DETROIT RESEARCH INSTITUTE

STATISTICAL BULLETIN

VOL. I, BULLETIN 7, SUPPLEMENT I, JAN. 1972

BIOMEDICAL STATISTICS

I Statistical Design

It is the purpose of this paper to introduce into medical research, both basic and clinical, some statistical methodology and procedures that are new in this area. The techniques are not new, some having been used in industry for several years, but their application to the biomedical area has been with a few exceptions neglected.

The rewarding attributes of these procedures are that they are realistic in their usefulness in "real life" situations, they are effective when dealing with small samples from large and unknown populations, they can and often do make use of intuition or previous information. They are also quite simple to apply "cook book" style, yet are understandable from a theoretical view point without requiring statistical background.

The procedures that are to be emphasized are methods of using order statistics, methods for determining the parameters of a three parameter Weibull distribution, and methods for using this distribution both in statistical analysis and statistical inference. In applying these methods to real problems there will be of necessity some discussion of Bayesian statistics or the art of using personal probabilities in the evaluation of problems. We do not wish to get involved in an argument concerning the semantics of using the word Bayesian in regard to intuitive or personal probabilities. Nor do we wish to argue with "classic" statisticians as to the rightness or wrongness in the Bayesian approach. We hope it will become evident that in solving a real problem in clinical medicine, the concept of a personal probability cannot be ignored.

This paper has been written primarily for the clinical investigator. While it is not our purpose to make a statistician of him, it is our purpose to make it clear to him that "statistics" concerning his work are of use only to him and no one else. Statistics is a very personal thing. The only function that statistics perform is to give him a sense of confidence about his feeling of certainty or uncertainty about some problem he is facing. The problem he is facing may be a universal problem; however, the statistics concerning the problem is the statistics of his feeling about the problem and all its aspects. The statistics then is not universal. These last statements must be modified. If the approach of classical statistics is used the statements generally hold true. If, on the other hand, order statistics is used, much can be said about the certainty or uncertainty of the information or data that is available about the problem under consideration regardless of the feeling or assumptions made by the investigator.

What are the requirements then for the statistical design of an experiment? The most important requirement is just good common sense. The entire experiment must be looked at with proper perspective. What is the purpose of the research, what will it cost in time and money, are the possible rewards worth the effort, and what is the methodology to be used? The questions must first be answered in an informal yet orderly and common sense fashion. Then a more rigorous approach to the design must be established. What is the purpose of the research? We here assume that with today's costs in time and money, a research project does have a real purpose, at least as far as the investigator is concerned. He must then, for a realistic pursuance of the research problem, state that problem in a concise, consistent and somewhat rigid manner. It is here that we first meet the concept of personal probability. The investigator has met some problem through his experience that challenges him. It is because he has some personal views of the problem that he wishes to pursue it to establish for his own benefit whether his ideas

are right or wrong. This then is the second requirement in a statistical design, a concise and proper statement of the problem.

The third requirement is the formulation of a plan of action based on the results of the solution of the problem. This requirement is seldom met in medical research and if it is considered it is usually on a haphazard basis. With today's economy and cost consciousness about research, a project which when completed does not provide a definite decision about further activity based on the results of the research has little or nothing to merit its undertaking. The pursuant action may be a simple decision to desist from further concern about the problem; it may be a decision to continue activity in a different direction; or it may be a decision to implement a new clinical procedure or treatment. In any case the rules for arriving at these decisions must be established before the research takes place. Unfortunately, this is not a simple step. In the current practice of classical biostatistics, emphasis is placed on preventing the investigator from rejecting a hypothesis when it is true, but little is done to protect him from accepting the hypothesis when it is false. This "significance" testing plays a major role in the current practice of statistical inference. This practice has many pitfalls, both in the use of "significance levels" as a basis of hypothesis testing and in the acceptance of several "assumptions" in applying the statistical models to the data available. It is in this area that the investigator has failed the most in his responsibility to the decision process. He must assign a "utility" to a correct decision and a "risk" to a wrong decision. Only he can do that and again we meet the reality of a personal probability, a personal feeling of how useful is the utility and how risky is the risk. This problem becomes much less of a problem if we consider hypothesis in the light of "confidence" rather than "significance". By doing this we consider at the same time not only the risk of accepting as wrong, what is right, but also the risk of accepting as right, what is wrong.

