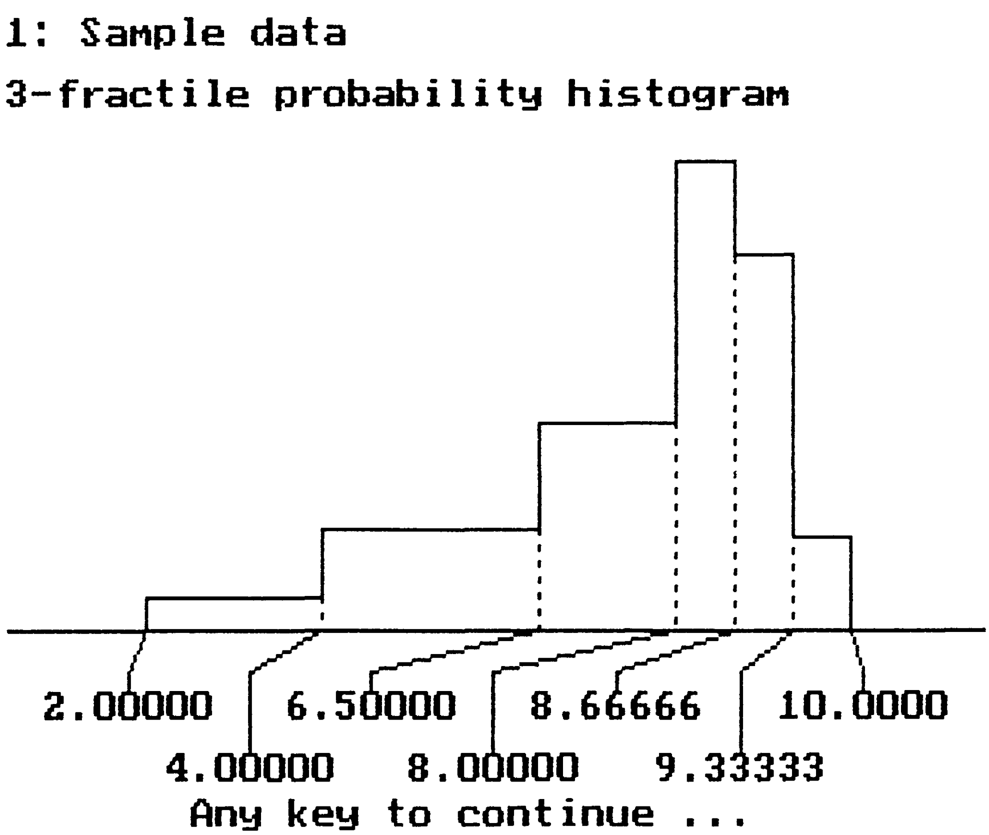


Figure 16. Output of PROBDIST for the 3-fractile histogram model.

PROBDIST 89.11 FIG17.PDL 14:39:43 3-May-1990

Project name : Open File Report
Estimation name : Test data
Units : none
Model : Min/max Normal Distribution

INPUT:

	PARAMETERS	
	Min	Max
VARIABLE NAME	F100	F0
Sample data	2.00000	10.0000

OUTPUT:

	ESTIMATES								
VARIABLE NAME	MEAN	S. D.	F100	F95	F75	F50	F25	F5	F0
Sample data	6.00000	1.46074	2.00000	3.80666	5.10000	6.00000	6.90000	8.19333	10.0000

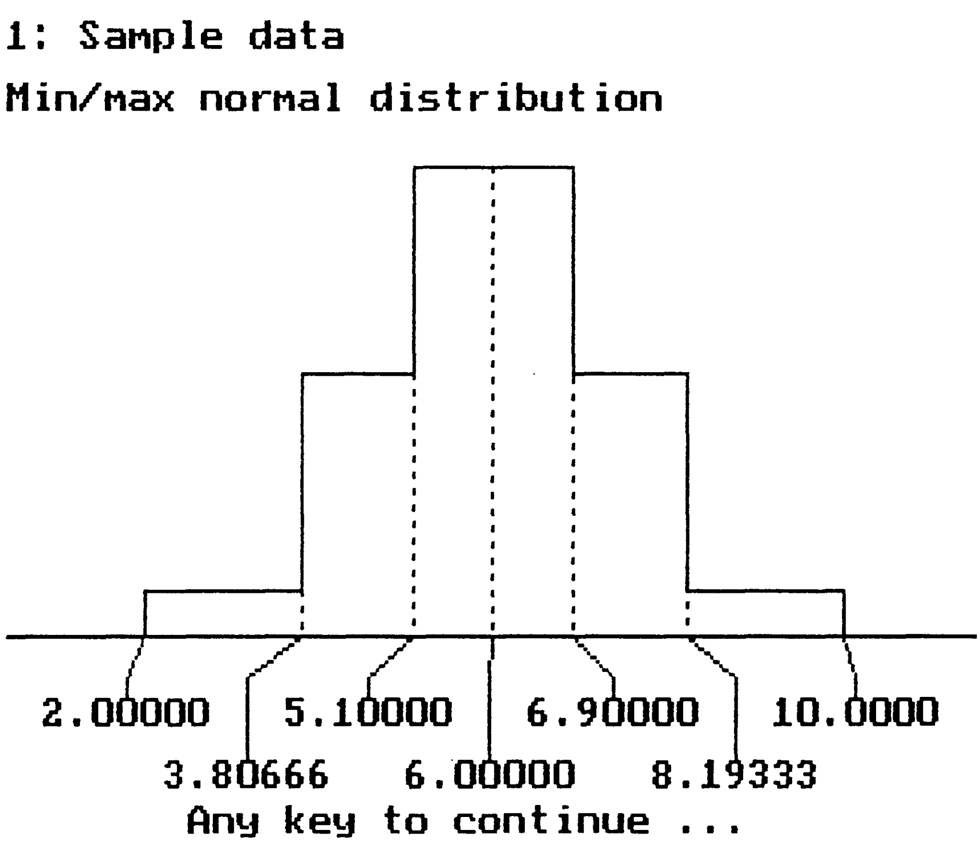


Figure 17. Output of PROBDIST for the minimum/maximum normal distribution model.

PROBDIST 89.11   FIG18.PDL   14:40:08   3-May-1990

Project name : Open File Report
Estimation name : Test data
Units : none
Model : Mean/SD Normal Distribution

INPUT:

VARIABLE NAME	PARAMETERS	
	Mean (Mu)	S.D. (Sigma)
Sample data	5.00000	1.00000

OUTPUT:

VARIABLE NAME	ESTIMATES								
	MEAN	S. D.	F100	F95	F75	F50	F25	F5	F0
Sample data	5.00000	1.09555	2.00000	3.35500	4.32500	5.00000	5.67500	6.64500	8.00000

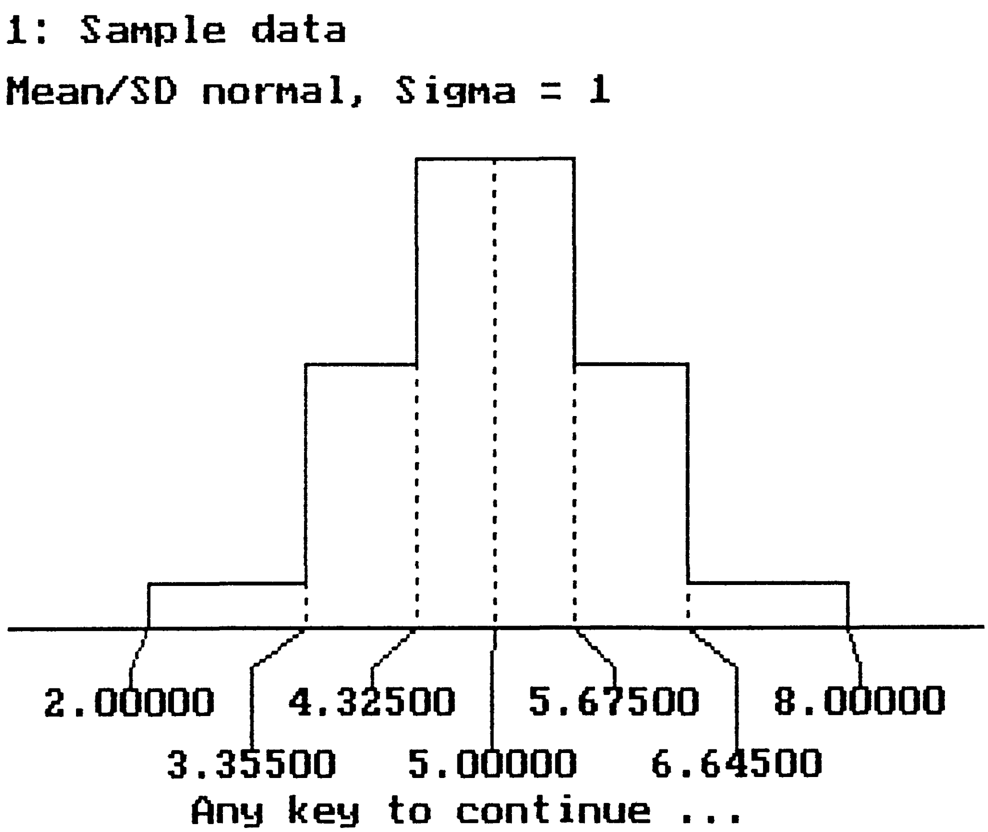


Figure 18. Output of PROBDIST for the mean/standard deviation normal distribution model.

```
PROBDIST  89.11                              FIG19.PDL   14:40:33    3-May-1990

        Project name : Open File Report
     Estimation name : Test data
               Units : none
               Model : Truncated Normal Distribution
```

INPUT:

PARAMETERS

VARIABLE NAME	Min (F100)	Mean (Mu)	Max (F0)	S.D. (Sigma)
Sample data	2.00000	5.00000	10.0000	1.00000

OUTPUT:

ESTIMATES

VARIABLE NAME	MEAN	S. D.	F100	F95	F75	F50	F25	F5	F0
Sample data	5.05290	1.23119	2.00000	3.36504	4.32778	5.00163	5.67623	6.64769	10.0000

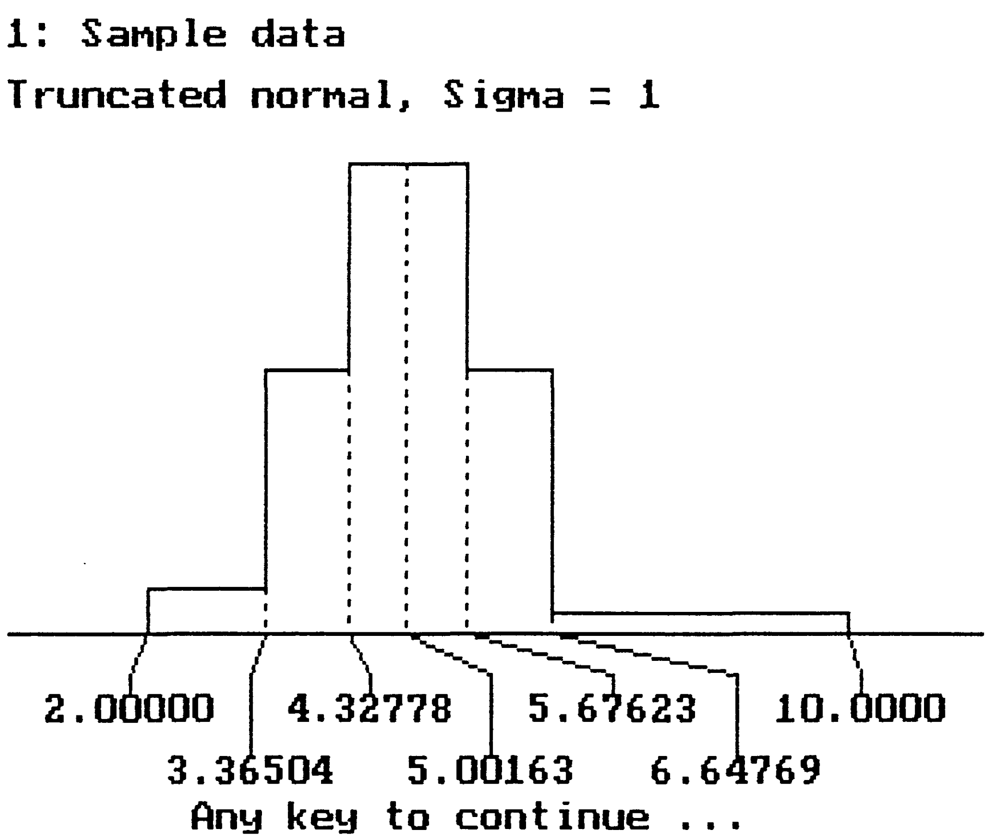


Figure 19. Output of PROBDIST for the truncated normal distribution model.

PROBDIST 89.11 FIG20.PDL 14:40:57 3-May-1990

Project name : Open File Report
Estimation name : Test data
Units : none
Model : Lognormal Distribution

INPUT: PARAMETERS

VARIABLE NAME	Min F100	Median F50	Max F0
Sample data	2.00000	4.00000	10.0000

OUTPUT: ESTIMATES

VARIABLE NAME	MEAN	S. D.	F100	F95	F75	F50	F25	F5	F0
Sample data	4.29568	1.28129	2.00000	2.93519	3.46408	4.00000	4.73208	6.27719	10.0000

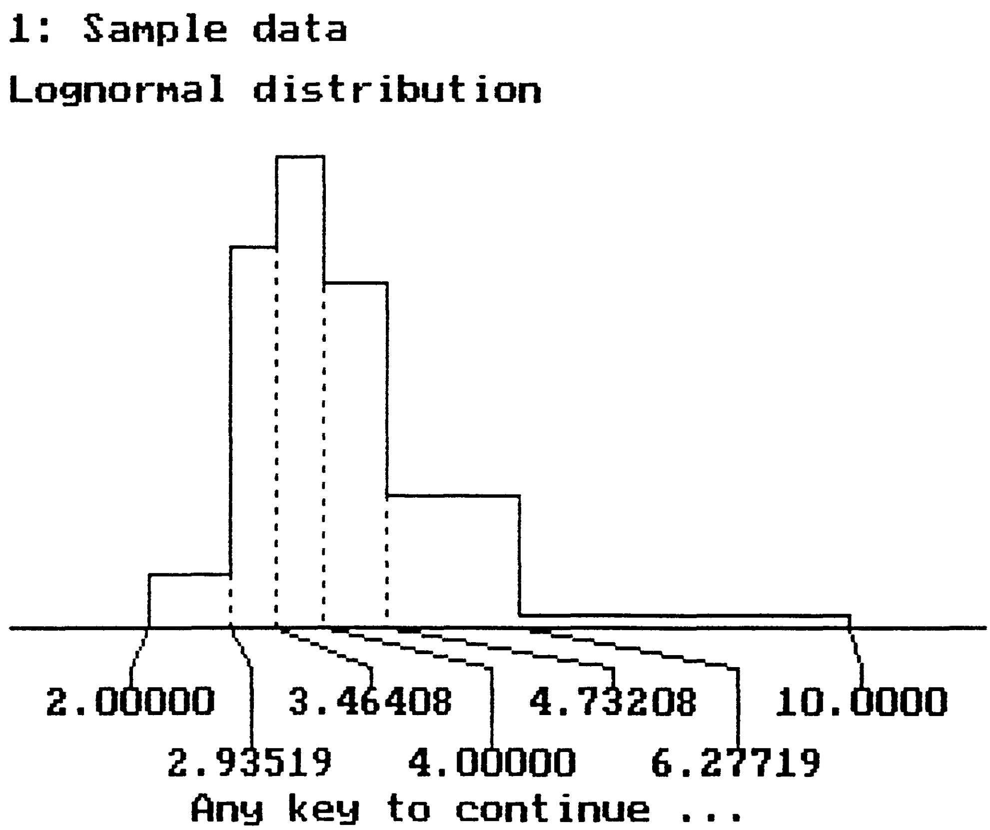


Figure 20. Output of PROBDIST for the lognormal distribution model.

```
PROBDIST  89.11                         FIG21.PDL   14:41:21    3-May-1990

      Project name : Open File Report
   Estimation name : Test data
             Units : none
             Model : Truncated Lognormal Distribution
```

INPUT: PARAMETERS

VARIABLE NAME	Min F100	Normal Mean (Mu)	Max F0	S.D. (Sigma)
Sample data	2.00000	5.00000	500.000	1.20000

OUTPUT: ESTIMATES

VARIABLE NAME	MEAN	S. D.	F100	F95	F75	F50	F25	F5	F0
Sample data	158.630	125.516	2.00000	18.7198	56.6487	117.261	223.345	411.409	500.000

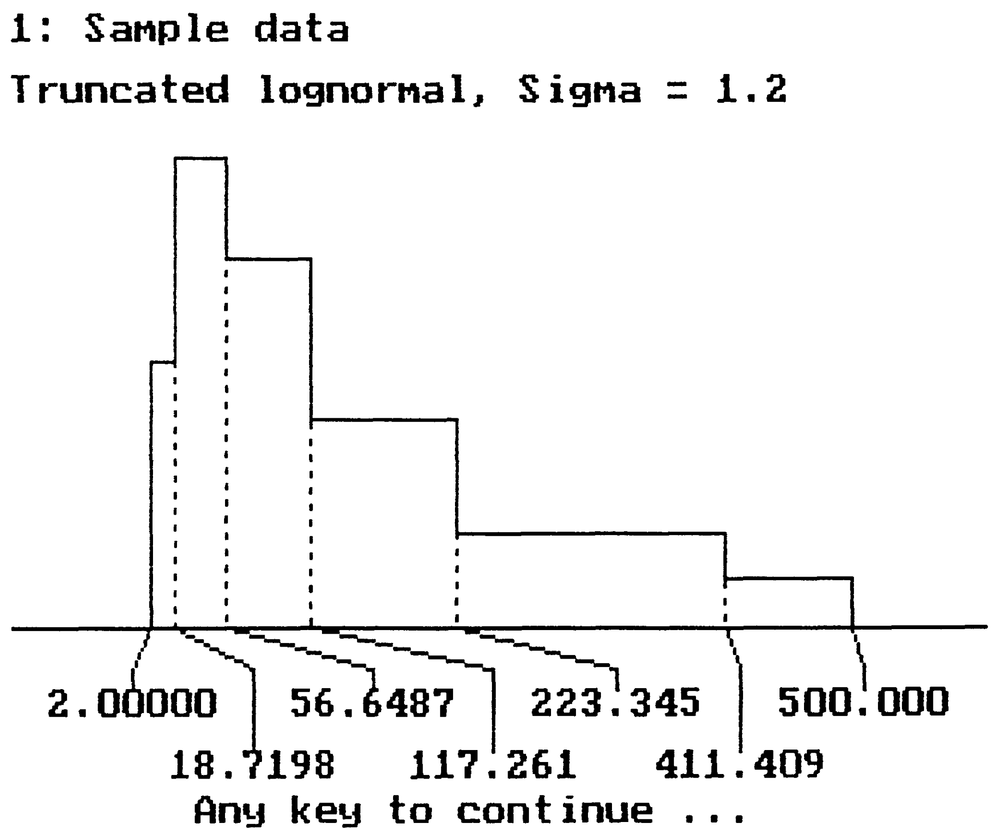


Figure 21. Output of PROBDIST for the truncated lognormal distribution model.

PROBDIST 89.11 FIG22.PDL 14:41:42 3-May-1990

Project name : Open File Report
Estimation name : Test data
Units : none
Model : Exponential Distribution

INPUT:

	PARAMETERS	
	Min	Max
VARIABLE NAME	F100	F0
Sample data	2.00000	10.0000

OUTPUT:

	ESTIMATES								
VARIABLE NAME	MEAN	S. D.	F100	F95	F75	F50	F25	F5	F0
Sample data	3.27798	1.38289	2.00000	2.05940	2.33316	2.80274	3.60549	5.46941	10.0000

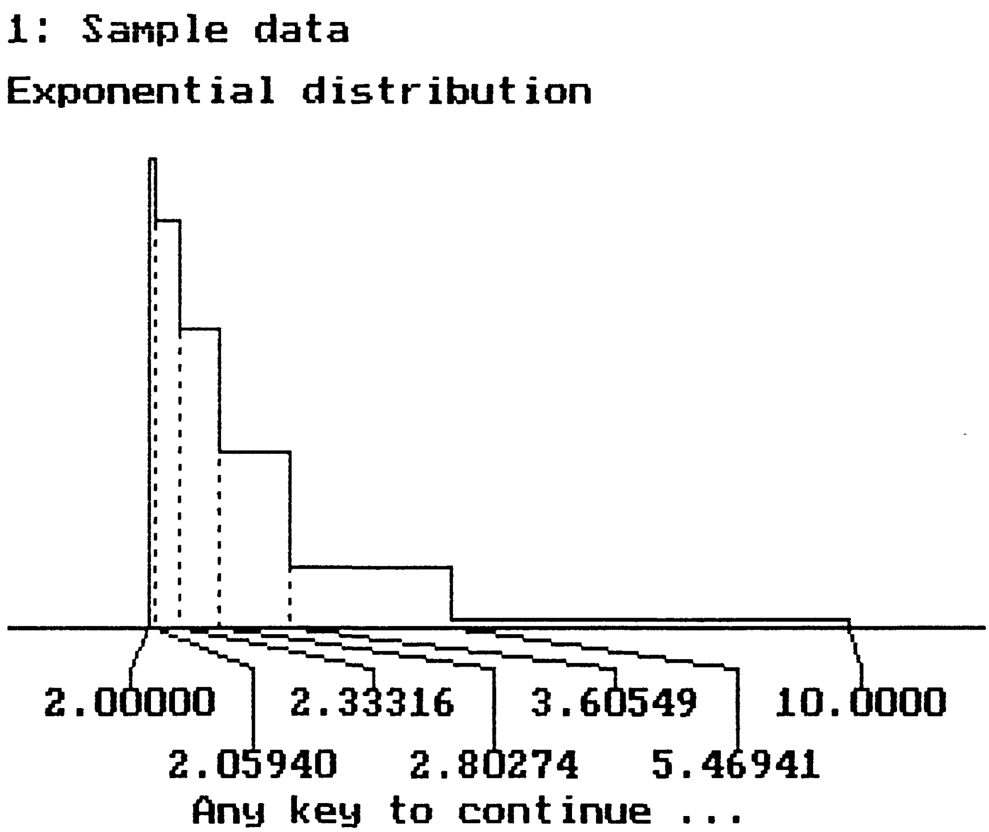


Figure 22. Output of PROBDIST for the exponential distribution model.