Order statistics as we shall see can be a very powerful tool in establishing these "confidences" and the further use of the Weibull distribution can give added latitude in their applications.

The next requirement for a statistical design is the method of collecting data or information to be used in solving the problem. It must be remembered that data collection devices are not, in the real world, noise free. In other words, when you collect data from an experiment that data represents, both the information that is generated by experiment, plus the noise generated by the data collector. This data collector may be instruments, people or a combination of both. In any case, it will have some noise. The statistical design then will have a data collection device that is as simple as it can be made to eliminate sources of noise. Also, the investigator must know the system so that he can best appraise the noise that is present.

It follows that the next stage in the method is one that provides a method of separating the information from the noise in the data. We must turn to statistical methods for this operation. We wish to call this statistical analysis. We expect from this operation to have some insight into the character of the information. The noise should be filtered out. We would hope to establish some order to the data, different modes should be identified if they are present and hopefully some knowledge of the distribution functions underlying the samples could be suggested.

The final step in the statistical design we shall call statistical inference. It is with this step that we come to conclusions about comparisons of our samples of information concerning the problem as it was stated. At this stage, we relate our data samples to statistical mathematical models and by inductive reasoning come to some conclusions, with a stated amount of certainty or uncertainty about the real world these samples represent.

II Order Statistics

It is time to discuss order statistics. First, however, we should generalize on statistics as used in analysis and inference. Statistics is used to make inferences about populations that are so large as to be for practical purposes immeasurable. We take, therefore, samples from these populations, measure the samples and by statistics refer the sample measurements to the total population. With classical statistics, this is at best a risky business. For example, we assume that the population does not change with time, what's true today is true tomorrow. We identify these populations by one or more characteristics that attract our interest. We measure these characteristics in the sample and we assume that they are independent of any other characteristics other than the ones that are measured in the sample. In other words, if we find that a certain measurement taken from the sample is independent of all other measurements of the sample, then we assume the same holds true of the entire population; i.e., that our sample is truly representative of the population in all respects. It takes courage to make this assumption.

Having taken a set of measurements of a sample, we assume that the total population has a similar set of measurements and these have certain identifying features that we call parameters. This set of population measurements we will call the distribution function and thus we have the parameters of the distribution function. These are used to identify one distribution function from another. As has been said, the populations are so large as to be immeasurable so that we cannot know the exact values of these parameters. We can, however, estimate them from the parameters of our samples. Whether or not these estimates are realistic, we will never know; however, to use classical statistics we say they are and let it go at that.

The parameters that we are talking about are named mean, variance, skewness and kurtosis. The mean represents a balance point in the distribution function, a fulcrum about which all the measurements of the population are balanced as to weight. If this fulcrum happens to coincide with a point in the distribution that equals half the total members of the population, we say that we have a symetrical distribution, if it does not, we say the distribution is skewed (another parameter).

Since nothing in this world seems perfect, so are the members of a population imperfect. They all differ, or vary from one another. The amount of this difference, or variance, is characteristic for a given population, so their distribution functions can be characterized by this variance (another parameter).

Classical statistics assumes that these parameters are known or correctly estimated and also they are the best descriptors of the population. If the investigator truly believes this, then classical statistical inference will provide him with more assurance of certainty or uncertainty about his problem based on his data. If he is in doubt about these assumptions, either because of past experience (personal probability again) or just plain reluctance to accept so many assumptions about a large unknown population, the classical statistical inference is of little assistance to him in his real problem.

In the classical approach to statistics, we define some population and we assign to this population a set of descriptors or parameters (mean, variance, etc.) Then we try to select a representative sample from this population so that we can estimate its parameters from the sample, as has been pointed out this can be risky business. In using order statistics we approach the problem from a different direction. We first collect a sample, any set of values of some unknown population. If we have such a sample, we assume only that the sample itself is

not the total population but has been drawn from some population. No further assumptions are made about the population, except that it can be made countable.

Having done this, the population thus must have percentiles. In other words, it can be divided into fractions, a fourth of the population, a half, three quarters, etc. Since this population is too large to count, we must estimate its percentiles from the sample. We can count the sample and divide it into percentiles. It is not realistic to believe that a percentile of the sample would exactly or even closely represent a corresponding percentile of the population. Since we assumed nothing about the population, our first approximation of what percentile of the population is represented by a percentile of the sample must be pure chance (a 50% probability).

The operation for achieving this is quite simple. First, each member of the sample must be ordered according to its value with its neighbor, the smallest being first, etc. Hence, we have a set of order statistics. A note should be made here concerning the value of each ordered statistic. Since we are dealing with percentiles of the population, the exact value of each member of the sample does not have to be known exactly. It is only required that it be known whether or not it is larger or smaller than its neighbor. This consideration can be most useful in clinical medicine where numerical values are sometimes difficult to establish, such as in grading of reflexes, amount of pain, etc.

To return to the problem at hand we have ordered the data set. For computational purpose we will label each member j, so we have a set of j order statistics and we will call the order number of each j, k. We can rank each member of the sample then by simply computing:

$$\text{sample rank} = \frac{j_k}{n}$$

where n = sample size

We have said, however, that this would not be a very realistic guess as to where the sample member j_k would rank in the total population. It turns out that a simple calculation:

$$\text{median rank} = \frac{j_k - .3}{n + .4} \text{ , where } n = \text{sample size}$$

yields the median rank of j_k in the total population. That is it establishes what percentile of the total population the sample member represents with a 50% probability of its actual rank being either higher or lower. The calculation of the actual median rank and its mathematical derivation is quite complicated and you are referred to an article by L.G. Johnson for a complete description (1). (The portion of this article dealing with the mathematical derivation of median ranks is included in Appendix III.)

Empirically it has been determined that when a sample thus ranked is plotted on Weibull probability paper (the abscissa is log and the ordinate is log log), and it produces a straight line, the population has a Weibull distribution function. When this is the case, it enables the use of the Weibull function in many techniques of statistical inference. We shall discuss these methods later.

To return to order statistics, we have thus far determined the pure chance ranking of our sample. Let us proceed to other probability levels for ranking the sample into the total population. For example, we shall take the 5% and 95% ranks (in clinical research a ranking of 1% and 99% may be preferred). For the 5% rank we shall determine at what percentile of the population a sample member has only a 5% probability being less.. For the 95% rank, we shall determine the percentile of the population that the sample has a 95% probability of being less.

Now we have drawn confidence bands for the population from which our sample was drawn. We have established with a 90% probability, the range of the popula-

tion from which our sample was drawn. This gives us a tool for making statistical inference about other samples we may have and their relation to the original sample. In the process, we have made no assumptions about the population other than it does exist and that it is countable.

The calculation of population ranks other than the median rank is not simple but is mathematically sound. There are computer programs available that will calculate any rank for any size sample. (2)

This then becomes a very powerful statistical tool for medicine because it makes no requirements about stationarity, apriori and repeatable probabilities, or functional dependence of the assumed population.

III The Weibull Distribution

We have just discussed order statistics and its usefulness because it is not required to know the distribution function and its parameters. Before that we talked about distribution functions in general and their parameters or identifying features. The point was made (hopefully) that in the game of statistics the problem is matching, by inductive reasoning, a mathematical statistical model to real life data. In classical statistics, this is difficult, if not impossible. We just arbitrarily say we have done so by saying that the data has been drawn from a population with a "normal" distribution. The risk of doing this has already been emphasized.