PROBDIST 89.11   FIG23.PDL 14:42:06 3-May-1990

Project name : Open File Report
Estimation name : Test data
Units : none
Model : Truncated Exponential Distribution

INPUT:

VARIABLE NAME	Min F100	Max F0	Beta
	PARAMETERS		
Sample data	2.00000	10.0000	3.00000

OUTPUT:

VARIABLE NAME	MEAN	S. D.	F100	F95	F75	F50	F25	F5	F0
					ESTIMATES				
Sample data	4.48180	2.02684	2.00000	2.14292	2.79435	3.87791	5.59086	8.46225	10.0000

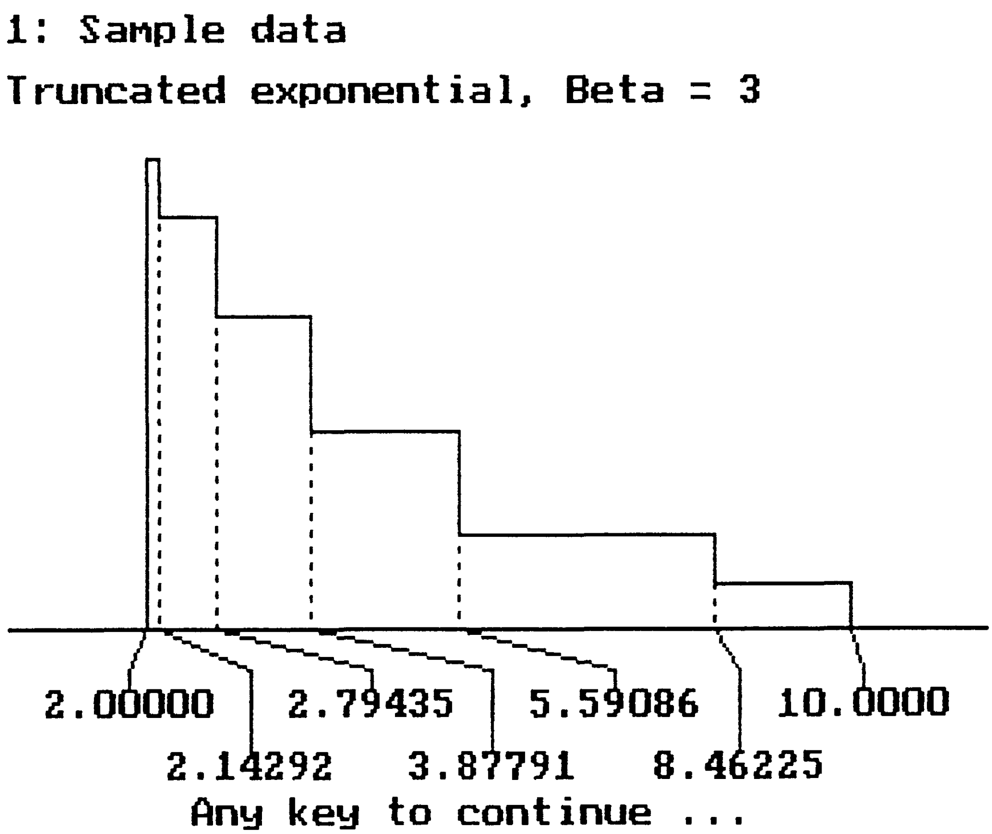


Figure 23. Output of PROBDIST for the truncated exponential distribution model.

PROBDIST 89.11 FIG24.PDL 14:42:35 3-May-1990

Project name : Open File Report
Estimation name : Test data
Units : none
Model : Pareto Distribution

INPUT: PARAMETERS

VARIABLE NAME	Min F100	Max F0
Sample data	2.00000	10.0000

OUTPUT: ESTIMATES

VARIABLE NAME	MEAN	S. D.	F100	F95	F75	F50	F25	F5	F0
Sample data	2.74582	1.16589	2.00000	2.02404	2.13864	2.35053	2.76251	4.01936	10.0000

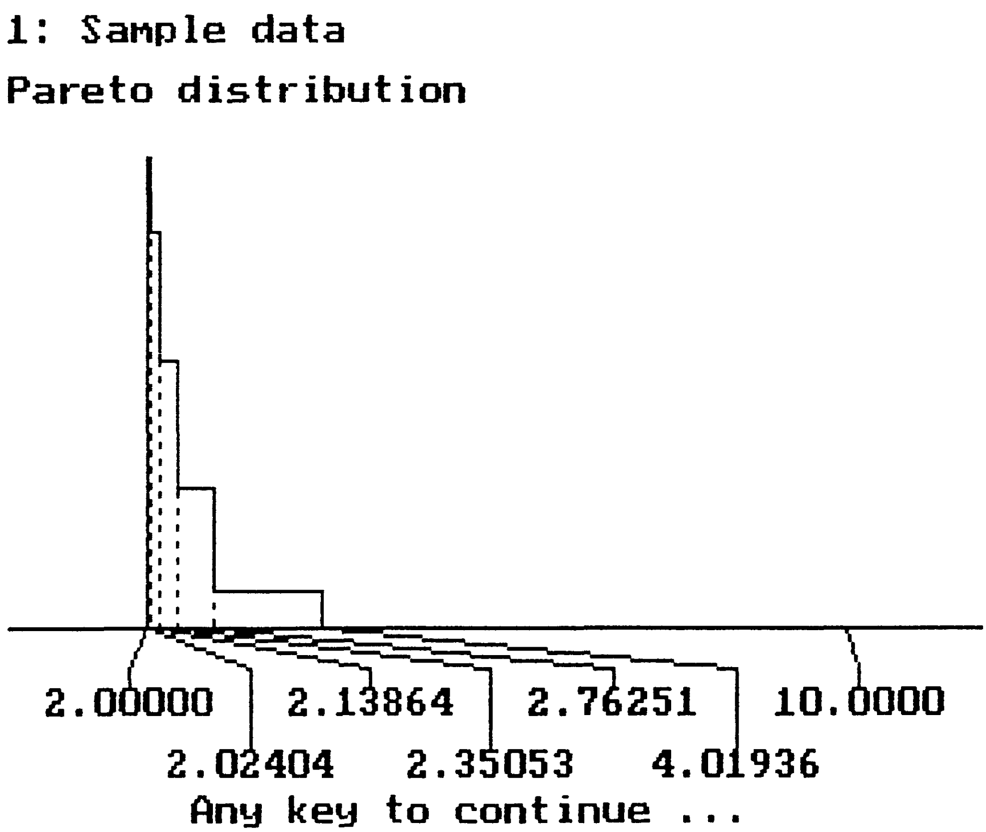


Figure 24. Output of PROBDIST for the Pareto distribution model.

PROBDIST 89.11 FIG25.PDL 14:42:56 3-May-1990

Project name : Open File Report
Estimation name : Test data
Units : none
Model : Truncated Pareto Distribution

INPUT: PARAMETERS

VARIABLE NAME	Min F100	Max F0	d
Sample data	2.00000	10.0000	0.50000

OUTPUT: ESTIMATES

VARIABLE NAME	MEAN	S. D.	F100	F95	F75	F50	F25	F5	F0
Sample data	4.54702	2.17003	2.00000	2.11531	2.69285	3.81966	5.83592	8.86975	10.0000

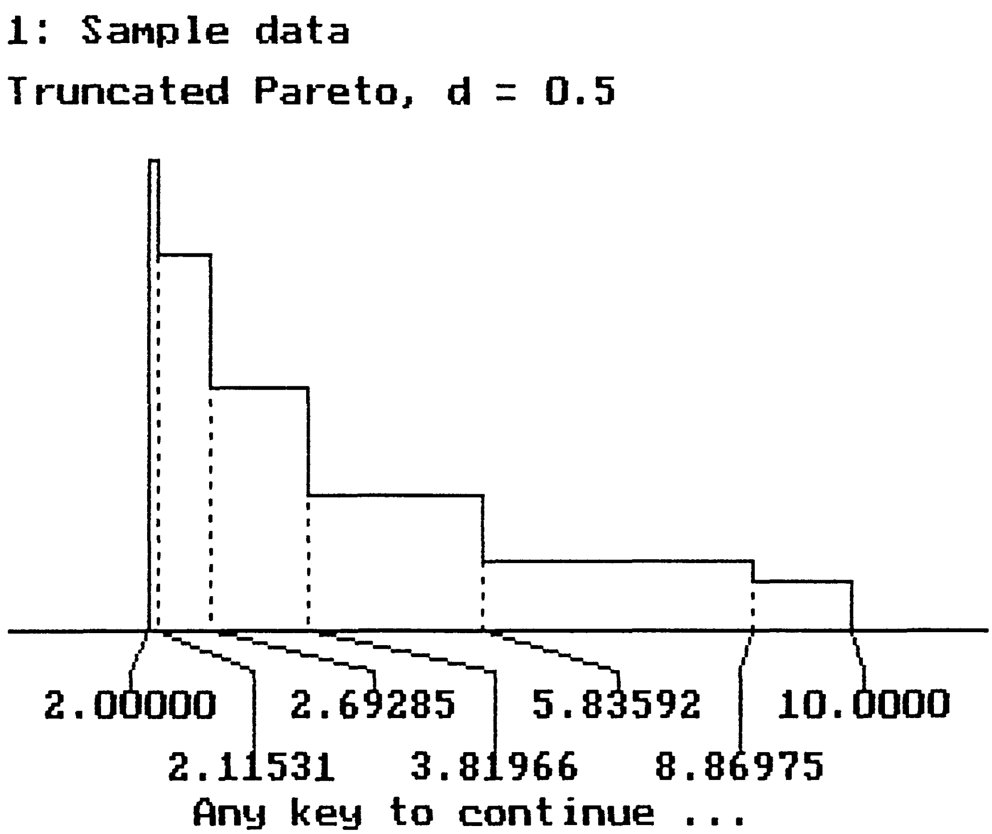


Figure 25. Output of PROBDIST for the truncated Pareto distribution model.

PROBDIST 89.11 FIG26.PDL 14:43:21 3-May-1990

Project name : Open File Report
Estimation name : Test data
Units : none
Model : Uniform Distribution

INPUT: PARAMETERS

VARIABLE NAME	Min F100	Max F0
Sample data	2.00000	10.0000

OUTPUT: ESTIMATES

VARIABLE NAME	MEAN	S. D.	F100	F95	F75	F50	F25	F5	F0
Sample data	6.00000	2.30940	2.00000	2.40000	4.00000	6.00000	8.00000	9.60000	10.0000

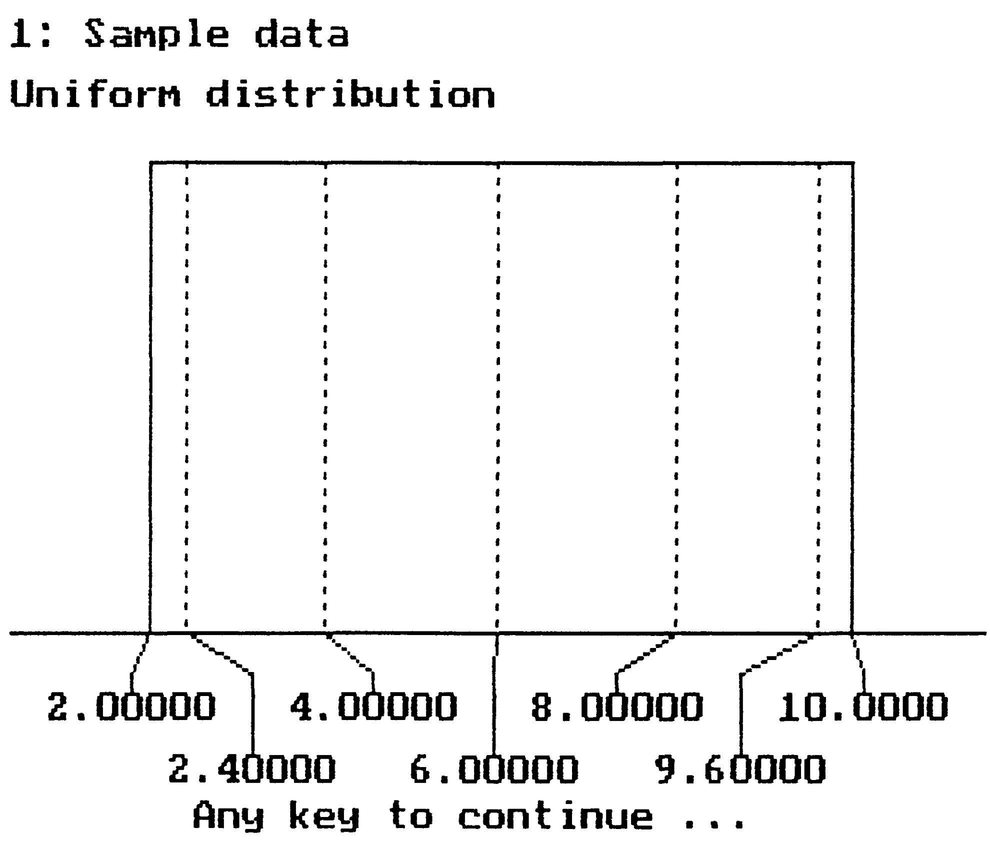


Figure 26. Output of PROBDIST for the uniform distribution model.

PROBDIST 89.11 FIG27.PDL 14:43:47 3-May-1990

Project name : Open File Report
Estimation name : Test data
Units : none
Model : Triangular Distribution

INPUT: PARAMETERS

VARIABLE NAME	Min F100	Mode	Max F0
Sample data	2.00000	4.00000	10.0000

OUTPUT: ESTIMATES

VARIABLE NAME	MEAN	S. D.	F100	F95	F75	F50	F25	F5	F0
Sample data	5.36398	1.78678	2.00000	2.89442	4.00000	5.10102	6.53589	8.45080	10.0000

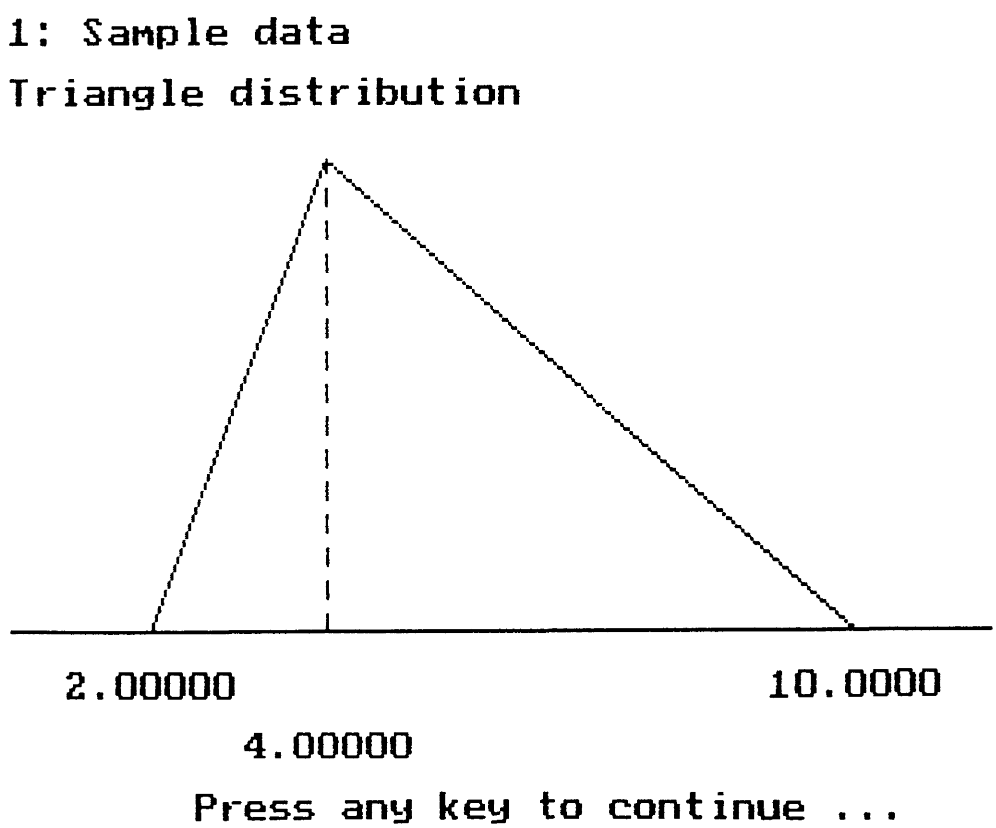


Figure 27. Output of PROBDIST for the triangular distribution model.

STATISTICS

II. Statistics

A. Sampling Concepts

1. Populations

Population of units: a set of units having some common characteristic.

Example: The population of oil fields in a play.

Population of observations: the set of all possible observations or values of a random variable X.

A population of observations is conceptualized from a population of units if each unit (e.g., oil field) were to be measured according to some random variable X (e.g., oil field size).

Population distribution: the probability distribution of a random variable X.

Example: Normal population means a population whose observations are values of a random variable having a normal distribution.

Population size: the number of observations in the population.

Parent population: the population of observations that we are interested in studying.

Example: The population of oil field sizes in a play. The parent population distribution could be modeled as a Pareto distribution.

Sampled population: the population of observations from which a sample is taken.

Sometimes, for various reasons (e.g., financial), the sampled population is more restricted than the parent population.

Example: The population of oil field sizes in a play with the constraint of a point of economic truncation. The sampled population distribution could be modeled as a lognormal distribution.

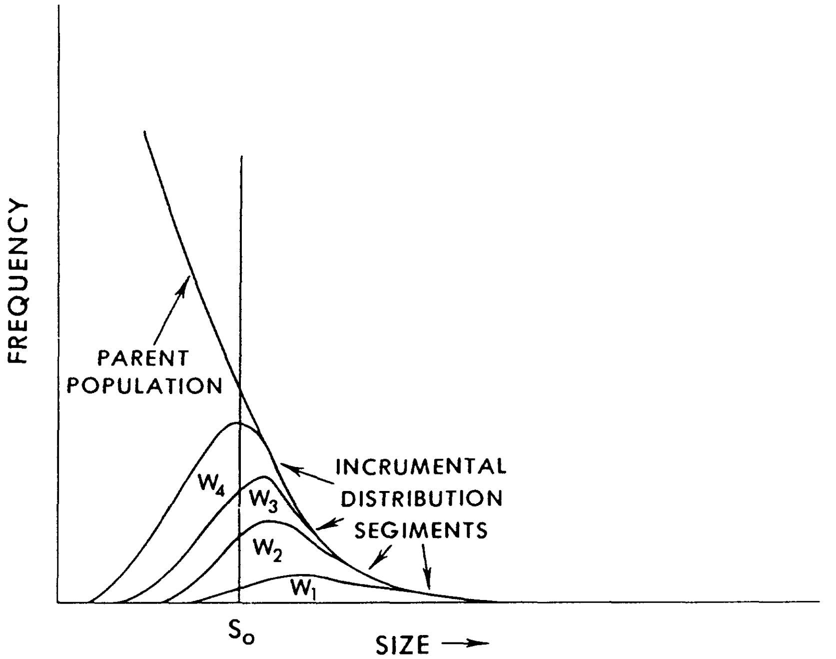


Figure 9.1. The progressive exhaustion of an oil and gas field size distribution by wildcat drilling. S_0 is point of economic truncation. W_1 to W_4 are sequential segments of the field size distribution added with W_1 to W_4 wildcat wells.

Source: Drew, 1990

2. Parameters

Population parameter: a parameter of the population distribution; i.e., any numerical quantity that characterizes or describes a population distribution.

Important: A parameter is a constant or fixed value.

Characterizing parameter: a population parameter that characterizes the population distribution.

Descriptive parameter: a population parameter that is a function of the characterizing parameters and describes the population distribution, e.g., the population mean.

Population mean: the mean μ of the population distribution.

Population variance: the variance σ^2 of the population distribution.

Population standard deviation: the standard deviation σ of the population distribution.

For simplicity,

"parameters" will often refer specifically to the characterizing parameters;

"moments" will refer specifically to the population mean and variance.

a. Binomial population

Population distribution: binomial distribution

Parameters:

Number of prospects: $n = 3$

Probability of field (success): $p = 0.2$

Moments:

Population mean: $\mu = np = (3)(0.2) = 0.6$

Population variance: $\sigma^2 = np(1 - p) = (3)(0.2)(0.8) = 0.48$

b. Uniform population

Population distribution: uniform distribution

Parameters:

Minimum value: $a = 0$

Maximum value: $b = 1$

Moments:

Population mean: $\mu = \frac{a + b}{2} = 0.5$

Population variance: $\sigma^2 = \frac{(b - a)^2}{12} = 0.083$

3. Samples

Sample: a subset of a population.

Sample size: the number of members in a sample.

Sample of units: a subset of n units from a population of units.

Example: A sample of n oil fields in a play.

Sample of observations: a subset of n observations from a sampled population of observations, i.e., a set of n random variables $X_1, X_2, ..., X_n$.

Example: A sample of n oil field sizes in a play.

Physical sample: in geology, a single unit.

Examples: A core sample or a water sample.

Important: Unless otherwise stated, a "sample" means a sample of observations--a set of data.

Two basic types of data

Discrete data: data resulting from a discrete random variable (count data).

Example: Number of new discoveries in each of the plays making up a basin.

Continuous data: data resulting from a continuous random variable (measured data).

Example: Discovered oil field sizes in a play.

Example 1: Oil and mixed oil and gas field sizes

Oil and mixed oil and gas field sizes (in million barrels known recovery) for 175 fields of 1 million BOE or more known recovery in the northern Michigan Silurian reef play are given below.

Source of data is the Significant Oil and Gas Fields of the United States Data Base, a product of NRG Associates, Inc. (1988). The version used included discoveries up to and including 1990. Known recovery refers to the sum of cumulative production plus reserves. Included in the NRG files are those fields with at least 1 million BOE of known recovery and also those smaller, but expected to eventually be revised to at least 1 million BOE.

Let X: Oil or mixed oil and gas field size (million barrels)

Oil and mixed oil and gas field sizes - 1990

0.11	0.74	0.96	1.27	1.82	3.30
0.18	0.75	0.96	1.28	1.85	3.30
0.24	0.75	0.99	1.30	1.90	3.30
0.33	0.75	1.00	1.30	1.90	3.40
0.34	0.77	1.00	1.30	1.95	3.40
0.38	0.78	1.00	1.30	2.00	3.45
0.40	0.78	1.02	1.30	2.00	3.60
0.40	0.80	1.02	1.35	2.08	3.60
0.45	0.80	1.06	1.35	2.10	3.65
0.49	0.80	1.08	1.40	2.25	3.70
0.50	0.82	1.10	1.40	2.35	3.80
0.55	0.83	1.13	1.40	2.35	3.90
0.58	0.83	1.14	1.42	2.38	3.95
0.58	0.84	1.15	1.45	2.40	4.20
0.60	0.85	1.15	1.50	2.40	4.50
0.60	0.85	1.15	1.55	2.40	4.50
0.62	0.86	1.15	1.60	2.45	4.60
0.63	0.87	1.16	1.60	2.60	4.60
0.63	0.88	1.16	1.60	2.60	5.10
0.65	0.90	1.17	1.60	2.60	5.15
0.65	0.90	1.18	1.65	2.70	5.20
0.65	0.90	1.18	1.65	2.75	5.40
0.66	0.91	1.20	1.66	2.75	7.80
0.66	0.91	1.20	1.68	2.80	12.00
0.68	0.93	1.20	1.68	2.90	14.25
0.68	0.93	1.21	1.70	2.95	
0.68	0.93	1.22	1.71	3.00	
0.70	0.95	1.22	1.75	3.00	
0.71	0.95	1.26	1.75	3.00	
0.73	0.96	1.27	1.80	3.15	

Example 2: Gas field sizes

Gas field sizes (in billion cubic feet known recovery) for 61 fields of 1 million BOE or more known recovery in the northern Michigan Silurian reef play are given below.

Source of data is the Significant Oil and Gas Fields of the United States Data Base, a product of NRG Associates, Inc. (1988). The version used included discoveries up to and including 1990. Known recovery refers to the sum of cumulative production plus reserves. Included in the NRG files are those fields with at least 1 million BOE of known recovery and also those smaller, but expected to eventually be revised to at least 1 million BOE.