In 1939 Waloddi Weibull postulated a very general cumulative distribution function whose only requirements were that it be non-zero and non-decreasing. In other words, the probability of accounting for the total population was ever increasing as you "added up" your total sample. This is done very simply by first ordering the sample, just as in order statistics and then determining the median rank of each member of the sample. Then a linear relationship holds between the

logarithm of the values of the order statistics and the logarithm of the logarithm of the cumulative percent of the population. It has been shown by experience, and this after all is the true test of worth, that data collected from experiments in industry and biology do fit a Weibull distribution (3) (4) (5).

As we learn more about this distribution it is not difficult to understand why Weibull called it a function of broad applicability for it is, in reality, a whole family of distribution functions. The "normal" distribution function is a Weibull distribution function, and so is the exponential. Thus, it is not only a very useful model for statistical inference but it also is useful in determining whether or not your data fits a more well known distribution function.

The Weibull distribution also has parameters or characteristics which describe it exactly. The first of these is the location parameter. The location parameter describes the point of origin of the probability density function. We will call this parameter alpha (α). The second Weibull parameter is the shape parameter. This parameter which we will call beta (β) gives a numerical value which equates to the general shape of the probability density function. The third parameter is the scale parameter, theta (θ) which defines the value of your sample (x) at which 63.2 percent of your density function has been accounted.

Since the shape parameter, beta, describes the shape of the probability density function, it is the slope of the linear function of the cumulative distribution function. It then becomes a simple procedure to estimate all three Weibull parameters. The best alpha will give the best linear fit of the data. This determines the beta and the theta is then self defined. From these parameters, the more familiar parameters of your distribution can be evaluated such as mean (mu), variance (sigma squared) and skewness.

Previously we discussed the use of statistics for evaluating the data and separating the information from the noise. The Weibull distribution function is of considerable use in this respect particularly when we use its graphical properties. If we have a sample set of data that represents a single function it will plot on Weibull probability paper (abscissa-log, ordinate log-log) as a straight line. If, however, the data represents more than one function, the Weibull plot will be a mixture of straight lines. (3) By trial and error method the sample points can be separated and replotted until each is identified with its particular function. This is a simple yet proven method of separating a mixed signal. This method can also be used to determine whether or not extreme values in a sample really belong to the population under study and if they should be discarded from further consideration.

For statistical inference there is a large set of procedures using the Weibull parameters that can be used in hypothesis testing, comparison of samples and predictions about sampling. This, as can be seen, is a very powerful function. There is no need to assume a normal distribution for the data. Whether or not it is normal can be determined by using the Weibull parameters. Therefore, by using the Weibull function there is available the techniques of "classical" statistical inference as well as many others which can be used when the classical techniques are inappropriate.

IV Bayesian Statistics

The last idea to be discussed is the idea of Bayesian statistics, in particular, the concept of "personal probability". There are many arguments concerning the Bayesian approach to statistics even among the Bayesians themselves; (6) however, in the realm of clinical investigation the idea and use of personal probability should not and cannot be avoided. The use of a statistical evaluation of a clinical experiment is simply an extension of a physicians intuition about some

problem based upon some defined observations. By the very institution of the research, the organizing of the problem and the method of data collection are unavoidably influenced by the clinician's intuition. It is more than proper that this intuition should be properly utilized in the follow through of the project. Indeed, in the final utilization of the results of such experimentation, the experiment will only prove useful if in fact the results can be made compatible or acceptable to the clinicians prior probability (intuition) of these results.

The arguments of classical statistics for the presence of apriori probabilities, repeatable probabilities and independence in the complicated populations of the real world are really based on a very simple and naive belief or intuition. This belief or assumption rarely can survive close scrutiny of the facts.

In summary of the foregoing ideas concerning the use of statistics in medical research it can be said that statistics should be simple, easily understood and based on common sense. First of all the investigators own feelings should be the basis for the design of the experiment and the evaluation of the results (Bayesian statistics). The simplest approach is order statistics, where it is not required that we make any assumptions about the character of the population that is being investigated. If it is desired or necessary that we have a more complicated model, then the Weibull distribution function is the most appropriate both from its simplicity of use and its broad character that encompasses most of the distribution function with which we are familiar.

REFERENCES

(1) Johnson, Leonard G., The Median Ranks of Sample Values in Their Population With Application to Certain Fatigue Studies, Industrial Mathematics, 2:1-9, 1951.

(2) Johnson, Leonard G., Calculating Rank Tables, Statistical Bulletin, of the Detroit Research Institute, 1:3, 1-4, 1971.

(3) Weibull, Waloddi, A Statistical Distribution Function of Wide Applicability, J. Applied Mech., 18:293-297, 1951.

(4) Dubes, Richard C., Data Reduction with Grouping and Weibull Models, Interim Report No. 7, Contract No. AFOSR-1023-67B, Div. Engineer. Res. Michigan State University, Jan. 30, 1970.

(5) Berguer, Ramon, J.D., & Smith, Roger F., M.D., Angiotension Infusion, Henry Ford Hospital Med. J. 16:2, 127-135, 1968.

(6) Edwards, Ward, Lindman, Harold, & Savage, Leonard J., Bayesian Statistical Inference for Psychological Research, Psychological Rev. 70:3, 193-242, 1963.

APPENDIX I

THE RANDOM SAMPLE

If there is some population of elements that has some characteristic X that we wish to investigate, and the population has some distribution function $F_0(x)$, we suppose that for k = 1, 2, ..., and for an arbitrary $x_1, x_2, \ldots, x_k$ there exists the conditional distribution function

$$F_k(x|x_1, x_2, \ldots x_k) = P(X'_{k+1}<x|X'_1 = x_1, \ldots X'_k = x_k)$$

If this supposition is true then we may choose some elements of a sample of the population by a random method if:

(1) for every x we have $P(X<x) = F_0(x)$

(2) for k = 1, 2, ..., n-1 and for arbitrary $x_1, x_2, \ldots, x_k$ we have the equalities - $P(X_{k+1}<x|X_1=x_1, \ldots, X_k=x_k) = F_k(x|x_1, \ldots, x_k)$

It must be noted that this equality is conditional:

$$F_k(x \text{ if } x_1, \ldots, x_k \text{ exists}) = P(X_{k+1}<x \text{ if } X_1=x_1, X_2=x_2, \ldots, X_k=x_k)$$

In classical statistical inference an assumption is made that indeed for X being a member of the population and x being a member of the sample population, $X_1=x_1, \ldots X_k=x_k$ or that $P(X_1=x_1, \ldots, X_k=x_k) \geq 0.99$ (almost certainty). In real life problems this is of necessity an intuitive or personal probability since it cannot be established by observation.

Further assumptions are required for most classical tests of statistical significance. One such assumption is that

$$X_1, X_2, \ldots X_k \rightarrow F_0(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\left[\frac{(x-\mu)^2}{2\sigma^2}\right] \quad \text{or}$$

that the characteristic X is normally distributed in the population. Again since this cannot be observed we assume that since $X_1 = x_1, \ldots X_k = x_k$

(intuition) then $x_1, \ldots x_k = F_0(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right]$ It has been observed that when $k<30$, the sample $x_1, \ldots x_k$ is not normally distributed. Therefore when small random samples are used as a basis for classical inference several intuitive assumptions must be accepted before beginning the testing of the samples.

It can be further observed that samples, regardless of how large, when drawn from biological populations are also not normally distributed. This should not be unexpected since many biological populations display infant mortality, random decay, or old age survival. In such cases the characteristic X of the population or a sample x drawn from such a population could not be normally distributed.