X: Gas field size (billion cubic feet)

Gas field sizes - 1990

4.44	5.61	6.66	9.00	13.50	25.80
4.50	5.73	6.84	9.00	13.65	30.00
4.50	5.79	6.90	9.15	13.80	33.00
4.80	5.85	6.90	9.60	14.26	33.00
4.95	5.88	6.90	9.90	14.70	35.70
5.16	5.88	6.99	10.05	15.00	46.50
5.28	6.00	7.50	10.20	15.60	
5.29	6.06	7.80	10.80	17.68	
5.40	6.09	7.95	11.09	20.40	
5.40	6.11	8.10	11.85	21.00	
5.55	6.30	8.70	12.90	21.00	

Example 3: Net pay thickness data

Suppose a geologist is studying a new drilling prospect in an area in which 20 wells have been drilled. One of the unknown variables to be considered in the new prospect is net pay thickness. To get an idea of the possible likelihoods and ranges of possible values he has tabulated the net pay thickness values from each of the completed wells, as shown in the table below. Source of data is Newendorp (1975).

X: Net pay thickness (feet)

Net Pay Thickness (Feet) of 20 Wells Completed in a Basin

Well No.	Thickness
1	111
2	81
3	142
4	59
5	109
6	96
7	124
8	139
9	89
10	129
11	104
12	186
13	65
14	95
15	54
16	72
17	167
18	135
19	84
20	154

4. Sampling techniques

The definitions on sampling techniques are stated in terms of observations, but could be stated in terms of units.

Random sampling: a method of selecting a sample of size n from the sampled population such that every possible sample of size n has an equal chance of being selected.

Random sample: a sample that results from random sampling. Alternatively, a set of n independent and identically distributed random variables $X_1, X_2, ..., X_n$ each having the same population distribution.

Sampling with replacement: sampling in which each observation of a sampled population can be selected more than once.

Sampling without replacement: sampling in which each observation of a sampled population cannot be selected more than once.

Stratified random sampling: the sampled population is divided into subpopulations and a random sample is taken from each subpopulation. Example: A series of random samples taken independently from various strata or depths.

Sampling proportional to size: biased sampling in which the chance of being selected is proportional to the size of the unit.

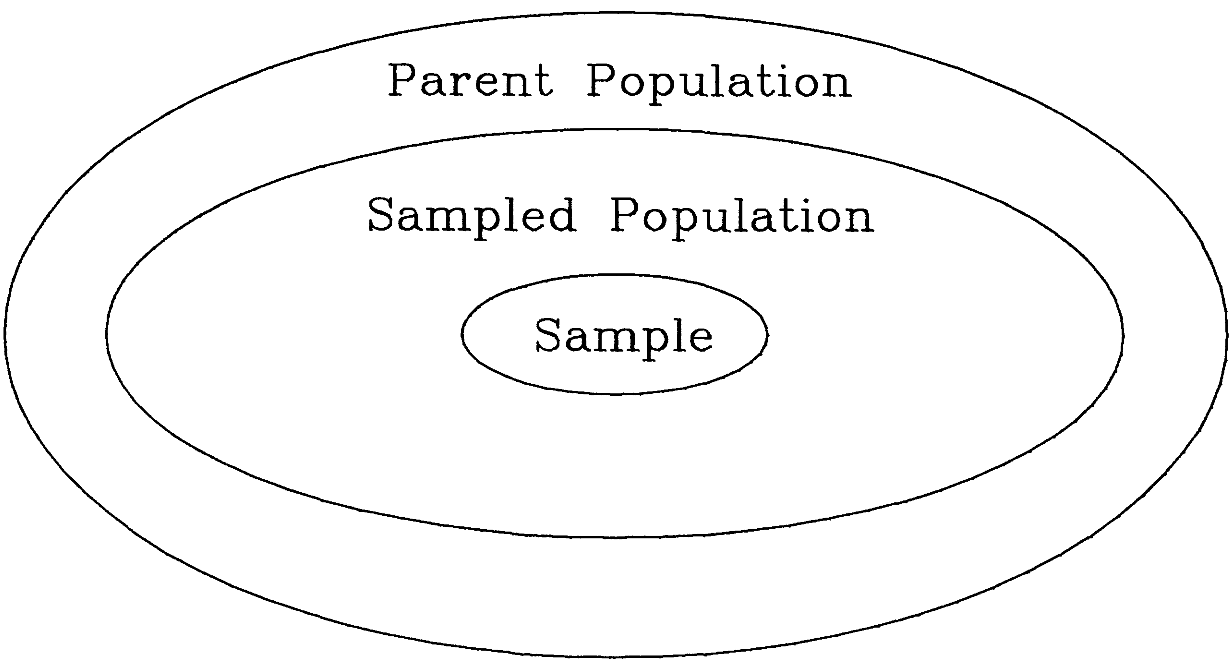
The discovery process modeling approach suggested by Barouch and Kaufman (1976) and Arps and Roberts (1958) relies on the following postulates:

1. The discovery of pools within an area of exploratory interest can be modeled statistically as sampling without replacement from an underlying population of pools.

2. The discovery of a particular pool within the available population of undiscovered pools is random with the probability of discovery being proportional to the areal extent of the pool.

This introduces a sampling bias toward the largest pools and produces a result where the largest pools are found more quickly.

If the sampled population is different from the parent population, then a random sample from the sampled population will be a biased sample from the parent population.



5. Statistics

Statistic: any function of the observations comprising a sample; i.e., any numerical quantity that is calculated from a sample $X_1, X_2, \ldots, X_n$. In general, a statistic Y can be expressed in functional notation as

$Y = g(X_1, X_2, \text{---}, X_n)$

Important: A statistic is a random variable, because a function of random variables is a random variable.

The reason the term statistic is defined so broadly, and not simply as a numerical quantity "describing" a sample, is that in statistical inference we calculate numerical values from the sample for the purpose of making inferences concerning the *population*, and not necessarily to describe the *sample*. It is important to realize that a parameter possesses a fixed value, whereas a statistic can assume one of many possible values since it depends on the sample; that is, the value of a statistic varies from sample to sample.

Example: Oil and mixed oil and gas field sizes.

Oil and mixed oil and gas field sizes (in million barrels known recovery) for 175 fields of 1 million BOE or more known recovery in the northern Michigan Silurian reef play. Three statistics are

Minimum value: $X_{min} = 0.11$

Maximum value: $X_{max} = 14.25$

Total: $\sum_{i=1}^{175} X_i = 317.56$

In summary, we are interested in the population parameters and sample statistics (defined later) shown in the chart.

Parameters	Statistics
μ population mean	$\overline{X}$ sample mean
σ^2 population variance	S^2 sample variance
σ population standard deviation	S sample standard deviation

<u>Sampling distribution</u>: the probability distribution of a statistic.

Example: The probability distribution of $\overline{X}$ is called the sampling distribution of the mean.

B. Descriptive Statistics

1. Tabular methods

Notation: $[x_1, x_2)$ means $x_1 \leq x < x_2$

a. Frequency distribution

X:Oil or mixed oil and gas field size (million barrels)

Frequency distribution of oil and mixed oil and gas field sizes

Class Interval	Frequency f
1. [0, 0.5)	10
2. [0.5, 1.0)	53
3. [1.0, 1.5)	41
4. [1.5, 2.0)	21
5. [2.0, 2.5)	12
6. [2.5, 3.0)	9
7. [3.0, 3.5)	10
8. [3.5, 4.0)	7
9. [4.0, 4.5)	1
10. [4.5, 5.0)	4
11. [5.0, 5.5)	4
12. [5.5, 15)	3
	n = 175

b. Relative frequency distribution

Relative frequency distribution of oil and mixed oil and gas field sizes

Class Interval	Frequency f	Relative Frequency f/n	Relative Frequency %
1. [0, 0.5)	10	10/175 = 0.06	6%
2. [0.5, 1.0)	53	53/175 = 0.30	30%
3. [1.0, 1.5)	41	41/175 = 0.23	23%
4. [1.5, 2.0)	21	21/175 = 0.12	12%
5. [2.0, 2.5)	12	12/175 = 0.07	7%
6. [2.5, 3.0)	9	9/175 = 0.05	5%
7. [3.0, 3.5)	10	10/175 = 0.06	6%
8. [3.5, 4.0)	7	7/175 = 0.04	4%
9. [4.0, 4.5)	1	1/175 = 0.01	1%
10. [4.5, 5.0)	4	4/175 = 0.02	2%
11. [5.0, 5.5)	4	4/175 = 0.02	2%
12. [5.5, 15)	3	3/175 = 0.02	2%
	n = 175	1.00	100%

c. Cumulative frequency distribution (less than)

Cumulative frequency distribution (less than) of oil and mixed oil and gas field sizes

Class Interval	Frequency f	Cumulative Frequency (less than)
1. [0, 0.5)	10	10
2. [0.5, 1.0)	53	63
3. [1.0, 1.5)	41	104
4. [1.5, 2.0)	21	125
5. [2.0, 2.5)	12	137
6. [2.5, 3.0)	9	146
7. [3.0, 3.5)	10	156
8. [3.5, 4.0)	7	163
9. [4.0, 4.5)	1	164
10. [4.5, 5.0)	4	168
11. [5.0, 5.5)	4	172
12. [5.5, 15)	3	175
	n = 175	

d. Relative cumulative frequency distribution (less than)

Also called cumulative proportion and cumulative percentage.

Relative cumulative frequency distribution (less than) of oil and mixed oil and gas field sizes

Class Interval	Frequency f	Cumulative Frequency (less than)	Relative Cumulative Frequency Proportion	Relative Cumulative Frequency Percentage
1. [0, 0.5)	10	10	0.06	6%
2. [0.5, 1.0)	53	63	0.36	36%
3. [1.0, 1.5)	41	104	0.59	59%
4. [1.5, 2.0)	21	125	0.71	71%
5. [2.0, 2.5)	12	137	0.78	78%
6. [2.5, 3.0)	9	146	0.83	83%
7. [3.0, 3.5)	10	156	0.89	89%
8. [3.5, 4.0)	7	163	0.93	93%
9. [4.0, 4.5)	1	164	0.94	94%
10. [4.5, 5.0)	4	168	0.96	96%
11. [5.0, 5.5)	4	172	0.98	98%
12. [5.5, 15)	3	175	1.00	100%
	n = 175			

e. Cumulative frequency distribution (more than)

Cumulative frequency distribution (more than) of oil and mixed oil and gas field sizes

Class Interval	Frequency f	Cumulative Frequency (more than)
1. [0, 0.5)	10	175
2. [0.5, 1.0)	53	165
3. [1.0, 1.5)	41	112
4. [1.5, 2.0)	21	71
5. [2.0, 2.5)	12	50
6. [2.5, 3.0)	9	38
7. [3.0, 3.5)	10	29
8. [3.5, 4.0)	7	19
9. [4.0, 4.5)	1	12
10. [4.5, 5.0)	4	11
11. [5.0, 5.5)	4	7
12. [5.5, 15)	3	3
	n = 175	

f. Relative cumulative frequency distribution (more than)

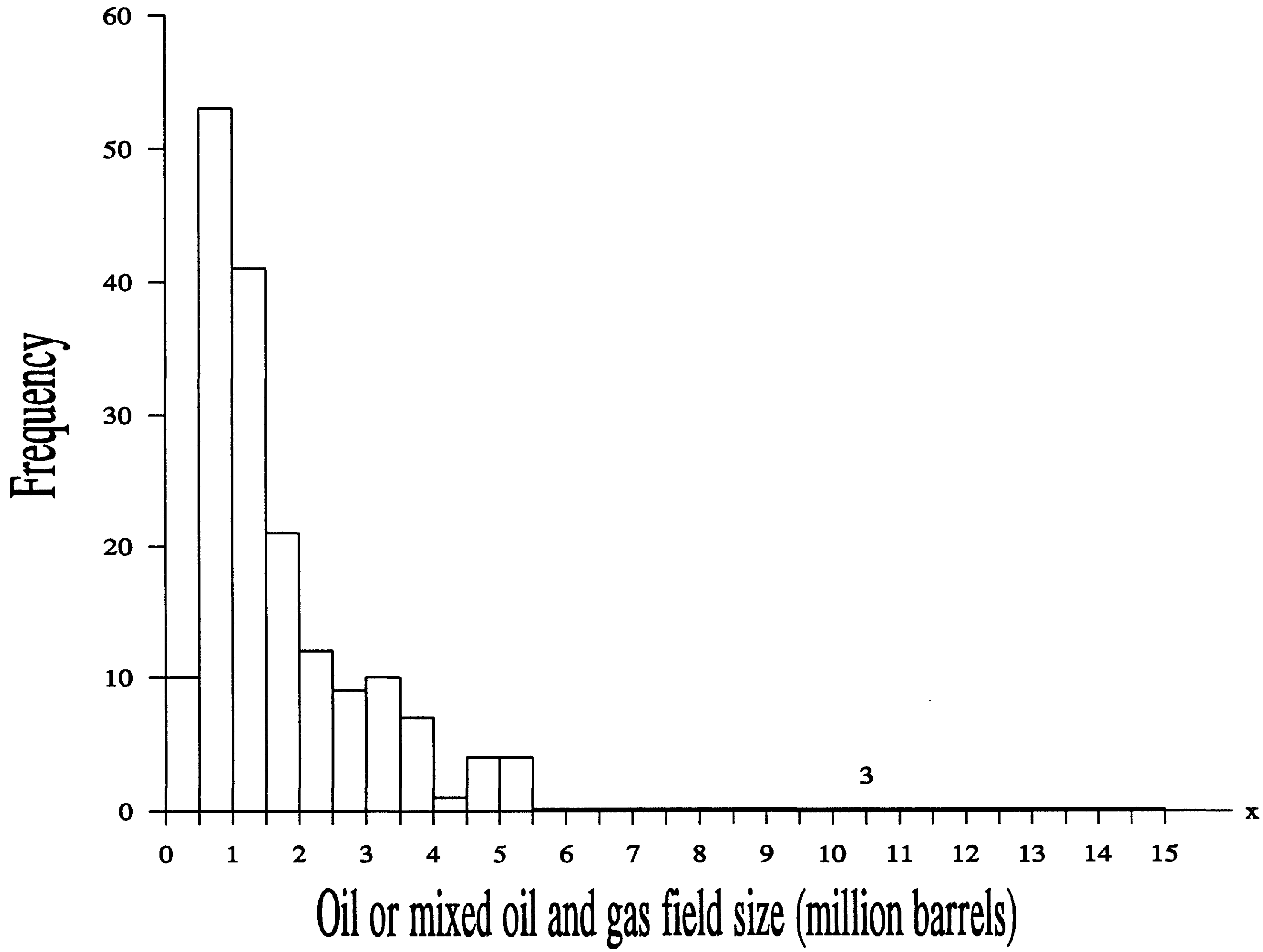
Also called cumulative proportion and cumulative percentage.

Relative cumulative frequency distribution (more than) of oil and mixed oil and gas field sizes

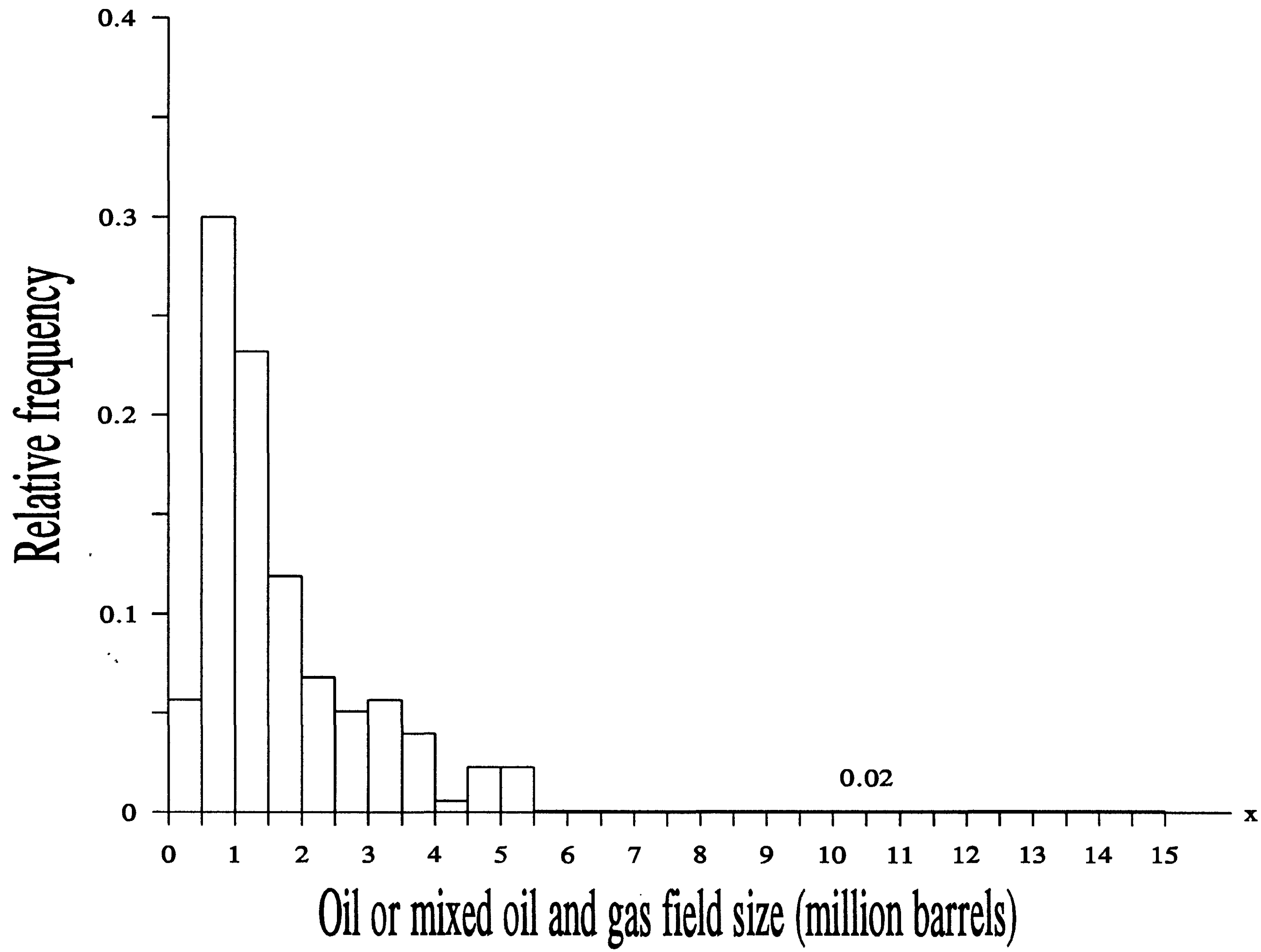
Class Interval	Frequency f	Cumulative Frequency (more than)	Relative Cumulative Frequency Proportion	Relative Cumulative Frequency Percentage
1. [0, 0.5)	10	175	1.00	100%
2. [0.5, 1.0)	53	165	0.94	94%
3. [1.0, 1.5)	41	112	0.64	64%
4. [1.5, 2.0)	21	71	0.41	41%
5. [2.0, 2.5)	12	50	0.29	29%
6. [2.5, 3.0)	9	38	0.22	22%
7. [3.0, 3.5)	10	29	0.17	17%
8. [3.5, 4.0)	7	19	0.11	11%
9. [4.0, 4.5)	1	12	0.07	7%
10. [4.5, 5.0)	4	11	0.06	6%
11. [5.0, 5.5)	4	7	0.04	4%
12. [5.5, 15)	3	3	0.02	2%
	n = 175			