APPENDIX II

Parameters of the Distribution of a Random Variable

There are four kinds of parameters of a distribution of a random variable:

(1) moments

(2) functions of moments

(3) order parameters

(4) functions of order parameters

Moments

The moment of order k of the random variable X is:

$$m_k = E\ (X^k)$$

For the discrete distribution

$$E\ (X^k) = \sum_1 x_1{}^k p_1$$

For the continuous distribution

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

If $m_k = E[X-c]^k$, and $c=m_1 = E(x)$, then $\mu_k = E[X-E(X)]^k$ are central moments when $c = m_1 = E(X) = 0$

Then $\mu_1 = E[X-E(X)] = E[X-m_1] = E[X]-m_1$

$= m_1-m_1 = 0$

The central moment of the first order, μ_1, is called the mean of the distribution function of X (the population). It follows that

$$\mu_2 = E[X-E(X)]^2 = E[(X-m_1)^2]$$

$$= E[X]^2-2m_1E(X)+m_1{}^2$$

The central moment of the second order, μ_2 is called the variance of X and is denoted by σ^2.

The central moment of the third order, μ_3 is the third power of μ_1.

$$\mu_3 = E[X-E(X)]^3 = E[(X-m_1)^3]$$
$$= E(X^3)-3m_1E(X^2) + 3m_1^2E[X]-m_1^3$$
$$= m_3-3m_1m_2+3m_1^3-m_1^3$$
$$= m_3-3m_1m_2+2m_1^3$$

If the distribution of X is symmetrical then all odd moments are zero, but if this is not the case then a function of the third central moment, μ_3, is defined as $\alpha = \frac{\mu_3}{\sigma^3}$ and is called the coefficient of skewness.

Order Parameters

The value x satisfying the inequalities

$$P(X\leq x) \geq 1/2,\ P(X\geq x)\geq 1/2$$

is called the median of the distribution of the random variable X.

The value x satisfying the inequalities

$$P(X\leq x)\geq p,\ P(X\geq x)\geq 1-p,\ (0<p<1)$$

is called the quantile of order p.

The value of x satisfying the equality $P(X\leq x) = 0$ is called the point of origin of the distribution of X. This is called the parameter α of a Weibull distribution function.

The value x satisfying the equality $P(X=x) = 0.623$ is called the location parameter of a Weibull distribution.

APPENDIX III

Order Statistics

Let $X_1, X_2, \ldots X_n$ be an n dimensional random variable.

Let $x_1, x_2, \ldots x_n$ be a sample of values drawn from $X_1, X_2, \ldots X_n$.

Arrange the sample $x_1, x_2, \ldots x_n$ in such a way that $x_{r_1}, x_{r_2}, \ldots x_{r_n}$ satisfies the inequalities $x_{r_1} \leq x_{r_2} \leq \ldots \leq x_{r_n}$ Then $f\,[x(-\infty<x<\infty)] = 0$ for $x \leq x_{r_1}$ and $f\,[x(-\infty<x<\infty)] = \frac{m}{n}$ $(m=1,2, \ldots n)$ for $x > x_{r_1}$. By definition $x_{r_n} < x = S_n(x)$ and is called the "empirical distribution function" of x.

From these assumptions it follows that $P(X_r<x) = F(x) = p$ = constant $(r = 1, 2, \ldots n)$

Hence, for a fixed value of x, $S_n(x)$ is the frequency of successes in the Bernoulli scheme. Thus

$$P\left[S_n(x) = \frac{m}{n}\right] = \frac{n!}{m!\,(n-m)!}\ [F(x)]^m\ [1-F(x)]^{n-m}$$

Let the function of $(X_1, X_2, \ldots X_n)$ which takes the value x_{r_k} in each possible sequence $x_1, x_2, \ldots x_n$ be called an "order statistic" and be noted by $\zeta_k^{(n)}$. The number k is called the "rank" of $\zeta_k^{(n)}$.

Let $\phi_{kn}(x) = F(\zeta_k^{(n)}) = P(\zeta_k^{(n)}<x) = P\left[S_n(x) \geq \frac{k}{n}\right]$

$$= \sum_{m=k}^{n} P\left[S_n(x) = \frac{m}{n}\right]$$

then $\phi_{kn}(x) = \sum_{m=k}^{n} \frac{n!}{m!\,(n-m)!}\ [F(x)]^m [1-F(x)]^{n-m}$

assume $f(x) = F'(x)$ exists

then $f_{kn}(x)$ of $\zeta_k^{(n)}$ exists

therefore $f_{kn}(x) = \frac{n!}{(k-1)!\,(n-k)!}\ [F(x)]^{k-1} [1-F(x)]^{n-k} f(x)$

Median Ranks of Order Statistics

For the derivation of median ranks we are reprinting the paper by Leonard G. Johnson from Industrial Mathematics, vol.2, 1951.

The c-Rank of Order Statistics

A method for approximating the c-rank of any order statistic from any sample of size N has been presented by L.G. Johnson in the Statistical Bulletin, vol.1, bull. 3, July 1971.

It is presented here:

Let N = sample size

let j = order statistic

let c = confidence index (0<c<1)

let $Z_c(j/N)$ = c-rank of jth order statistic

let $Z_{.50}$ = median rank of jth order statistic in N

$$= \frac{j-.3}{N+.4} \text{ (Bernard's formula)}$$

$$\text{define } A_N(j) = 1 + \frac{.45N \cdot 6c(j-1)(N-j)}{(N-1)^2}$$

$$\text{define } \mu = \frac{\text{Log}\left(1-c\,\frac{1}{jA_N(j)}\right)}{\text{Log}\left(1-.5\,\frac{1}{jA_N(j)}\right)}$$

$$\text{then } Z_c(j/N) = 1 - [1 - Z_{.50}(j/N)]^{\mu}$$

DERIVATION OF THE MEDIAN RANKS

Deriving the Median Ranks

Assume the following situation is given:

A sample of n observations in numerical order: ${}_0X_1$, ${}_0X_2$, ${}_0X_3$, ..., ${}_0X_n$.

Probability density function of population: f(x) (unknown)

Cumulative distribution function of population: F(x) (unknown)

We define,

True rank of ${}_0X_1 = {}_nZ_1 = F({}_0X_1)$ (unknown)

True rank of ${}_0X_2 = {}_nZ_2 = F({}_0X_2)$ (unknown)

True rank of ${}_0X_3 = {}_nZ_3 = F({}_0X_3)$ (unknown)

. . .

. . .

. . .

In general, True rank of ${}_0X_j = {}_nZ_j = F({}_0X_j)$ (unknown)

Since the true rank of an observation is unknown, the best that we can do is to estimate what that true rank is. Consider next the set of all possible samples of size n from the same population. The table below is a partial list of these samples:

Sample 1: ${}_0X_1(1)$, ${}_0X_2(1)$, ${}_0X_3(1)$ ...., ${}_0X_n(1)$

Sample 2: ${}_0X_1(2)$, ${}_0X_2(2)$, ${}_0X_3(2)$ ...., ${}_0X_n(2)$

Sample 3: ${}_0X_1(3)$, ${}_0X_2(3)$, ${}_0X_3(3)$ ...., ${}_0X_n(3)$

. . . .

. . . .