2. Pictorial methods

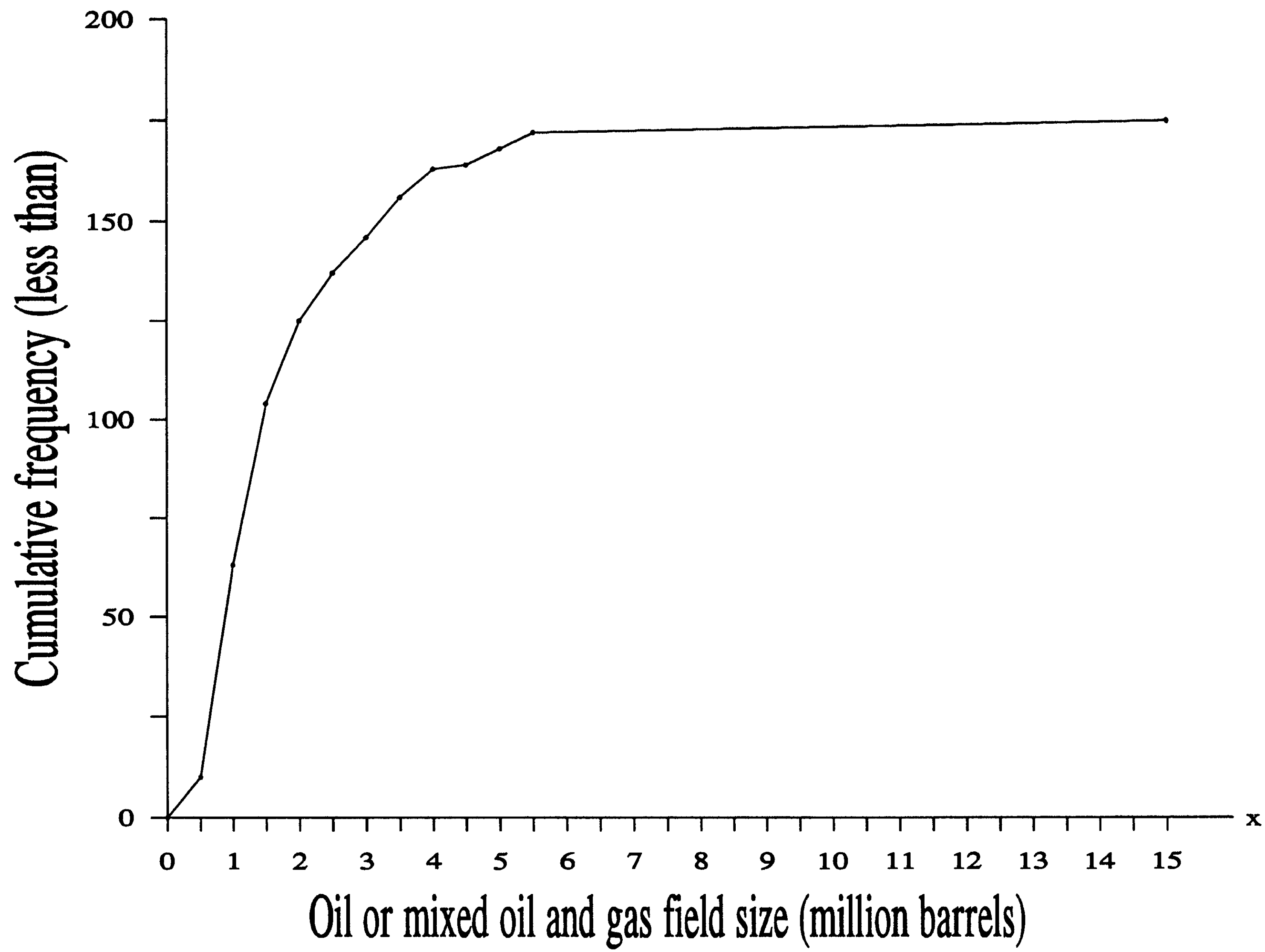
a. Frequency histogram



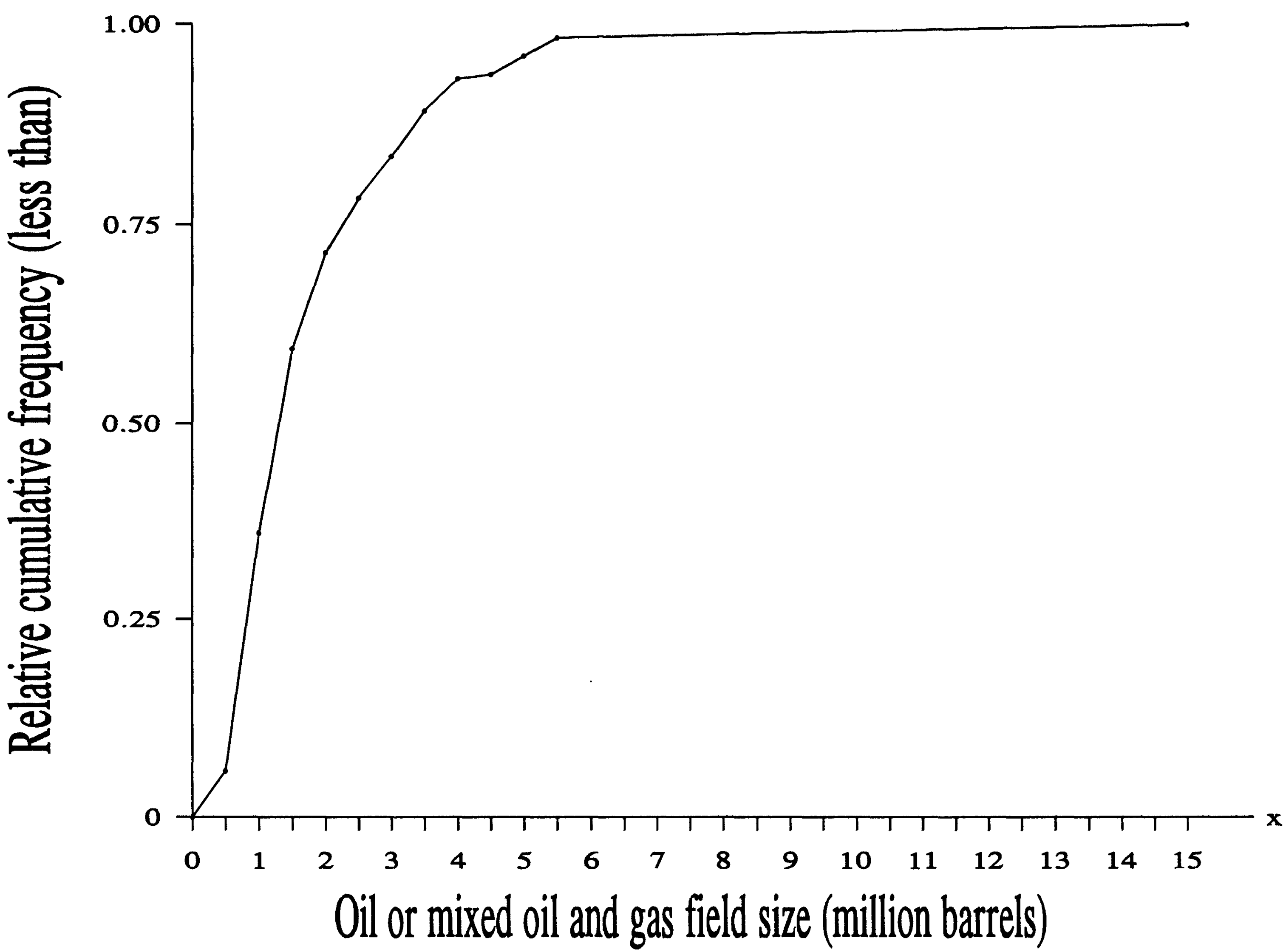
b. Relative frequency histogram



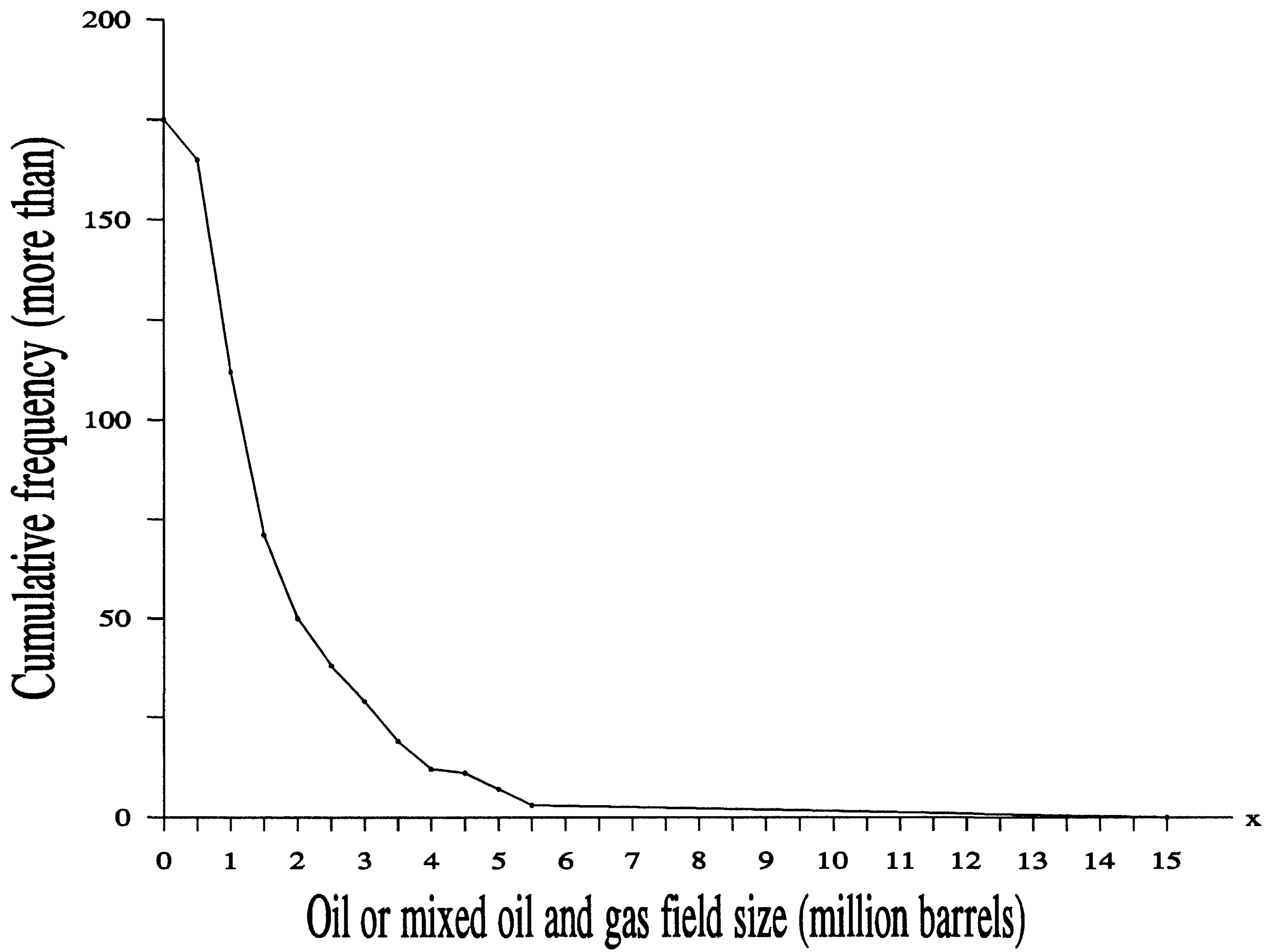
c. Cumulative frequency polygon (less than)



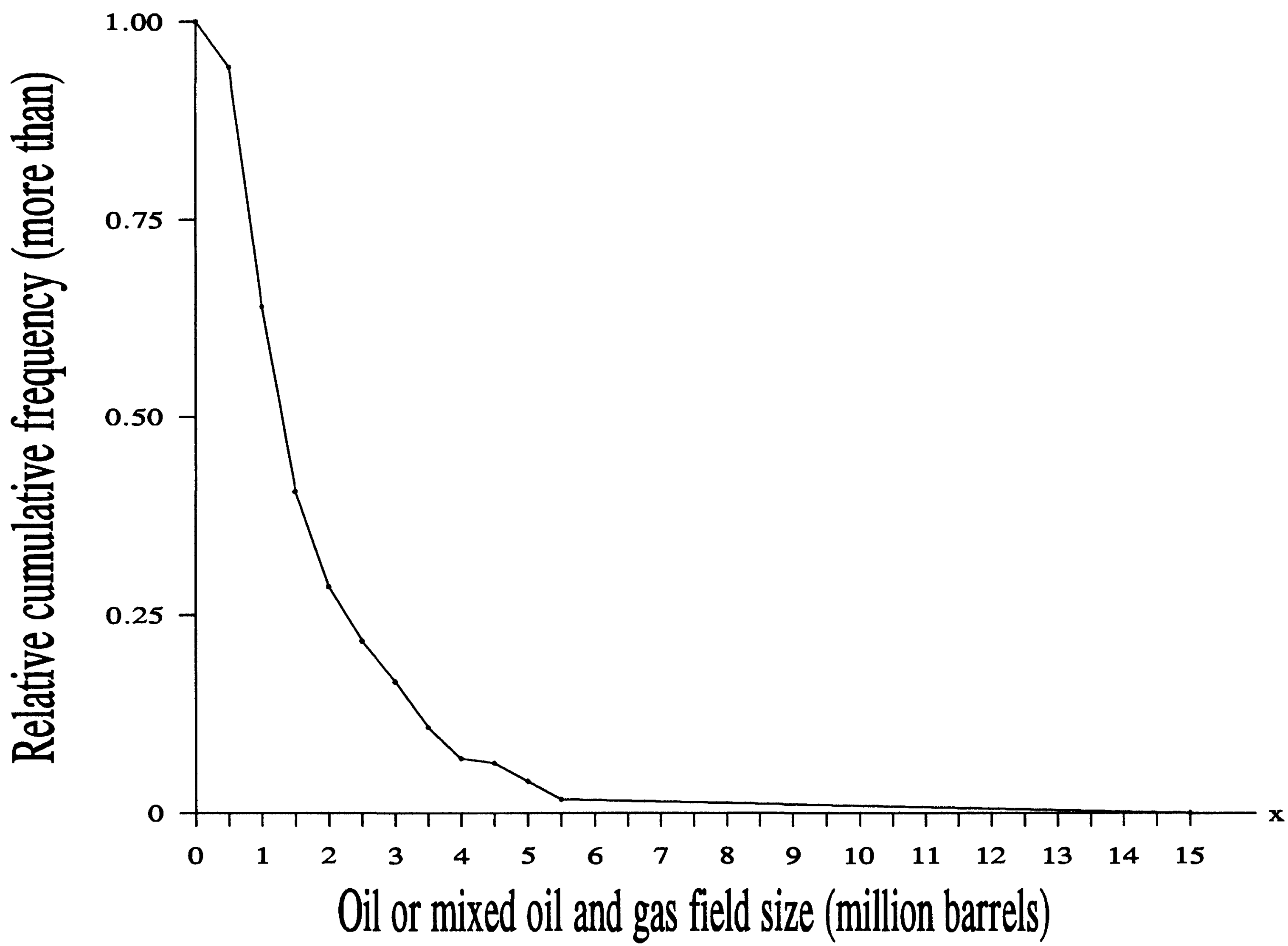
d. Relative cumulative frequency polygon (less than)



e. Cumultive frequency polygon (more than)



f. Relative cumulative frequency polygon (more than)



3. Measures of central location

Also called measures of central tendency or averages.

a. Sample mean

Definition: If $X_1, X_2, ..., X_n$ represent a sample of size n, then the <u>sample mean</u> is defined by the statistic

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Example 1: Net pay thickness data

X: Net pay thickness (feet)

Data from 20 wells completed in a basin:

111, 81, 142, 59, 109, 96, 124, 139, 89, 129, 104, 186, 65, 95, 54, 72, 167, 135, 84, 154.

Because n = 20 and $\sum X_i = 2195$,

sample mean:

$$\overline{X} = \frac{\sum_{i=1}^{20} X_i}{20} = \frac{2195}{20} = 109.75$$

Example 2: Oil and mixed oil and gas field sizes

X:Oil or mixed oil and gas field size (million barrels)

Because n = 175 and $\sum X_i = 317.56$,

sample mean:

$$\overline{X} = \frac{\sum_{i=1}^{175} X_i}{175} = \frac{317.56}{175} = 1.81$$

b. Sample median

Definition: If $X_1, X_2, \ldots, X_n$ represent a sample of size n, arranged in increasing order of magnitude, then the sample median is defined by the statistic

$$\tilde{X} = \begin{cases} X_{(n+1)/2} & \text{if n is odd} \\ \dfrac{X_{n/2} + X_{(n/2)+1}}{2} & \text{if n is even} \end{cases}$$

Example 1: Net pay thickness data

X: Net pay thickness (feet)

Arranging the 20 observations in increasing order of magnitude,

54, 59, 65, 72, 81, 84, 89, 95, 96, 104, 109, 111, 124, 129, 135, 139, 142, 154, 167, 186.

Because n = 20 is even,

sample median:

$$\tilde{X} = \frac{X_{10} + X_{11}}{2} = \frac{104 + 109}{2} = 106.50$$

Example 2: Oil and mixed oil and gas field sizes

X:Oil or mixed oil and gas field size (million barrels)

Because n = 175 is odd,

sample median:

$$\tilde{X} = X_{88} = 1.22$$

c. Sample mode

Definition: If $X_1, X_2, ..., X_n$ represent a sample of size n, then the sample mode M is that value of the sample that occurs most often or with the maximum frequency. The mode may not exist, and when it does, it is not necessarily unique.

Example 1: Net pay thickness data

X: Net pay thickness (feet)

Data from 20 wells completed in a basin:

111, 81, 142, 59, 109, 96, 124, 139, 89, 129, 104, 186, 65, 95, 54, 72, 167, 135, 84, 154.

Because all of the values in the sample are different,

sample mode: M does not exist

Example 2: Oil and mixed oil and gas field sizes

X:Oil or mixed oil and gas field size (million barrels)

Because the value of 1.30 occurs most often with frequency of 5,

sample mode: M = 1.30

The sample mode is usually more useful in terms of a frequency distribution, where the modal class is defined to be the class with maximum frequency.

Modal class is [0.5, 1.0)

4. Measures of variation

a. Sample variance

Definition: If $X_1, X_2, \ldots, X_n$ represent a sample of size n, then the <u>sample variance</u> is defined by the statistic

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

The reason for using n-1 as the divisor will be explained later.

Theorem:

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

Example 1: Net pay thickness data

X: Net pay thickness (feet)

Data from 20 wells completed in a basin:

111, 81, 142, 59, 109, 96, 124, 139, 89, 129, 104, 186, 65, 95, 54, 72, 167, 135, 84, 154.

Because $n = 20$, $\Sigma X_i = 2195$ and $\Sigma X_i^2 = 266531$,

sample variance:

$$S^2 = \frac{20(266531) - (2195)^2}{20(19)} = 1348.93$$

Example 2: Oil and mixed oil and gas field sizes

X: Oil or mixed oil and gas field size (million barrels)

Because $n = 175$, $\Sigma X_i = 317.56$ and $\Sigma X_i^2 = 1105.62$,

sample variance:

$$S^2 = \frac{175(1105.62) - (317.56)^2}{175(174)} = 3.04$$

b. Sample standard deviation

Definition: The <u>sample standard deviation</u>, denoted by S, is the positive square root of the sample variance, i.e.,

$$S = \sqrt{S^2}$$

Example 1: Net pay thickness data

X: Net pay thickness (feet)

Because the sample variance $S^2 = 1348.93$,

sample standard deviation:

$$S = \sqrt{1348.93} = 36.73$$

Example 2: Oil and mixed oil and gas field sizes

X:Oil or mixed oil and gas field size (million barrels)

Because the sample variance $S^2 = 3.04$,

sample standard deviation:

$$S = \sqrt{3.04} = 1.74$$

Note: The sample standard deviation S has the same units as the random variable X.

c. Sample range

Definition: If $X_1, X_2, \ldots, X_n$ represent a sample of size n, and $X_{(n)}$ and $X_{(1)}$ are, respectively, the largest and smallest observations in the sample, then the <u>sample range</u> is defined by the statistic

$$R = X_{(n)} - X_{(1)}$$

Example 1: Net pay thickness data

X: Net pay thickness (feet)

Data from 20 wells completed in a basin:

111, 81, 142, 59, 109, 96, 124, 139, 89, 129, 104, 186, 65, 95, 54, 72, 167, 135, 84, 154.

Because $X_{(20)} = 186$ and $X_{(1)} = 54$,

sample range:

$$R = X_{(20)} - X_{(1)} = 186 - 54 = 132$$

Example 2: Oil and mixed oil and gas field sizes

X:Oil or mixed oil and gas field size (million barrels)

Because n = 175, $X_{(175)} = 14.25$ and $X_{(1)} = 0.11$,

sample range:

$$R = X_{(175)} - X_{(1)} = 14.25 - 0.11 = 14.14$$

C. Sampling Distributions

Recall: A statistic is a random variable.

A sampling distribution is the probability distribution of a statistic.

1. Sampling distribution of the mean

   a. Mean and standard deviation of $\overline{X}$

      Theorem: Given a random sample of size n from <u>any</u> population with mean μ and standard deviation σ, then

      $$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \sigma / \sqrt{n}$$

      Example: Given a random sample of size n = 9 from <u>any</u> population of porosity (%) with $\mu = 18$ and $\sigma = 3$, then

      $$\mu_{\bar{x}} = 18 \text{ and } \sigma_{\bar{x}} = 3 / \sqrt{9} = 1$$

   b. Sampling distribution of $\overline{X}$

      Theorem: Given a random sample of size n from a <u>normal</u> population with mean μ and standard deviation σ, then

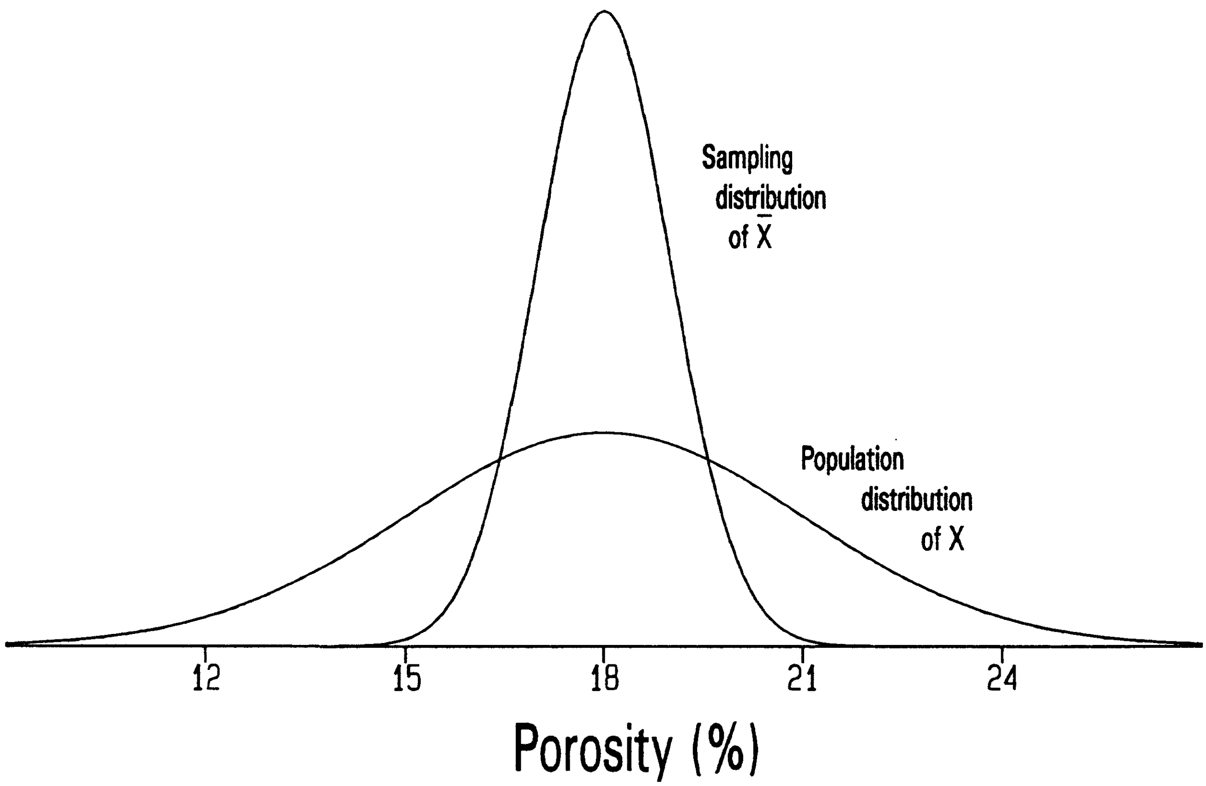
      $\overline{X}$ has a normal distribution with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma / \sqrt{n}$.

      Therefore, $P(\overline{X} < \bar{x}) = P(Z < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = z)$

      Example: Given a random sample of size n = 9 from a <u>normal</u> population of porosity (%) with $\mu = 18$ and $\sigma = 3$, then

      $\overline{X}$ has a normal distribution with $\mu_{\bar{x}} = 18$ and $\sigma_{\bar{x}} = 1$

      $P(\overline{X} < 20) = P(Z < \frac{20 - 18}{1} = 2) = 0.9772$ from Table A.1



X has a normal distribution ($\mu = 18$, $\sigma = 3$)

$\bar{X}$ has a normal distribution ($\mu_{\bar{X}} = 18$, $\sigma_{\bar{X}} = 1$) for n = 9

2. Central Limit Theorem

a. One of the most amazing theorems in all of mathematics

Theorem: Given a random sample of size n from <u>any</u> population with mean μ and standard deviation σ, then

$\overline{X}$ has <u>approximately</u> a normal distribution with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma / \sqrt{n}$

if n is large ($n \geq 30$).

Therefore,

$$P(\overline{X} < \bar{x}) \doteq P(Z < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = z)$$

b. Example: Given a random sample of size n = 36 from a <u>uniform</u> population of porosity (%) with parameters a = 12.8 and b = 23.2.

The population mean and standard deviation are

$$\mu = \frac{a+b}{2} = \frac{12.8 + 23.2}{2} = 18$$

$$\sigma = \frac{b-a}{\sqrt{12}} = \frac{23.2 - 12.8}{\sqrt{12}} = 3$$

$\overline{X}$ has approximately a normal distribution with

$\mu_{\bar{x}} = \mu = 18$ and $\sigma_{\bar{x}} = \sigma / \sqrt{n} = 3 / \sqrt{36} = 0.5$.

$$P(\overline{X} < 19) \doteq P(Z < \frac{19 - 18}{0.5} = 2) = 0.9772 \text{ from Table A.1}$$

c. Even small sample size n gives bell-shaped distribution

Example: Given a random sample of size n = 9 from a <u>uniform</u> population of porosity (%) with parameters a = 12.8 and b = 23.2.

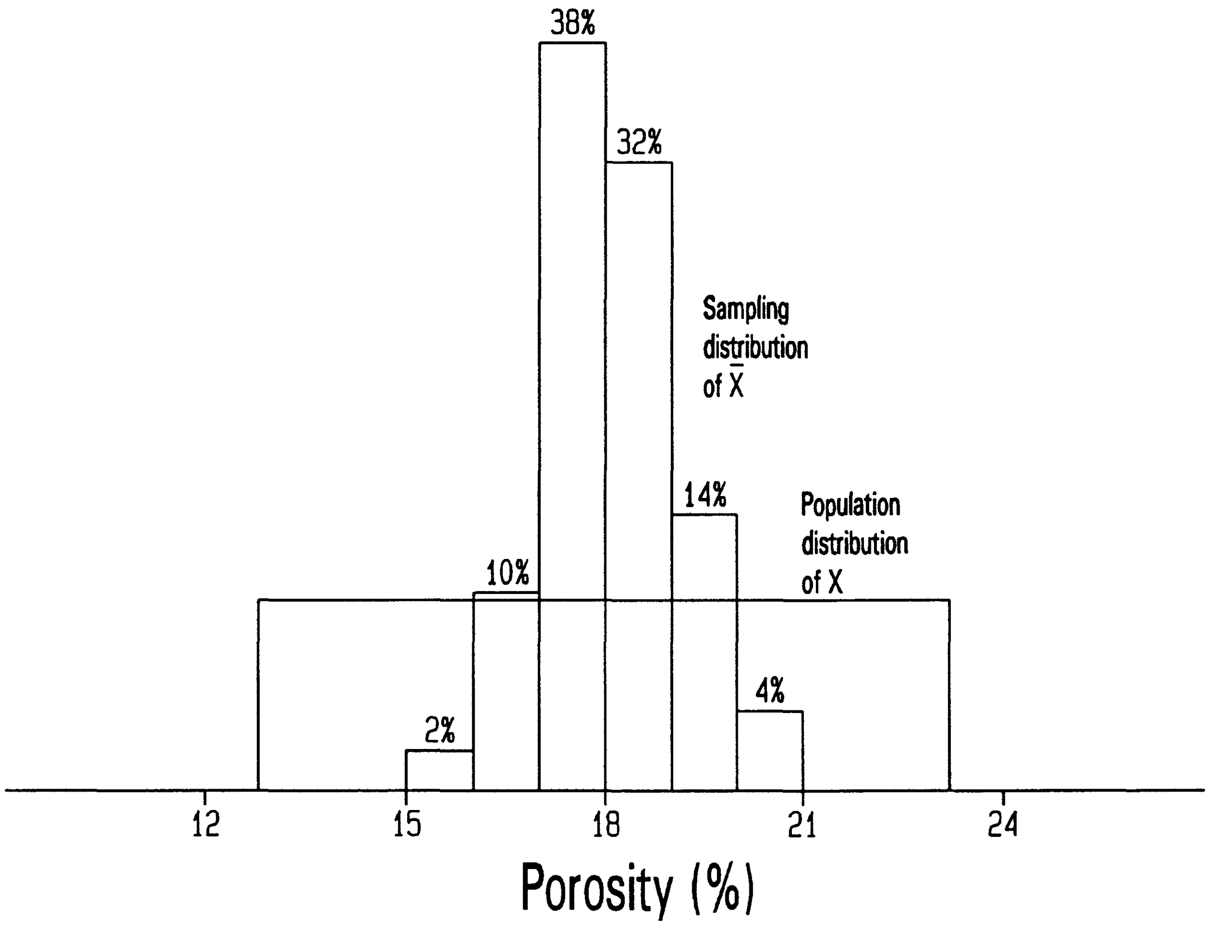
X has a uniform distribution ($\mu = 18$, $\sigma = 3$)

$\overline{X}$ has approximately a bell-shaped distribution with

$\mu_{\bar{x}} = \mu = 18$ and $\sigma_{\bar{x}} = \sigma / \sqrt{n} = 3 / \sqrt{9} = 1$.

50 samples of size n = 9 from Uniform (a = 12.8, b = 23.2)

Sample	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*Mean*
1	14.38	13.01	17.18	21.24	21.04	13.92	15.61	15.61	15.88	16.43
2	16.64	18.49	22.01	17.93	17.69	13.58	20.78	21.27	16.29	18.30
3	16.68	20.83	20.12	16.29	14.96	22.82	18.81	17.03	14.18	17.97
4	16.90	18.49	16.49	14.43	16.80	20.54	18.08	23.10	21.95	18.53
5	23.03	13.58	14.00	17.48	13.90	22.46	21.61	14.28	15.09	17.27
6	17.81	15.96	20.44	14.88	16.65	22.35	16.75	16.97	18.10	17.77
7	12.99	21.35	14.75	21.04	21.37	19.55	13.57	19.20	21.76	18.40
8	12.87	19.51	22.34	22.86	18.20	22.66	21.59	21.61	19.33	20.11
9	15.94	19.47	16.14	15.05	22.21	16.93	14.72	16.89	16.46	17.09
10	15.10	19.53	22.90	19.56	17.97	21.69	20.25	13.92	21.66	19.18
11	22.67	14.05	13.27	16.26	15.07	15.76	19.33	21.09	22.33	17.76
12	22.00	13.65	21.37	19.35	21.07	21.04	19.39	19.26	14.28	19.05
13	15.63	21.77	18.65	21.33	16.12	14.79	18.90	16.98	15.00	17.69
14	17.08	14.27	22.16	14.64	17.07	13.16	21.55	17.57	19.34	17.43
15	22.81	20.46	15.45	14.49	15.04	22.00	13.91	17.58	16.03	17.53
16	15.53	20.08	21.56	17.38	18.27	18.72	14.39	17.43	20.32	18.19
17	13.01	19.61	21.98	20.90	15.92	13.63	21.30	15.14	19.41	17.88
18	19.18	15.01	20.65	13.23	22.54	16.30	22.23	16.04	22.84	18.67
19	17.12	18.45	13.95	18.61	22.94	14.87	21.97	17.20	15.08	17.80
20	19.84	17.08	16.46	15.35	18.65	15.82	13.55	13.90	12.96	15.95
21	15.68	22.33	16.65	14.51	14.41	21.47	22.93	19.70	13.62	17.92
22	13.46	20.58	16.90	20.43	20.30	22.88	17.77	16.36	17.55	18.47
23	14.31	19.84	17.05	21.27	14.44	18.78	20.21	21.50	19.15	18.51
24	15.28	16.86	18.55	13.48	20.34	22.05	12.83	14.57	14.85	16.53
25	20.25	20.55	22.50	14.04	21.47	17.48	20.02	14.85	21.74	19.21
26	19.91	19.52	21.39	21.28	22.46	21.21	17.97	22.88	19.24	20.65
27	13.02	19.99	14.25	16.18	19.69	21.50	18.46	18.86	22.65	18.29
28	15.16	20.41	13.76	16.82	21.75	19.77	19.70	15.85	15.92	17.68
29	18.88	20.42	21.49	14.08	19.72	17.15	16.31	17.91	18.44	18.27
30	13.30	13.09	13.28	20.33	19.25	18.45	21.45	21.80	17.46	17.60
31	22.34	20.51	17.68	19.51	17.13	16.45	16.63	19.21	14.63	18.23
32	14.93	13.60	16.16	13.48	13.11	20.06	19.46	20.60	15.17	16.29
33	20.70	19.93	13.12	22.24	13.69	17.43	16.12	20.38	17.04	17.85
34	19.02	13.01	20.98	19.44	15.26	22.28	13.49	16.16	20.56	17.80
35	21.44	14.21	21.78	19.03	17.22	23.17	22.71	19.22	18.90	19.74
36	15.76	21.59	21.02	18.81	19.60	20.23	18.48	16.44	15.73	18.63
37	13.88	17.17	14.65	13.97	21.41	19.22	15.08	13.87	20.92	16.68
38	22.91	21.47	19.82	22.28	14.39	22.01	17.16	16.87	16.49	19.27
39	14.58	16.96	21.37	19.98	22.41	15.33	22.86	14.50	21.72	18.86
40	15.94	20.42	20.01	21.24	20.14	21.38	19.32	13.53	16.38	18.71
41	18.45	16.40	21.48	23.08	15.33	14.36	16.88	19.94	14.68	17.84
42	14.27	16.67	13.30	14.71	23.01	14.76	15.65	19.59	21.00	17.00
43	23.06	21.63	22.18	21.26	13.54	13.83	16.96	20.83	19.59	19.21
44	21.10	14.95	18.53	18.03	14.64	17.18	20.72	20.84	20.87	18.54
45	20.11	20.29	21.60	14.21	14.33	17.63	17.09	21.80	17.28	18.26
46	22.22	14.06	18.14	13.72	16.44	14.96	19.87	19.09	21.24	17.75
47	15.37	12.89	17.83	18.78	14.46	18.71	20.58	19.50	18.35	17.39
48	17.31	15.17	14.09	15.15	14.32	13.83	16.77	19.78	20.84	16.36
49	17.42	19.83	15.89	19.83	16.42	18.86	14.16	20.00	20.16	18.06
50	19.50	22.30	20.58	20.21	22.32	14.21	15.80	16.39	22.05	19.26



X has a uniform distribution ($\mu = 18$, $\sigma = 3$)

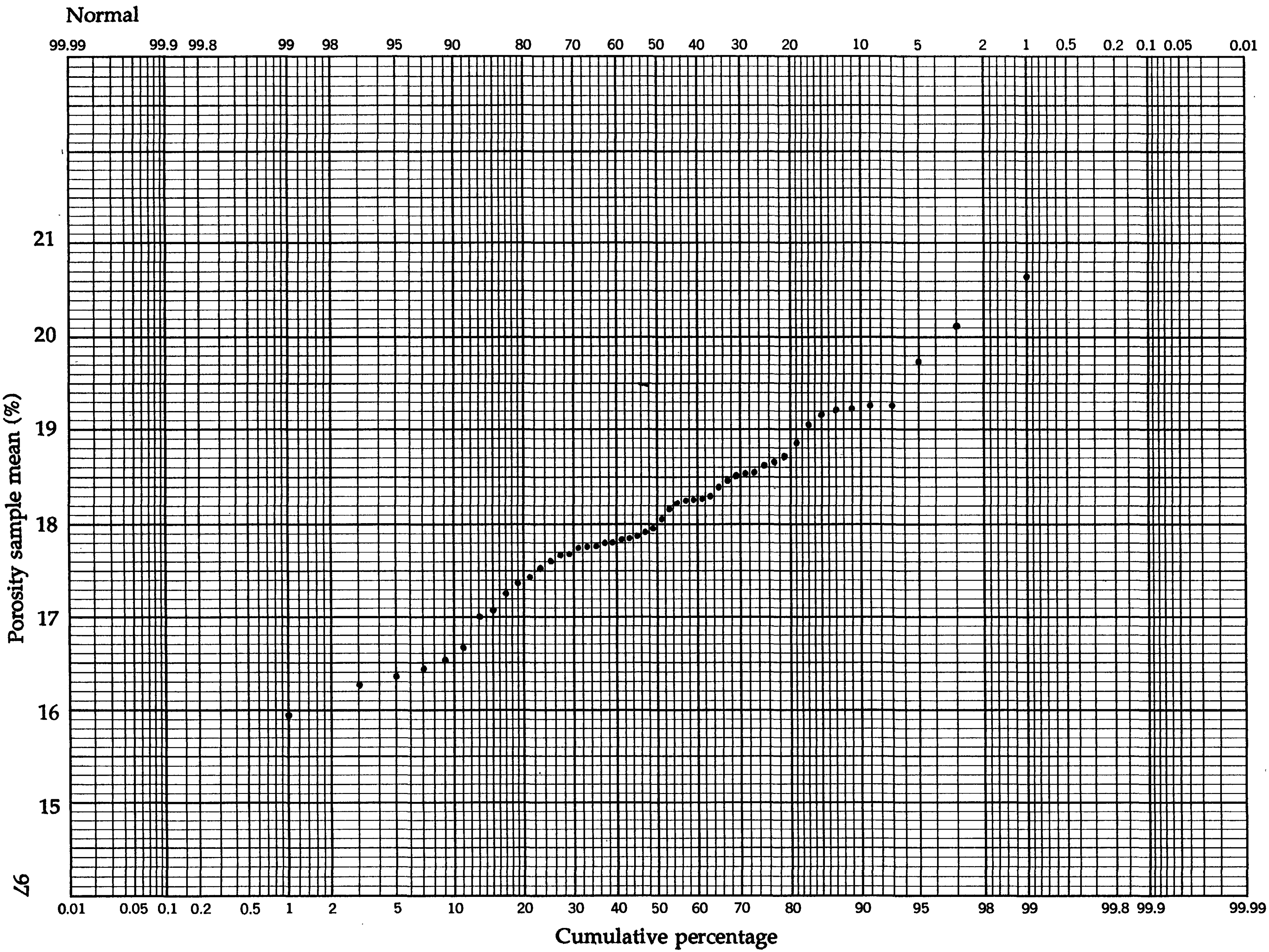
$\bar{X}$ has a bell-shaped distribution ($\mu_{\bar{X}} = 18$, $\sigma_{\bar{X}} = 1$) even for n = 9

3. Normal probability paper

A linear pattern of plotted points suggests a normal distribution.

50 random sample means $\bar{x}_i$ with n = 9 ($\mu_{\bar{X}}$ = 18, $\sigma_{\bar{X}}$ = 1) from a uniform population (μ = 18, σ = 3).

Index No. i	Sample Mean $\bar{x}_i$	Cumulative Percentage $(i - 0.5)\,100/50$
1	15.95	1
2	16.29	3
3	16.36	5
4	16.43	7
5	16.53	9
6	16.68	11
7	17.00	13
8	17.09	15
9	17.27	17
10	17.39	19
11	17.43	21
12	17.53	23
13	17.60	25
14	17.68	27
15	17.69	29
16	17.75	31
17	17.76	33
18	17.77	35
19	17.80	37
20	17.80	39
21	17.84	41
22	17.85	43
23	17.88	45
24	17.92	47
25	17.97	49
26	18.06	51
27	18.19	53
28	18.23	55
29	18.26	57
30	18.27	59
31	18.29	61
32	18.30	63
33	18.40	65
34	18.47	67
35	18.51	69
36	18.53	71
37	18.54	73
38	18.63	75
39	18.67	77
40	18.71	79
41	18.86	81
42	19.05	83
43	19.18	85
44	19.21	87
45	19.21	89
46	19.26	91
47	19.27	93
48	19.74	95
49	20.11	97
50	20.65	99


Normal
99.99 99.9 99.8 99 98 95 90 80 70 60 50 40 30 20 10 5 2 1 0.5 0.2 0.1 0.05 0.01
21
20
19
18
17
16
15
Porosity sample mean (%)
0.01 0.05 0.1 0.2 0.5 1 2 5 10 20 30 40 50 60 70 80 90 95 98 99 99.8 99.9 99.99
Cumulative percentage

D. Inferential Statistics

There are two basic areas of statistics:

1. Descriptive statistics
2. Inferential statistics

Descriptive statistics: the study of techniques for describing a given set of data.

- Probability is not utilized.
- Simple statistical methods are used.

  e.g., pie charts, bar charts, etc.

Note: The term "descriptive statistics" is also used to refer to statistics, e.g., the sample mean.

Inferential statistics: the study of procedures for making inferences about a population on the basis of a sample.

- Probability is utilized.
- Inferences are made about aspects of a population:
  1. Population distribution

     e.g., normal or lognormal distribution
  2. Population parameters

     e.g., population mean or variance
- Complex statistical methods are used:
  1. Parametric statistics
  2. Nonparametric statistics

Parametric statistics: statistical methods based on the assumption of a normal population (normality assumption), and inference is made about the population parameters, e.g., population mean μ or standard deviation σ.

Nonparametric (or distribution-free) statistics: statistical methods based on no assumption of the population distribution (distribution-free assumption).

There are two basic types of inference:

1. Statistical estimation
2. Tests of hypotheses

Statistical estimation: The type of inference is an estimate of a population aspect.

Example: Estimate the population mean μ.

Tests of hypotheses: The type of inference is a test of hypothesis of a population aspect.

Example: Test the hypothesis of a lognormal population.

Assumptions of a statistical method: the conditions under which the statistical method is valid.

Robust statistical method: a statistical method that is still valid even with moderate violation of the assumptions.

Throughout the rest of this section, a random sample is assumed.

1. Statistical estimation

a. Point estimation

Point estimate of a population parameter θ: a single number that can be regarded as the most plausible value of θ.

Example: A point estimate of the porosity population mean μ is 18.5% (say).

Point estimator of a population parameter θ: a statistic Y that is used to obtain a point estimate of θ.

Examples:

1. A point estimator of the population mean μ is the sample mean $\overline{X}$.
2. A point estimator of the population mean μ is the sample median $\tilde{X}$.

Because a point estimator is a statistic, a point estimator has a sampling distribution.

Example: For a normal population $\overline{X}$ has a normal sampling distribution.

Standard error of a point estimator Y: the standard deviation of Y.

Example: The standard error of $\overline{X}$ is $\sigma / \sqrt{n}$.

Unbiased estimator of a population parameter θ: a point estimator Y such that $E(Y) = \theta$ for every possible value of θ.

Examples:

1. An unbiased estimator of μ is $\overline{X}$ because $E(\overline{X}) = \mu$.
2. An unbiased estimator of σ^2 is S^2 because $E(S^2) = \sigma^2$.

   This explains why n-1 is used as the divisor.

Principle of minimum variance unbiased estimation: Among all unbiased estimators of θ, choose the estimator that has minimum variance.

Minimum variance unbiased estimator (MVUE) of θ:

the unbiased estimator that has minimum variance.

Example: For a normal population,

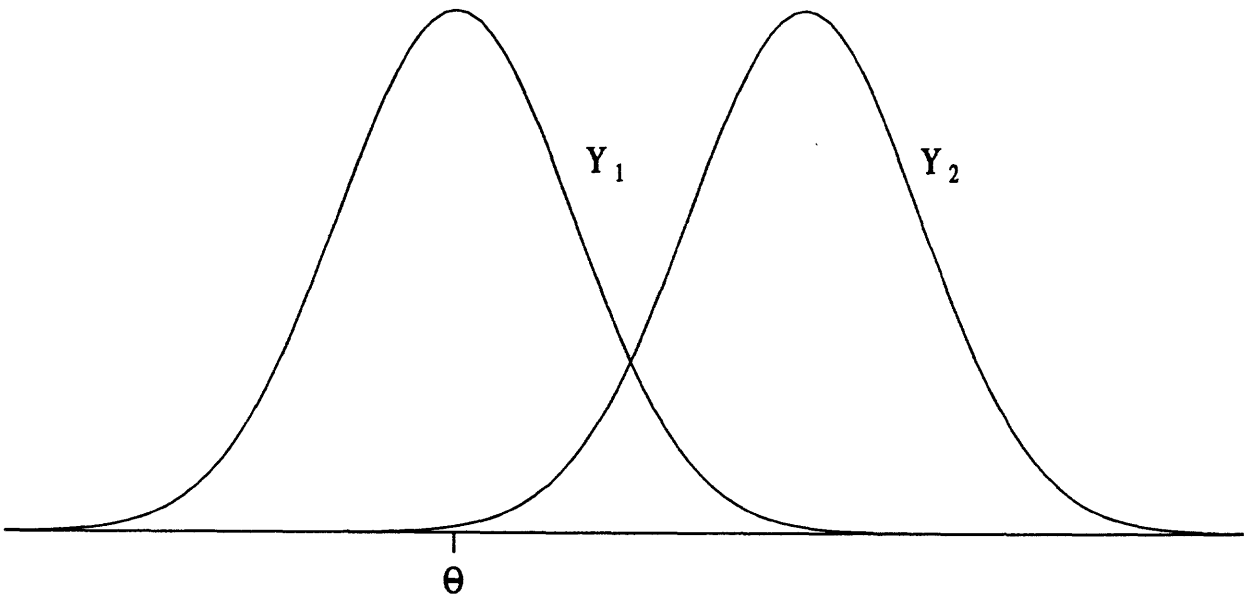
1. Both $\overline{X}$ and $\tilde{X}$ are unbiased estimators of the population mean μ.

2. The variance of $\overline{X}$ is smaller than the variance of $\tilde{X}$.

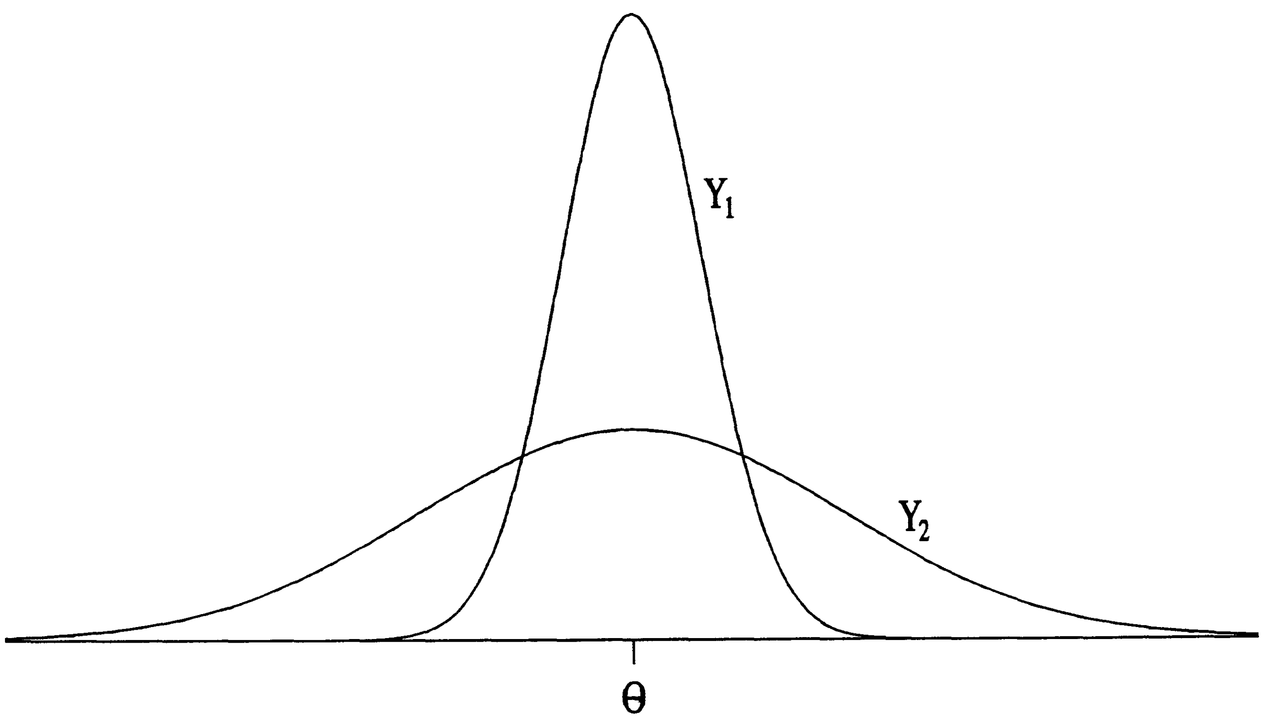
3. $\overline{X}$ is the MVUE of μ.

Important: The best estimator of μ depends on the population distribution.

If the population is not normal, the best estimator may not be $\overline{X}$.



Y_1 is an unbiased estimator of θ, while Y_2 is not.



Variance of unbiased estimator Y_1 < variance of unbiased estimator Y_2

b. Interval esimation

Interval estimate of a population parameter θ: an interval of numbers of the form $y_L < \theta < y_U$

Example: An interval estimate of the porosity population mean μ is $17.5 < \mu < 19.2$ (say).

Confidence interval of a population parameter θ: an interval such that $P(Y_L < \theta < Y_U) = 1-\alpha$, for $0 < \alpha < 1$.

The end points Y_L and Y_U are called the lower and upper confidence limits.

Example: A 95% confidence interval for μ

Suppose that the parameter of interest is a population mean μ and the following assumptions are made

1. The population distribution is normal, and
2. The value of the population standard deviation σ is known.

Recall: The sample mean $\overline{X}$ is normally distributed such that

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

has a standard normal distribution. Then

$$P(-1.96 < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < 1.96) = 0.95$$

Manipulating the inequalities to get the form $Y_L < \mu < Y_U$,

$$P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

Example: A 95% confidence interval for the porosity population mean μ with the above assumptions and given quantities

$\sigma = 3.0$, $n = 16$, and $\overline{x} = 18.5$.

$$\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 18.5 \pm (1.96)\frac{3.0}{\sqrt{16}} = 18.5 \pm 1.5 = (17.0,\ 20.0)$$

A 95% confidence interval for μ is $17.0 < \mu < 20.0$.

2. Tests of hypotheses

a. Z test for μ

Parameter of interest: population mean μ

Assumptions:

1. The population distribution is normal, and

2. The value of the population standard deviation σ is known.

Null hypothesis H_0: $\mu = \mu_0$

Alternative hypothesis H_1: $\mu < \mu_0$, $\mu > \mu_0$, or $\mu \neq \mu_0$

Test statistic:

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

Significance level: α = P(H_0 is rejected when H_0 is true) = α_0

Critical region:

Reject H_0 if $z < -z_\alpha$, $z > z_\alpha$, or $z < -z_{\alpha/2}$ and $z > z_{\alpha/2}$

Take random sample of size n: $\overline{x}$

Compute test statistic: z

Decision: Decide whether or not H_0 should be rejected

Example: Porosity (%)

Parameter of interest: true mean porosity μ

Assumptions: A normal population with known $\sigma = 3$

H_0: $\mu = 17$

H_1: $\mu < 17$ (left-tailed test)

Test statistic:

$$Z = \frac{\overline{X} - 17}{3 / \sqrt{n}}$$

$\alpha = 0.05$

Critical region: $z < -1.645$

Random sample of size n = 25 : $\bar{x} = 16.2$

Compute: $$z = \frac{16.2 - 17}{3 / \sqrt{25}} = -1.33$$

Decision: Do not reject H_0 and conclude that the mean porosity is not significantly less than 17%.

b. Chi-squared goodness-of-fit test

Definition: The lognormal distribution

A nonnegative r.v. X has a lognormal distribution if

the r.v.Y = ln(X) has a normal distribution.

Parameters: mean μ and standard deviation σ of the <u>normal</u> distribution

Example: Oil and mixed oil and gas field sizes

H_0: the population distribution is lognormal

H_1: the population distribution is not lognormal

Test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(F_i - E_i)^2}{E_i}$$

where

χ^2 has approximately a chi-squared distribution.

k is the number of classes.

F_i is the observed frequency of class i.

$E_i = n\hat{p}_i$ is the estimated expected frequency ($E_i \geq 5$ for every i).

$\hat{p}_i$ is the estimated probability of an observation falling in class i.

n is the sample size.

Significance level: α = P(H_o is rejected when H_o is true) = 0.05

Critical region:

if $\chi^2 \geq \chi^2_{\alpha,\, k-1}$ reject H_o

if $\chi^2 \leq \chi^2_{\alpha,\, k-1-m}$ don't reject H_o

if $\chi^2_{\alpha,\, k-1-m} < \chi^2 < \chi^2_{\alpha,\, k-1}$ withhold judgment

where m is the number of independent parameters estimated.

m = 2 because μ and σ are estimated from a sample.

From the sample of size n = 175, take logarithms $y_i = \ln(x_i)$ and

compute estimates of μ and σ: $\bar{y}$ = 0.31 and s= 0.74.

Chi-squared goodness-of-fit test

Class Interval	Observed Frequency f_i	Boundary b_i	Ln of Boundary $\ln(b_i)$	$z_i = \frac{\ln(b_i) - 0.31}{0.74}$	$P(Z < z_i)$	$\hat{p}_i$	$e_i = n\hat{p}_i$	Combined e_i	$\frac{(f_i - e_i)^2}{e_i}$
1. [0, 0.5)	10	0.5	–0.69	–1.35	0.0885	0.0885	15.49		1.95
2. [0.5, 1.0)	53	1.0	0	–0.42	0.3372	0.2487	43.52		2.07
3. [1.0, 1.5)	41	1.5	0.41	0.14	0.5557	0.2185	38.24		0.20
4. [1.5, 2.0)	21	2.0	0.69	0.51	0.6950	0.1393	24.38		0.47
5. [2.0, 2.5)	12	2.5	0.92	0.82	0.7939	0.0989	17.31		1.63
6. [2.5, 3.0)	9	3.0	1.10	1.07	0.8577	0.0638	11.17		0.42
7. [3.0, 3.5)	10	3.5	1.25	1.27	0.8980	0.0403	7.05		0.54
8. [3.5, 4.0)	7	4.0	1.39	1.46	0.9278	0.0298	5.22		0.61
9. [4.0, 4.5)	1	4.5	1.50	1.61	0.9463	0.0185	3.24		
10. [4.5, 5.0)	4	5.0	1.61	1.76	0.9608	0.0144	2.52	7.35	0.37
11. [5.0, 5.5)	4	5.5	1.70	1.88	0.9699	0.0091	1.59		
12. [5.5, 15)	3	∞	∞	∞	1	0.0301	5.27		0.98
	n = 175					0.9999	175.00		$\chi^2 = 9.24$

Note that for class 2:
P(0.5 < X < 1.0) = P(–0.69 < Y < 0) = P(–1.35 < Z < –0.42) = 0.3372 – 0.0885 = 0.2487.

After combining classes,
$k = 10$
$\chi^2_{\alpha, k-1} = \chi^2_{0.05, 9} = 16.919$
$\chi^2_{\alpha, k-1-m} = \chi^2_{0.05, 7} = 14.067$

Decision: Because $\chi^2 = 9.24 < 14.067$, H_0 is not rejected, so the lognormal model provides a good fit for the distribution of oil and mixed oil and gas field sizes.

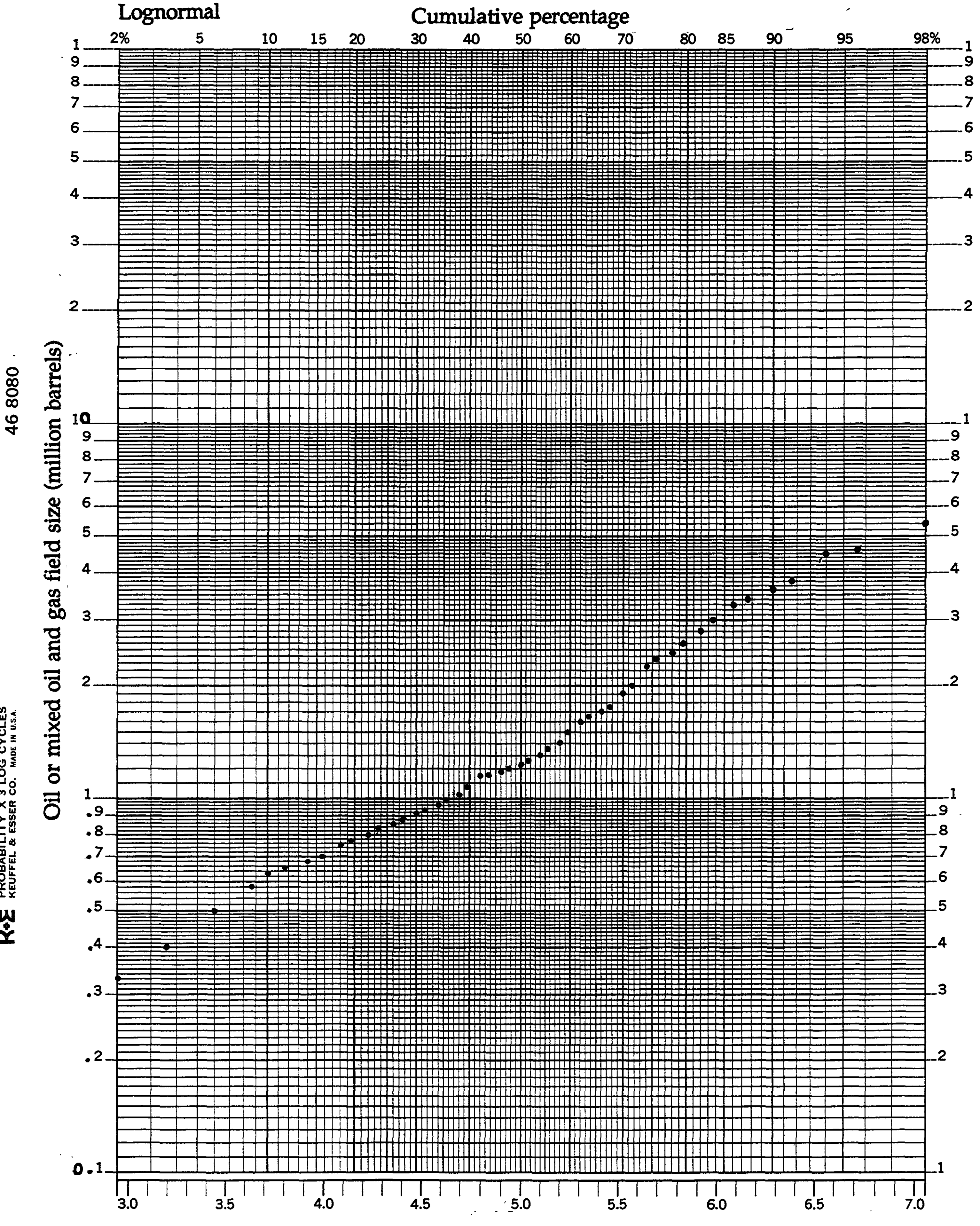
c. Lognormal probability paper

A linear pattern of plotted points suggests a lognormal distribution.

Example: Oil and mixed oil and gas field sizes (million barrels) for 175 fields of 1 million BOE or more known recovery in the northern Michigan Silurian reef play.

X: Oil or mixed oil and gas field size (million barrels)

Index No. i	Ordered Observation x_i	Cumulative Percentage $(i - 0.5)\ 100/175$
1	0.11	0.286
2	0.18	0.857
3	0.24	1.429
4	0.33	2.000
5	0.34	2.571
6	0.38	3.143
7	0.40	3.714
8	0.40	4.286
9	0.45	4.857
10	0.49	5.429
11	0.50	6.000
12	0.55	6.571
13	0.58	7.143
14	0.58	7.714
15	0.60	8.286
16	0.60	8.857
17	0.62	9.429
18	0.63	10.000
19	0.63	10.571
20	0.65	11.143
21	0.65	11.714
22	0.65	12.286
23	0.66	12.857
24	0.66	13.429
25	0.68	14.000
26	0.68	14.571
27	0.68	15.143
28	0.70	15.714
29	0.71	16.286
30	0.73	16.857
31	0.74	17.429
32	0.75	18.000
33	0.75	18.571
34	0.75	19.143
35	0.77	19.714
36	0.78	20.286
37	0.78	20.857
38	0.80	21.429
39	0.80	22.000
40	0.80	22.571
41	0.82	23.143
42	0.83	23.714
43	0.83	24.286
44	0.84	24.857
45	0.85	25.429
46	0.85	26.000
47	0.86	26.571
48	0.87	27.143
49	0.88	27.714
50	0.90	28.286
51	0.90	28.857
52	0.90	29.429
53	0.91	30.000
54	0.91	30.571
55	0.93	31.143
56	0.93	31.714
57	0.93	32.286
58	0.95	32.857
59	0.95	33.429
60	0.96	34.000
61	0.96	34.571
62	0.96	35.143
63	0.99	35.714
64	1.00	36.286
65	1.00	36.857
66	1.00	37.429
67	1.02	38.000
68	1.02	38.571
69	1.06	39.143
70	1.08	39.714
71	1.10	40.286
72	1.13	40.857
73	1.14	41.429
74	1.15	42.000
75	1.15	42.571
76	1.15	43.143
77	1.15	43.714
78	1.16	44.286
79	1.16	44.857
80	1.17	45.429
81	1.18	46.000
82	1.18	46.571
83	1.20	47.143
84	1.20	47.714
85	1.20	48.286
86	1.21	48.857
87	1.22	49.429
88	1.22	50.000
89	1.26	50.571
90	1.27	51.143
91	1.27	51.714
92	1.28	52.286
93	1.30	52.857
94	1.30	53.429
95	1.30	54.000
96	1.30	54.571
97	1.30	55.143
98	1.35	55.714
99	1.35	56.286
100	1.40	56.857
101	1.40	57.429
102	1.40	58.000
103	1.42	58.571
104	1.45	59.143
105	1.50	59.714
106	1.55	60.286
107	1.60	60.857
108	1.60	61.429
109	1.60	62.000
110	1.60	62.571
111	1.65	63.143
112	1.65	63.714
113	1.66	64.286
114	1.68	64.857
115	1.68	65.429
116	1.70	66.000
117	1.71	66.571
118	1.75	67.143
119	1.75	67.714
120	1.80	68.286
121	1.82	68.857
122	1.85	69.429
123	1.90	70.000
124	1.90	70.571
125	1.95	71.143
126	2.00	71.714
127	2.00	72.286
128	2.08	72.857
129	2.10	73.429
130	2.25	74.000
131	2.35	74.571
132	2.35	75.143
133	2.38	75.714
134	2.40	76.286
135	2.40	76.857
136	2.40	77.429
137	2.45	78.000
138	2.60	78.571
139	2.60	79.143
140	2.60	79.714
141	2.70	80.286
142	2.75	80.857
143	2.75	81.429
144	2.80	82.000
145	2.90	82.571
146	2.95	83.143
147	3.00	83.714
148	3.00	84.286
149	3.00	84.857
150	3.15	85.429
151	3.30	86.000
152	3.30	86.571
153	3.30	87.143
154	3.40	87.714
155	3.40	88.286
156	3.45	88.857
157	3.60	89.429
158	3.60	90.000
159	3.65	90.571
160	3.70	91.143
161	3.80	91.714
162	3.90	92.286
163	3.95	92.857
164	4.20	93.429
165	4.50	94.000
166	4.50	94.571
167	4.60	95.143
168	4.60	95.714
169	5.10	96.286
170	5.15	96.857
171	5.20	97.429
172	5.40	98.000
173	7.80	98.571
174	12.00	99.143
175	14.25	99.714


Lognormal
Cumulative percentage
2%
5
10
15
20
30
40
50
60
70
80
85
90
95
98%
Oil or mixed oil and gas field size (million barrels)
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
46 8080
K+E PROBABILITY X 3 LOG CYCLES
KEUFFEL & ESSER CO. MADE IN U.S.A.

Regression and Correlation

Regression

1. Simple linear regression

Straight line

$$y = a + bx$$

where

$$b = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

2. Simple linear regression - no constant

Straight line through origin

$$y = bx$$

where

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

3. Quadratic regression or polynomial regression of degree two

Parabola

$$y = b_0 + b_1x + b_2x^2$$

4. The reduced major axis line

Straight line

$$y = a + bx$$

where

$$b = s_y/s_x = \sqrt{\frac{n\,\Sigma y^2 - (\Sigma y)^2}{n\,\Sigma x^2 - (\Sigma x)^2}} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

with b given the sign of r below.

Correlation

1. Pearson product moment correlation coefficient

$$r = \frac{n \Sigma xy - \Sigma x \Sigma y}{\sqrt{[n \Sigma x^2 - (\Sigma x)^2]\,[n \Sigma y^2 - (\Sigma y)^2]}}$$

measures the strength of the linear relationship between two variables, y and x.

Nonparametric: Spearman rank correlation coefficient

2. Coefficient of Determination

$$r^2 = \frac{SSR}{SST}$$

measures what proportion of the total variation in the response y is accounted for by the fitted regression model and is reported as a percentage $r^2(100\%)$ and interpreted as percentage variation explained by the postulated model.

SSR is called the ***regression sum of squares*** and reflects the amount of variation in the *y* values explained by the model, e.g., a postulated straight line.
SST is called the ***total corrected sum of squares*** and reflects the total amount of variation in the *y* values.

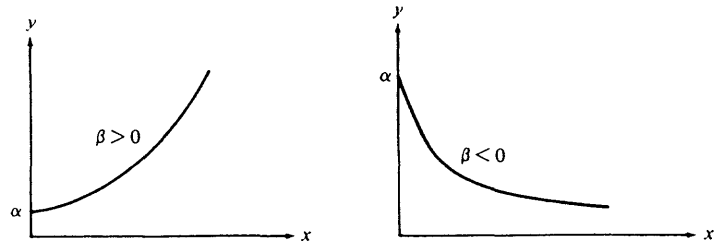
Transformations

A function relating y to x is ***intrinsically linear*** if by means of a transformation on x and/or y, the function can be expressed as $y' = \beta_0 + \beta_1 x'$, where x' is the transformed independent variable and y' is the transformed dependent variable.

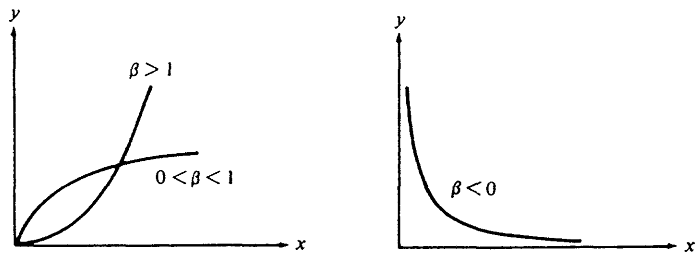
Some useful transformations to linearize:

Function	Transformation	Linear form
(a) Exponential: $y = \alpha e^{\beta x}$	$y' = \ln(y)$	$y' = \ln(\alpha) + \beta x$
(b) Power: $y = \alpha x^{\beta}$	$y' = \log(y)$, $x' = \log(x)$	$y' = \log(\alpha) + \beta x'$
(c) Reciprocal: $y = \alpha + \beta \frac{1}{x}$	$x' = \frac{1}{x}$	$y = \alpha + \beta x'$
(d) Hyperbolic: $y = \frac{x}{\alpha + \beta x}$	$y' = \frac{1}{y}$, $x' = \frac{1}{x}$	$y' = \beta + \alpha x'$
(e) Logarithmic: $y = \alpha + \beta \log(x)$	$x' = \log(x)$	$y = \alpha + \beta x'$

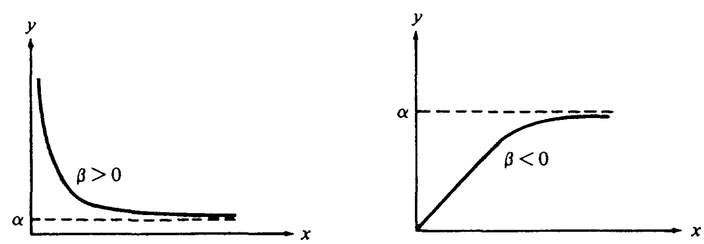
When log (...) appears, either a base 10 or base e logarithm can be used.



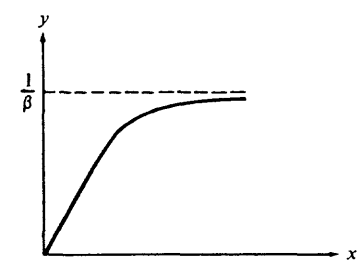
(a) Exponential function



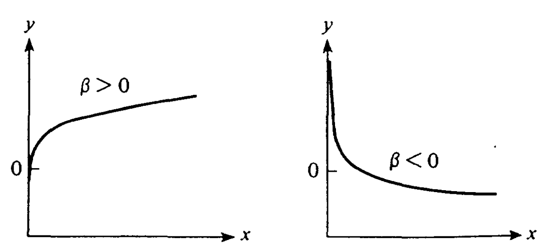
(b) Power function



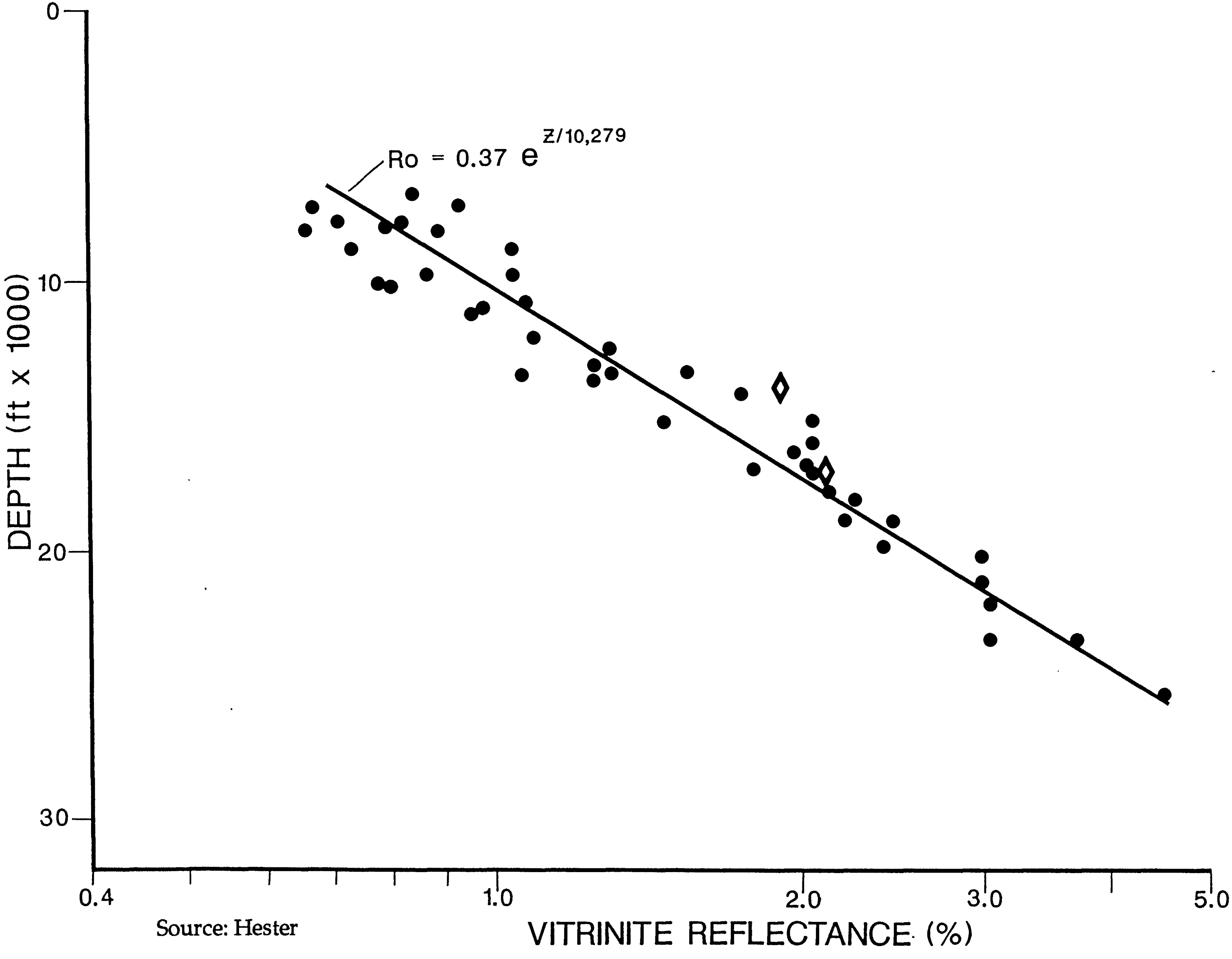
(c) Reciprocal function



(d) Hyperbolic function



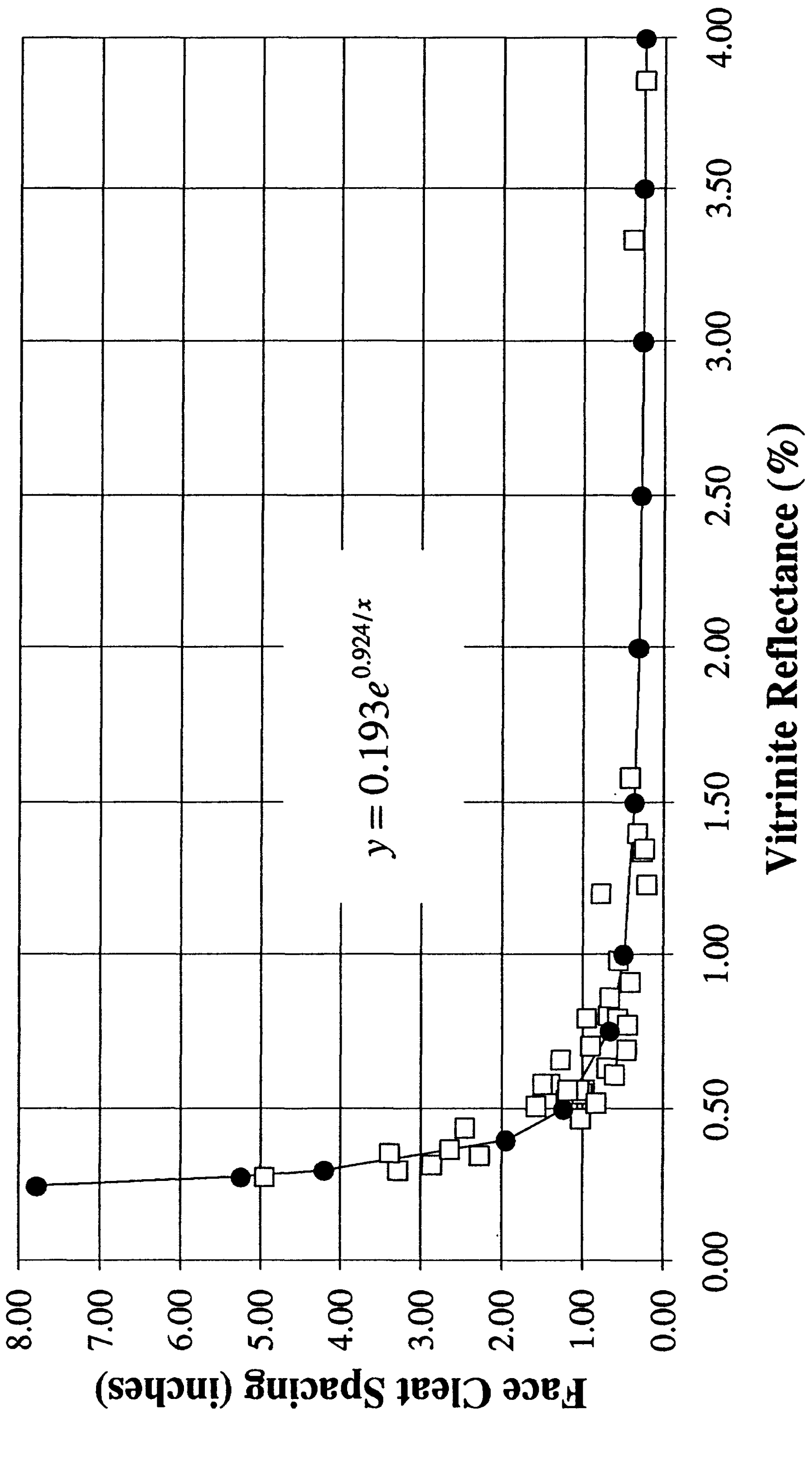
(e)



Source: Hester

Vitrinite Ref. (X axis)		Face Cleat Spacing (Y axis)
1	0.58	1.4
2	0.80	0.68
3	0.58	1.49
4	0.86	0.66
5	0.53	1.2
6	0.28	4.95
7	0.30	3.28
8	1.23	0.20
9	0.56	0.96
10	0.52	1.17
11	0.55	1.00
12	0.51	1.57
13	0.63	0.69
14	0.56	1.04
15	0.47	1.02
16	0.52	0.83
17	0.56	1.18
18	0.70	0.90
19	0.79	0.56
20	1.4	0.32
21	1.34	0.26
22	1.35	0.23
23	0.69	0.45
24	0.32	2.86
25	0.35	2.27
26	0.37	2.64
27	0.36	3.39
28	0.44	2.46
29	0.66	1.26
30	0.98	0.55
31	0.79	0.94
32	3.33	0.39
33	3.86	0.24
34	1.20	0.78
35	1.58	0.41
36	0.77	0.44
37	0.91	0.40
38	0.61	0.60

Source of data: Ben Law



$y = 0.193e^{0.924/x}$
Face Cleat Spacing (inches)
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
Vitrinite Reflectance (%)
0.00
0.50
1.00
1.50
2.00
2.50
3.00
3.50
4.00

An exponential relationship between face cleat spacing y and the reciprocal of vitrinite reflectance x.

$$y = 0.193\, e^{0.924/x}$$

x	y
0.25	7.78
0.28	5.23
0.3	4.2
0.4	1.94
0.5	1.22
0.75	0.66
1	0.49
1.5	0.36
2	0.31
2.5	0.28
3	0.26
3.5	0.25
4	0.24

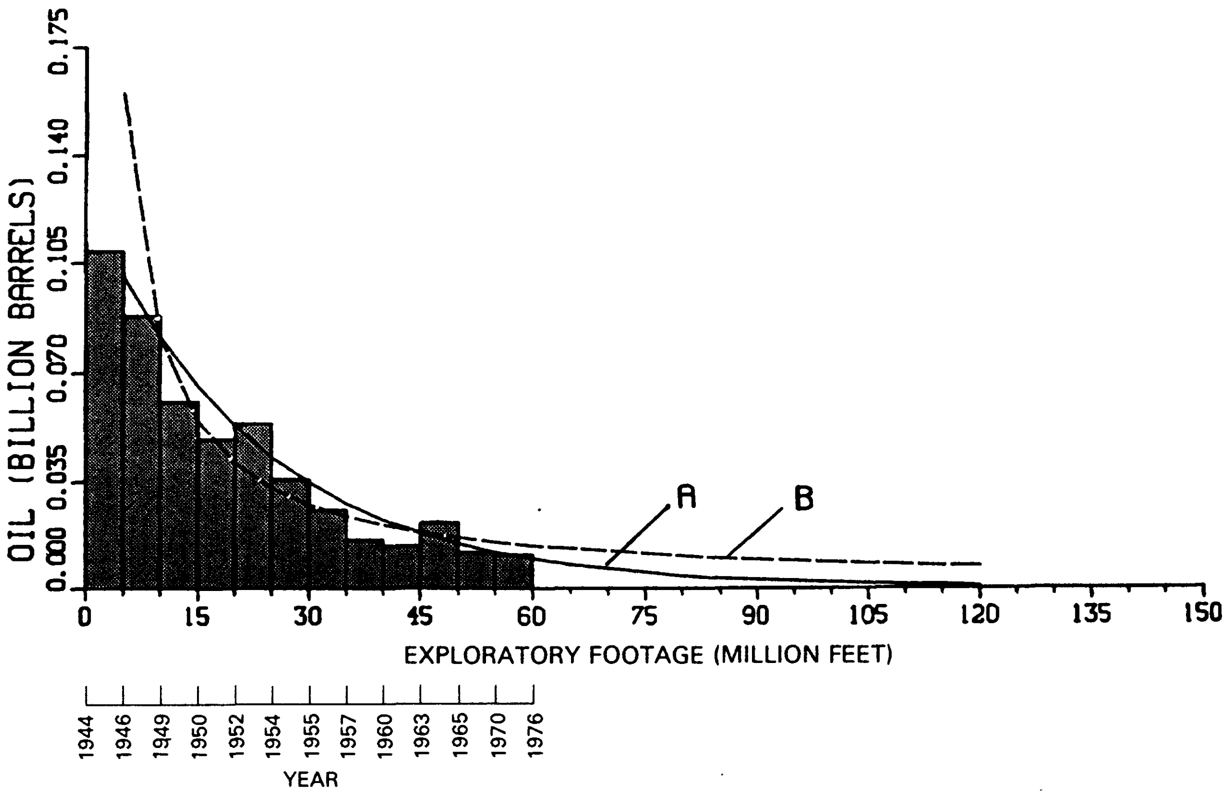
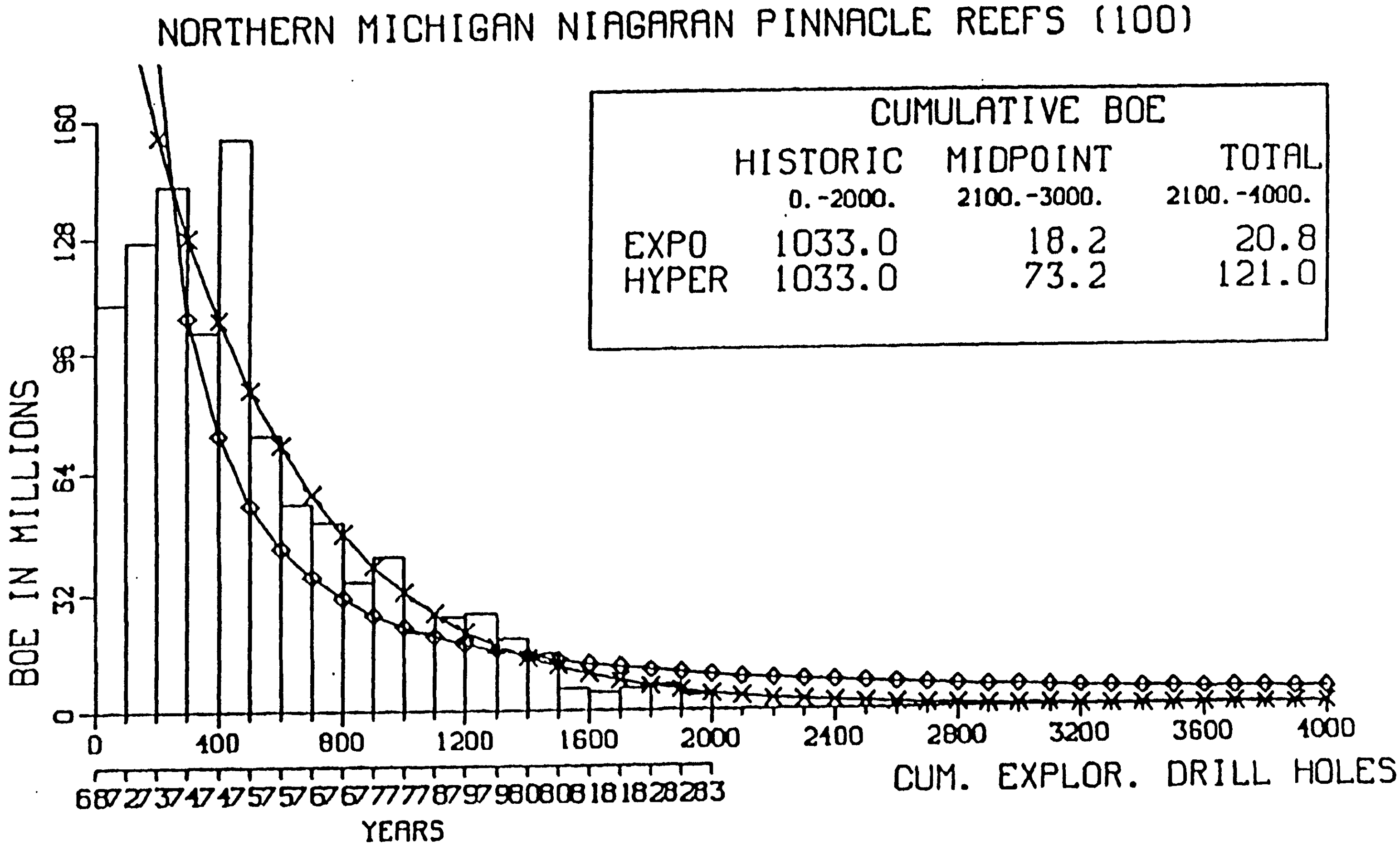
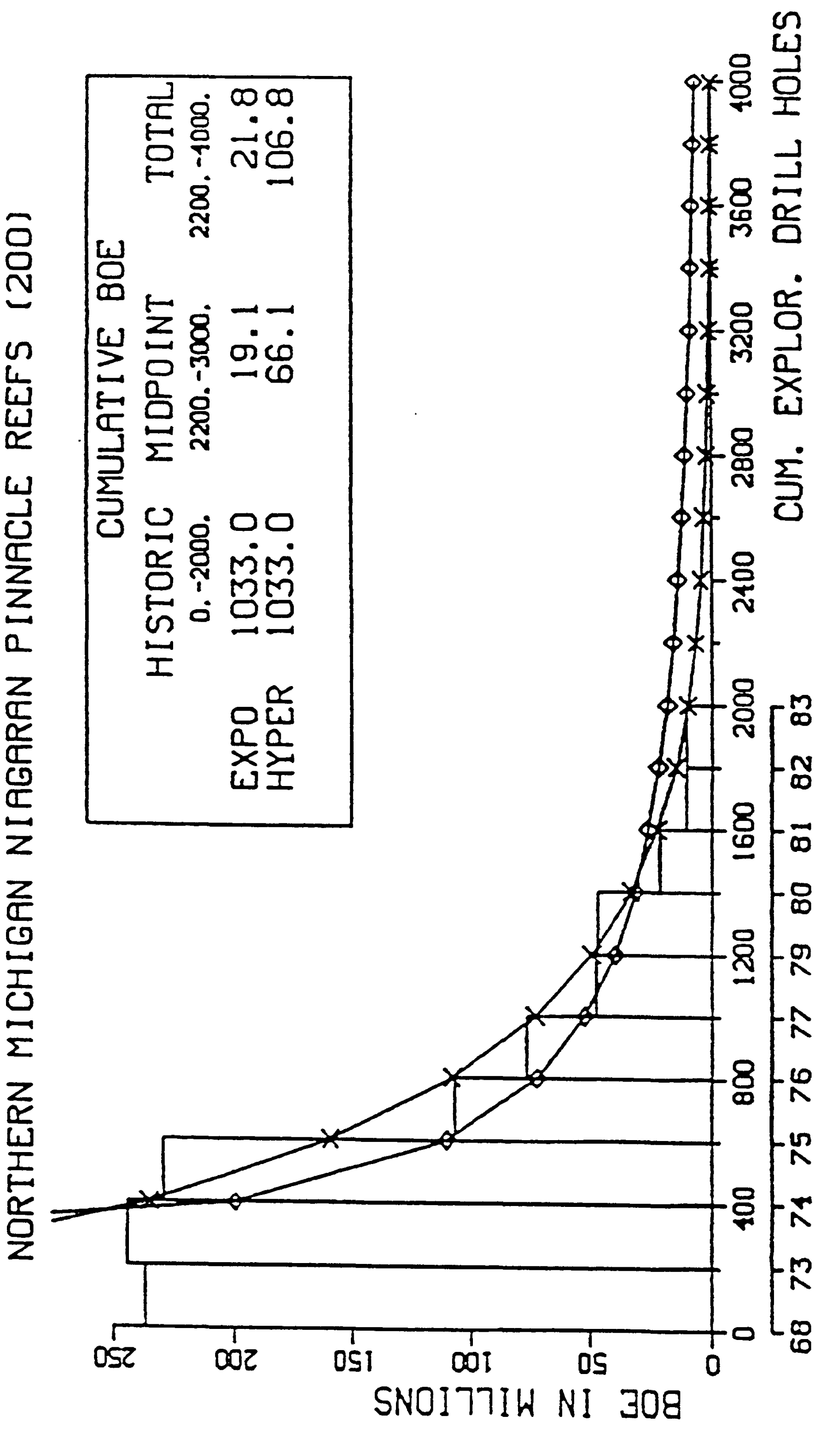
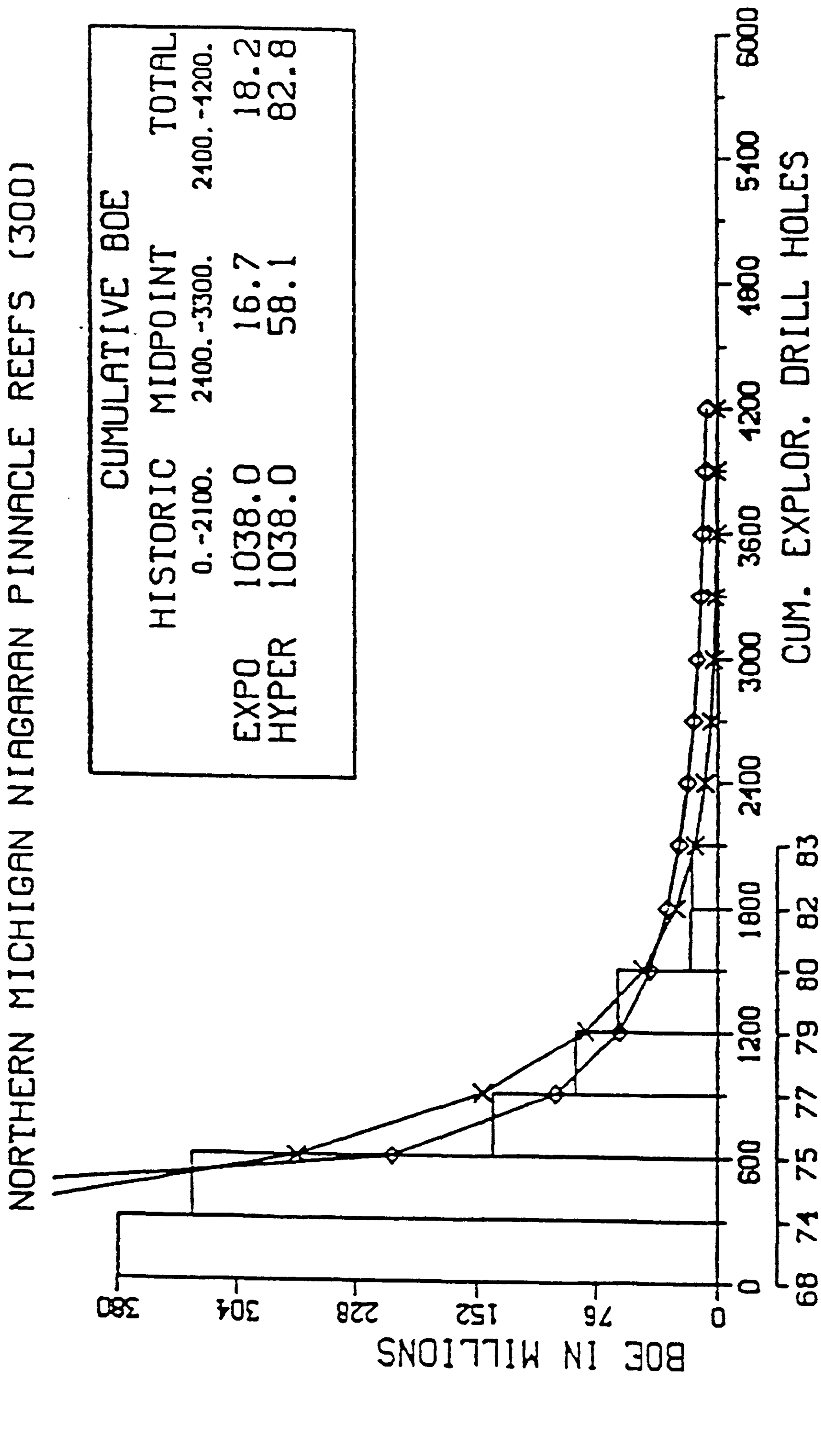


Figure 7.--Example of finding-rate curves showing extrapolation of exponential and hyperbolic curves. Historical data (data from Illinois basin) is from 1944 to 1976. The areas under the projected curves represent estimates of undiscovered recoverable oil to be found with the next 60 million feet of exploratory drilling. For the exponential curve A, the estimated amount of undiscovered recoverable oil is 0.038 billion barrels, and for the hyperbolic curve B, it is 0.115 billion barrels.

Source: Dolton and others, 1981


NORTHERN MICHIGAN NIAGARAN PINNACLE REEFS (100)
CUMULATIVE BOE
HISTORIC 0.-2000.
MIDPOINT 2100.-3000.
TOTAL 2100.-4000.
EXPO 1033.0 18.2 20.8
HYPER 1033.0 73.2 121.0
BOE IN MILLIONS
0 32 64 96 128 160
0 400 800 1200 1600 2000 2400 2800 3200 3600 4000
CUM. EXPLOR. DRILL HOLES
YEARS


NORTHERN MICHIGAN NIAGARAN PINNACLE REEFS (200)
CUMULATIVE BOE
HISTORIC MIDPOINT TOTAL
0.-2000. 2200.-3000. 2200.-4000.
EXPO 1033.0 19.1 21.8
HYPER 1033.0 66.1 106.8
BOE IN MILLIONS
0
50
100
150
200
250
0 400 800 1200 1600 2000 2400 2800 3200 3600 4000
CUM. EXPLOR. DRILL HOLES
68 73 74 75 76 77 79 80 81 82 83
YEARS


NORTHERN MICHIGAN NIAGARAN PINNACLE REEFS (300)
CUMULATIVE BOE
HISTORIC
0.-2100.
MIDPOINT
2400.-3300.
TOTAL
2400.-4200.
EXPO 1038.0 16.7 18.2
HYPER 1038.0 58.1 82.8
BOE IN MILLIONS
0
76
152
228
304
380
0
600
1200
1800
2400
3000
3600
4200
4800
5400
6000
CUM. EXPLOR. DRILL HOLES
68
74
75
77
79
80
82
83
YEARS

Power laws

Power laws are associated with fractals.

Power function:

$$y = \alpha x^{\beta}$$

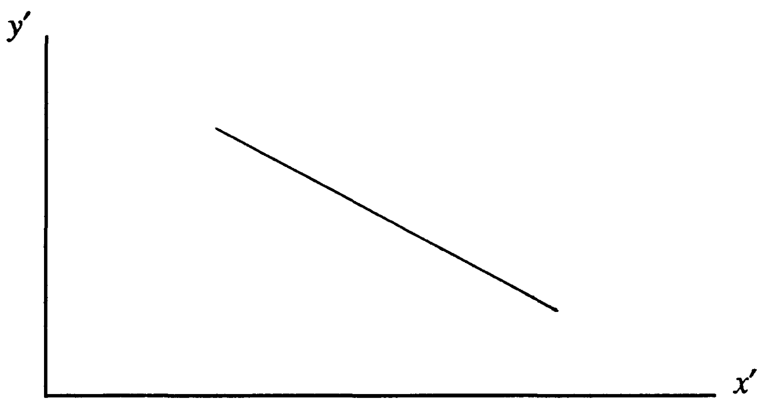
Taking logarithms,

$$\log y = \log \alpha + \beta \log x$$

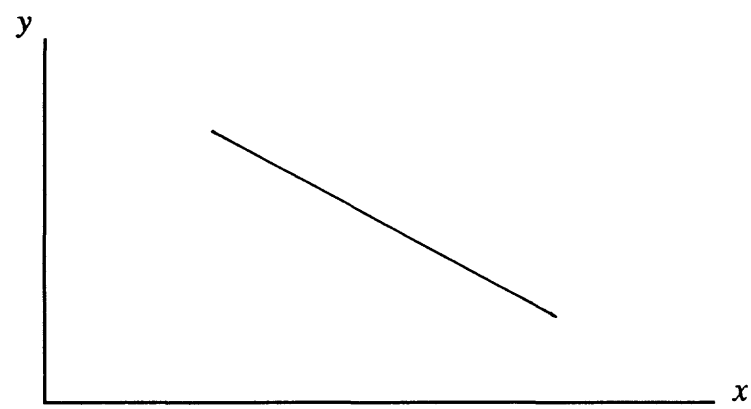
Obtain linear form,

$$y' = \alpha' + \beta x'$$

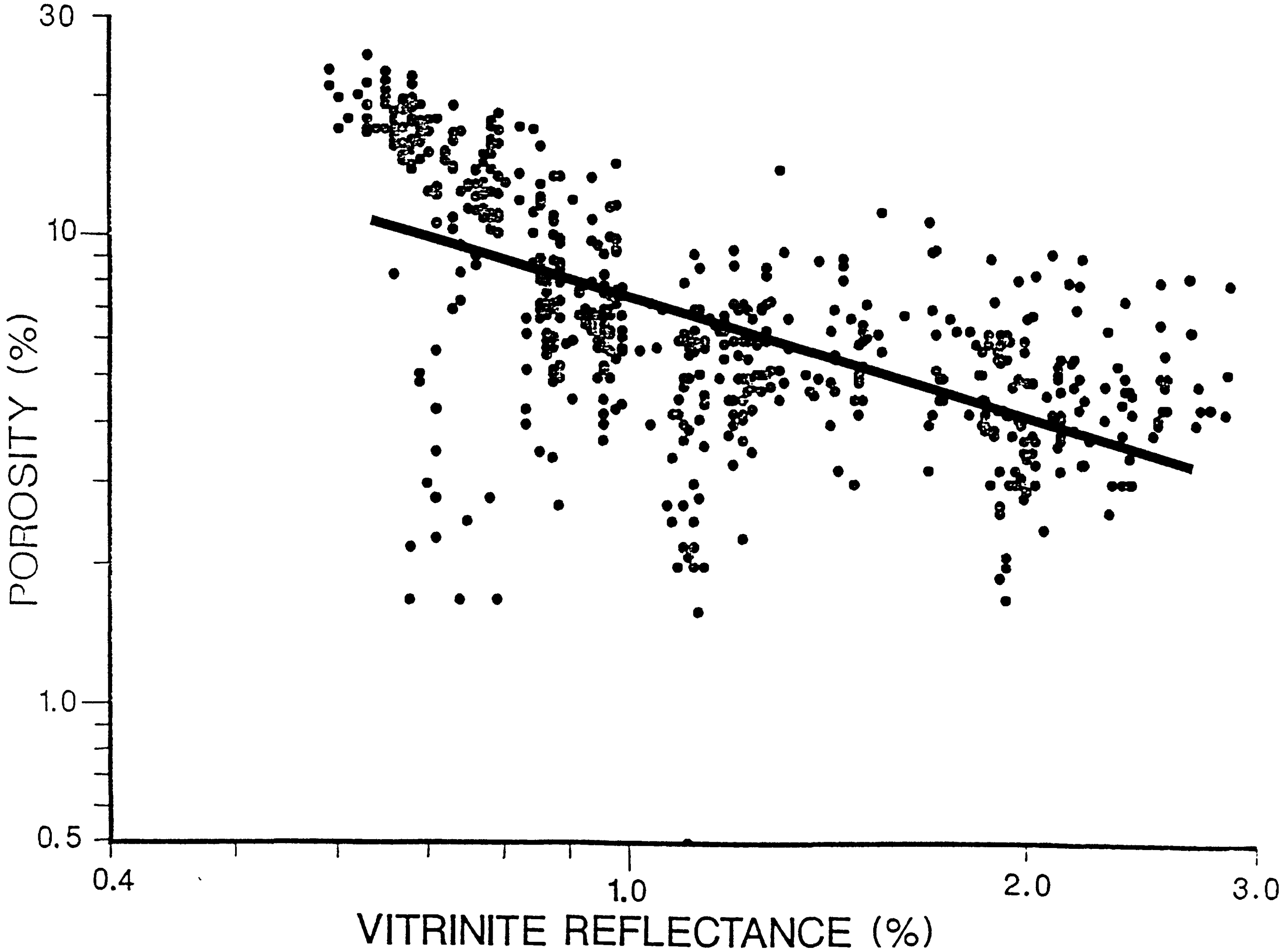
Same line in each case:



1. Both axes have arithmetic scales



2. Both axes have logarithmic scales



Source: Hester

e. Fractals

If the cumulative frequency distribution (more than) is plotted on log-log axes, a linear pattern of points suggests a Pareto (or fractal) distribution.

Even if the linear pattern of points is restricted to the right side of the cumulative frequency distribution due to an economic truncation, a Pareto (or fractal) distribution is suggested as a model for the parent population.

The Pareto distribution theory for this approach is established in Crovelli and Barton (submitted).

Fit a straight line to the linear pattern on the right side of the distribution.

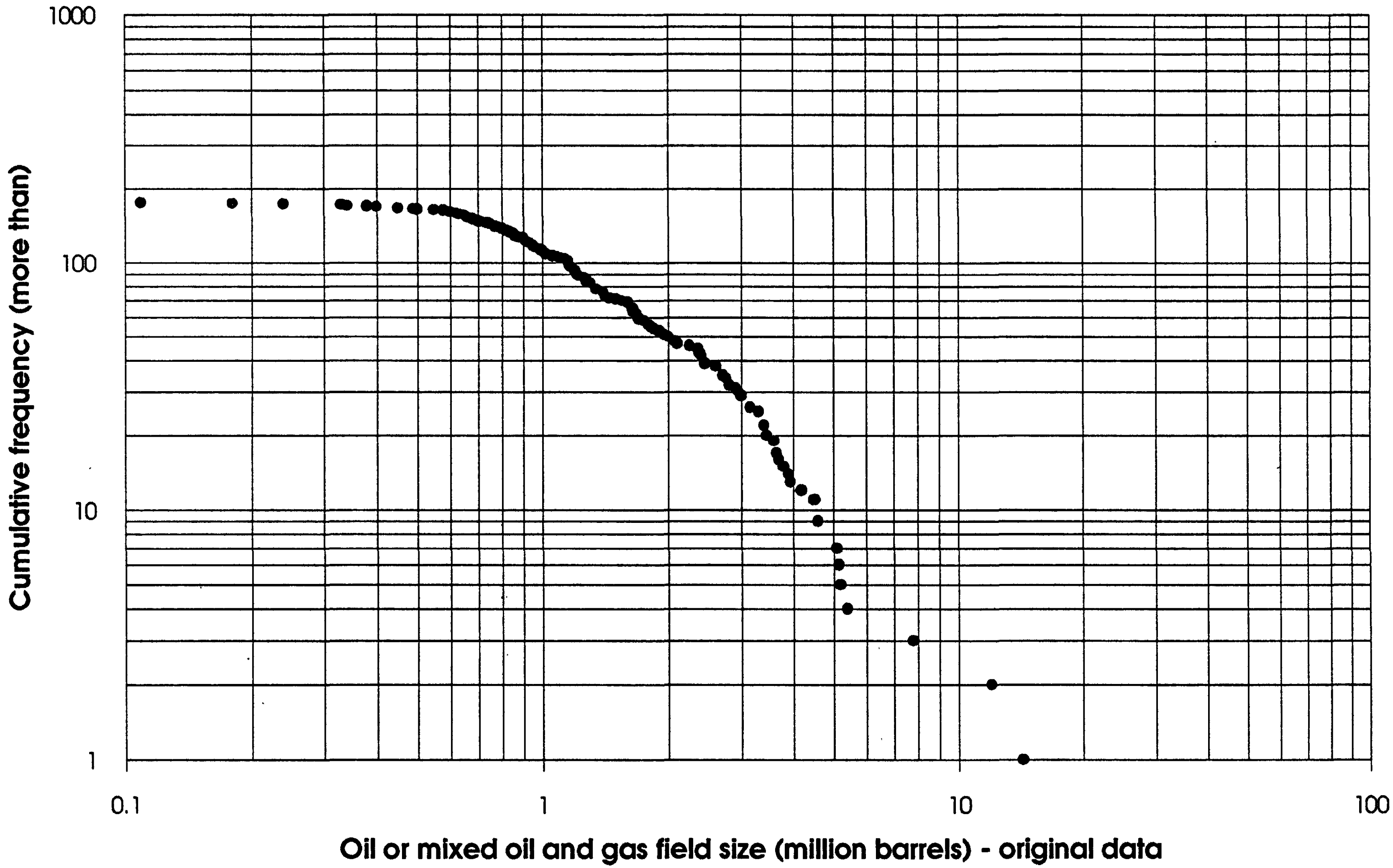
Model: $N(x) = \alpha x^{\beta} \qquad x > 0$

where $N(x)$ is the cumulative frequency distribution (more than).

Cumulative frequency distribution (more than) - original data of oil and mixed oil and gas field sizes

Boundary	Cumulative Frequency (more than)	Boundary	Cumulative Frequency (more than)	Boundary	Cumulative Frequency (more than)
0.11	175	0.99	113	2.08	48
0.18	174	1.00	112	2.10	47
0.24	173	1.02	109	2.25	46
0.33	172	1.06	107	2.35	45
0.34	171	1.08	106	2.38	43
0.38	170	1.10	105	2.40	42
0.40	169	1.13	104	2.45	39
0.45	167	1.14	103	2.60	38
0.49	166	1.15	102	2.70	35
0.50	165	1.16	98	2.75	34
0.55	164	1.17	96	2.80	32
0.58	163	1.18	95	2.90	31
0.60	161	1.20	93	2.95	30
0.62	159	1.21	90	3.00	29
0.63	158	1.22	89	3.15	26
0.65	156	1.26	87	3.30	25
0.66	153	1.27	86	3.40	22
0.68	151	1.28	84	3.45	20
0.70	148	1.30	83	3.60	19
0.71	147	1.35	78	3.65	17
0.73	146	1.40	76	3.70	16
0.74	145	1.42	73	3.80	15
0.75	144	1.45	72	3.90	14
0.77	141	1.50	71	3.95	13
0.78	140	1.55	70	4.20	12
0.80	138	1.60	69	4.50	11
0.82	135	1.65	65	4.60	9
0.83	134	1.66	63	5.10	7
0.84	132	1.68	62	5.15	6
0.85	131	1.70	60	5.20	5
0.86	129	1.71	59	5.40	4
0.87	128	1.75	58	7.80	3
0.88	127	1.80	56	12.00	2
0.90	126	1.82	55	14.25	1
0.91	123	1.85	54		
0.93	121	1.90	53		
0.95	118	1.95	51		
0.96	116	2.00	50		

Cumulative frequency distribution



Computation of sums for
transformation of power function to linear form

x_i	y_i	$x_i' = \log x_i$	$y_i' = \log y_i$	$(x_i')^2$	$(y_i')^2$	$x_i'y_i'$
2.35	45	0.371	1.653	0.138	2.733	0.613
2.38	43	0.377	1.633	0.142	2.668	0.615
2.40	42	0.380	1.623	0.145	2.635	0.617
2.45	39	0.389	1.591	0.151	2.531	0.619
2.60	38	0.415	1.580	0.172	2.496	0.656
2.70	35	0.431	1.544	0.186	2.384	0.666
2.75	34	0.439	1.531	0.193	2.345	0.673
2.80	32	0.447	1.505	0.200	2.265	0.673
2.90	31	0.462	1.491	0.214	2.224	0.690
2.95	30	0.470	1.477	0.221	2.182	0.694
3.00	29	0.477	1.462	0.228	2.139	0.698
3.15	26	0.498	1.415	0.248	2.002	0.705
3.30	25	0.519	1.398	0.269	1.954	0.725
3.40	22	0.531	1.342	0.282	1.802	0.713
3.45	20	0.538	1.301	0.289	1.693	0.700
3.60	19	0.556	1.279	0.309	1.635	0.711
3.65	17	0.562	1.230	0.316	1.514	0.692
3.70	16	0.568	1.204	0.323	1.450	0.684
3.80	15	0.580	1.176	0.336	1.383	0.682
3.90	14	0.591	1.146	0.349	1.314	0.677
3.95	13	0.597	1.114	0.356	1.241	0.665
4.20	12	0.623	1.079	0.388	1.165	0.673
4.50	11	0.653	1.041	0.427	1.084	0.680
4.60	9	0.663	0.954	0.439	0.911	0.632
5.10	7	0.708	0.845	0.501	0.714	0.598
5.15	6	0.712	0.778	0.507	0.606	0.554
5.20	5	0.716	0.699	0.513	0.489	0.500
5.40	4	0.732	0.602	0.536	0.362	0.441
7.80	3	0.892	0.477	0.796	0.228	0.426
12.00	2	1.079	0.301	1.165	0.091	0.325
14.25	1	1.154	0.000	1.331	0.000	0.000
	Total =	18.132	36.475	11.670	48.240	18.997

Computation of sums for
transformation of power function to linear form

n = 31

$\sum x_i' =$ 18.132

$\sum y_i' =$ 36.475

$\sum (x_i')^2 =$ 11.670

$\sum (y_i')^2 =$ 48.240

$\sum x_i' y_i' =$ 18.997

$$b = \frac{n\sum x_i' y_i' - \left(\sum x_i'\right)\left(\sum y_i'\right)}{n\sum (x_i')^2 - \left(\sum x_i'\right)^2} = \quad -2.194$$

|b| = 2.194 is an estimate of the shape parameter of the Pareto distribution.

$$a' = \bar{y}' - b\bar{x}' = \quad 2.460$$

$$y' = a' + bx' = 2.460 - 2.194x'$$

$$a = 10^{a'} = \quad 288.246$$

$$y = ax^b = 288.246x^{-2.194}$$

x = 2.35, y = 44.230
x = 13, y = 1.038

$$r = \frac{n\sum x_i' y_i' - \sum x_i' \sum y_i'}{\sqrt{\left[n\sum (x_i')^2 - \left(\sum x_i'\right)^2\right]\left[n\sum (y_i')^2 - \left(\sum y_i\right)^2\right]}} = \quad -0.981$$

$r^2 =$ 0.963

Selected References

Arps, J.J. and Roberts, T.G., 1958, Economics of drilling for Cretaceous oil on east flank of Denver-Julesburg Basin: American Association of Petroleum Geologists Bulletin, v. 42, p. 2549-2566.

Barouch, E. and Kaufman, G.M., 1976, Oil and gas discovery modelled as sampling without replacement and proportional to random size: Sloan School Working Paper No. 888-76.

Charpentier, R.R., 1989, A statistical analysis of the larger Silurian reefs in the northern part of the Lower Peninsula of Michigan: U.S. Geological Survey Open-File Report 89-216, 34 p.

Crovelli, R.A. and Balay, R.H., 1990, PROBDIST: Probability distributions for modeling and simulation in the absence of data: U.S. Geological Survey Open-File Report 90-446-A, Documentation (paper copy) 51 p.; Open-File Report 90-446-B, Executable program (5.25" diskette).

Crovelli, R.A. and Balay, R.H., 1992, LOGRAF: Lognormal graph for resource assessment forecast: U.S. Geological Survey Open-File Report 92-679-A, Documentation (paper copy) 30 p.; Open-File Report 92-679-B, Executable program (5.25" diskette).

Crovelli, R.A. and Barton, C.C. [submitted], Fractals and the Pareto distribution applied to petroleum field-size distributions, *in* Barton, C.C. and LaPointe, P.R., eds., Fractal geometry and its uses in the petroleum industry: American Association of Petroleum Geologists Book Series Volume, 26 ms. p., 10 figs. (Director's approval 28-Jan-1991, and BTR log-in 26-Dec-1990.)

Crovelli, R.A., 1984, Procedures for petroleum resource assessment used by the U.S. Geological Survey--statistical and probabilistic methodology, *in* Masters, C.D., ed., Petroleum resource assessment: International Union of Geological Sciences, pub. no. 17, p. 24-38.

Crovelli, R.A., 1984, U.S. Geological Survey probabilistic methodology for oil and gas resource appraisal of the United States: Journal of the International Association for Mathematical Geology, v. 16, no. 8, p. 797-808.

Dolton, G.L., Carlson K.H., Charpentier, R.R., Coury, A.B., Crovelli, R.A., Frezon, S.E., Khan, A.S., Lister, J.H., McMullin, R.H., Pike, R.S., Powers, R.B., Scott, E.W., and Varnes, K.L., 1981, Estimates of undiscovered recoverable conventional resources of oil and gas in the United States: U.S. Geological Survey Circular 860, 87 p.

Drew, L.J., 1990, Oil and gas forecasting--Reflections of a petroleum geologist: New York, Oxford University Press, International Association for Mathematical Geology, Studies in Mathematical Geology No. 2, 252 p.

Harbaugh, J.W., Doveton, J.H., and Davis, J.C., 1977, Probability methods in oil exploration: New York, John Wiley, 269 p.

Hester, T.C., [submitted] Porosity trends of Pennsylvanian sandstones with respect to thermal maturity and thermal regimes in the Anadarko basin, Oklahoma: U.S. Geological Survey Bulletin. (Branch Chief approval and CTR log-in)

Hunter, R.L. and Mann, C.J., eds., 1992, Techniques for determining probabilities of geologic events and processes: New York, Oxford University Press, International Association for Mathematical Geology, Studies in Mathematical Geology, No. 4, 364 p.

Mast, R.F., Dolton, G.L., Crovelli, R.A., Root, D.H., Attanasi, E.D., Martin, P.E., Cooke, L.W., Carpenter, G.B., Pecora, W.C., and Rose, M.B., 1989, Estimates of undiscovered conventional oil and gas resources in the United States--A part of the Nation's energy endowment: U.S. Geological Survey and Minerals Management Service, 44 p.

Newendorp, P.D., 1975, Decision analysis for petroleum exploration: Tulsa, Oklahoma, Petroleum Publishing Company, 668 p.

NRG Associates, Inc., 1992, The significant oil and gas fields of the United States [through December 31, 1990]: Available from Nehring Associates, Inc., P.O. Box 1655, Colorado Springs, Colorado 80901.

Walpole, R.E. and Myers, R.H., 1989, Probability and statistics for engineers and scientists: New York, Macmillan Publishing Company, 4th ed., 765 p.

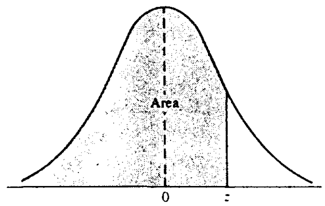


Table A.1 Areas Under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Table A.1 *(continued)* Areas Under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.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	.9278	.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

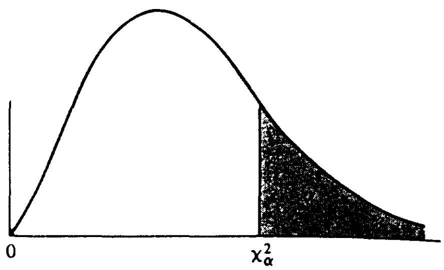


Table A.2 Critical Values of the Chi-Squared Distribution

ν	α = .995	.99	.98	.975	.95	.90	.80	.75	.70	.50
1	$.0^4393$	$.0^3157$	$.0^3628$	$.0^3982$	.00393	.0158	.0642	.102	.148	.455
2	.0100	.0201	.0404	.0506	.103	.211	.446	.575	.713	1.386
3	.0717	.115	.185	.216	.352	.584	1.005	1.213	1.424	2.366
4	.207	.297	.429	.484	.711	1.064	1.649	1.923	2.195	3.357
5	.412	.554	.752	.831	1.145	1.610	2.343	2.675	3.000	4.351
6	.676	.872	1.134	1.237	1.635	2.204	3.070	3.455	3.828	5.348
7	.989	1.239	1.564	1.690	2.167	2.833	3.822	4.255	4.671	6.346
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	5.071	5.527	7.344
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	5.899	6.393	8.343
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	6.737	7.267	9.342
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	7.584	8.148	10.341
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	8.438	9.034	11.340
13	3.565	4.107	4.765	5.009	5.892	7.042	8.634	9.299	9.926	12.340
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	10.165	10.821	13.339
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	11.036	11.721	14.339
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	11.912	12.624	15.338
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	12.792	13.531	16.338
18	6.265	7.015	7.906	8.231	9.390	10.865	12.857	13.675	14.440	17.338
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	14.562	15.352	18.338
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	15.452	16.266	19.337
21	8.034	8.897	9.915	10.283	11.591	13.240	15.445	16.344	17.182	20.337
22	8.643	9.542	10.600	10.982	12.338	14.041	16.314	17.240	18.101	21.337
23	9.260	10.196	11.293	11.688	13.091	14.848	17.187	18.137	19.021	22.337
24	9.886	10.856	11.992	12.401	13.848	15.659	18.062	19.037	19.943	23.337
25	10.520	11.524	12.697	13.120	14.611	16.473	18.940	19.939	20.867	24.337
26	11.160	12.198	13.409	13.844	15.379	17.292	19.820	20.843	21.792	25.336
27	11.808	12.879	14.125	14.573	16.151	18.114	20.703	21.749	22.719	26.336
28	12.461	13.565	14.847	15.308	16.928	18.939	21.588	22.657	23.647	27.336
29	13.121	14.256	15.574	16.047	17.708	19.768	22.475	23.567	24.577	28.336
30	13.787	14.953	16.306	16.791	18.493	20.599	23.364	24.478	25.508	29.336

Table A.2 (*continued*) Critical Values of the Chi-Squared Distribution

ν	α									
	.30	.25	.20	.10	.05	.025	.02	.01	.005	.001
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	7.231	7.841	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	8.383	9.037	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	9.524	10.219	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	10.656	11.389	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	11.781	12.549	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	12.899	13.701	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	14.011	14.845	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	15.119	15.984	16.985	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	16.222	17.117	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.123
15	17.322	18.245	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	18.418	19.369	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	19.511	20.489	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	20.601	21.605	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	21.689	22.718	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	22.775	23.828	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.315
21	23.858	24.935	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.797
22	24.939	26.039	27.301	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	26.018	27.141	28.429	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	27.096	28.241	29.553	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	28.172	29.339	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.620
26	29.246	30.434	31.795	35.563	38.885	41.923	42.856	45.642	48.290	54.052
27	30.319	31.528	32.912	36.741	40.113	43.194	44.140	46.963	49.645	55.476
28	31.391	32.620	34.027	37.916	41.337	44.461	45.419	48.278	50.993	56.893
29	32.461	33.711	35.139	39.087	42.557	45.722	46.693	49.588	52.336	58.302
30	33.530	34.800	36.250	40.256	43.773	46.979	47.962	50.892	53.672	59.703

Index

R

S