Sample r: ${}_0X_1(r)$, ${}_0X_2(r)$, ${}_0X_3(r)$ ...., ${}_0X_n(r)$

The true ranks of the jth observations in these samples of size n may be listed as follows:

$$F({}_0X_j^{(1)}) = {}_nZ_j^{(1)}$$

$$F({}_0X_j^{(2)}) = {}_nZ_j^{(2)}$$

$$F({}_0X_j^{(3)}) = {}_nZ_j^{(3)}$$

$$\vdots \qquad \vdots$$

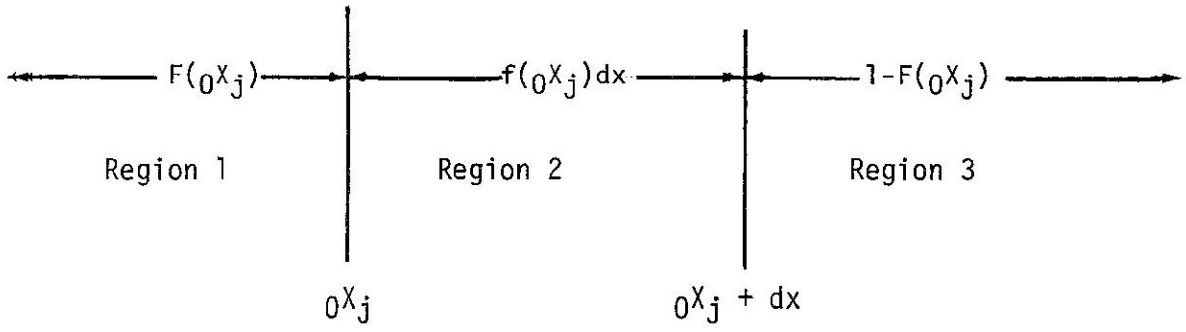
$$F({}_0X_j^{(r)}) = {}_nZ_j^{(r)}$$

We shall show that these ranks are distributed according to the probability density function

$$g({}_nZ_j) = \frac{n!}{(j-1)!\,(n-j)!}\,{}_nZ_j^{\,j-1}\,(1-{}_nZ_j)^{n-j}$$

This follows from the multinomial theorem, as demonstrated below:

Divide the entire population into the following three regions:



The expressions $F({}_0x_j)$, $f({}_0x_j)dx$, and $1 - F({}_0x_j)$ represent the probabilities that a single observation would fall into regions 1, 2, and 3, respectively. Therefore, it follows that the probability of exactly (j-1) values falling

into region 1, and exactly one value falling into region 2, and exactly (n-j) values falling into region 3 is given by the expression

$$\frac{n!}{(j-1)!\ 1!\ (n-j)!}\ [F({}_0X_j)]^{j-1}.f({}_0X_j)dx\ .\ [1 - F({}_oX_j)]^{n-j}$$

(This is an application of the multinomial theorem on joint probabilities.)

Now put $F({}_0X_j) = {}_nZ_j$. Then, $f({}_0X_j)dx = d_nZ_j$.

Hence, the above probability becomes

$$\frac{n!}{(j-1)!\ (n-j)!}\ {}_nZ_j{}^{j-1}\ (1 - {}_nZ_j)^{n-j}d_nZ_j$$

Therefore, the probability density function of ${}_nZ_j$ is

$$g({}_nZ_j) = \frac{n!}{(j-1)!\ (n-j)!}\ {}_nZ_j{}^{j-1}(1-{}_nZ_j)^{n-j} \qquad (A)$$

By definition, the mean value of ${}_nZ_j$ is

$$\int_0^1 {}_nZ_j.g({}_nZ_j)d_nZ_j \ = \ \int_0^1 {}_nZ_j.\ \frac{n!}{(j-1)!\ (n-j)!}\ {}_nZ_j{}^{j-1}(1-{}_nZ_j)^{n-j}d_nZ_j$$

This integral can be reduced to a Beta function, and its value turns out to be $j/(n+1)$. This is the mean rank of the jth observation in n. We could use this as an estimate of the true rank of the jth observation in n, but in certain applications the median value of ${}_nZ_j$ has been found to be more useful as an estimate. Let us now proceed to find the median value of ${}_nZ_j$, i.e., the median rank of the jth value in n.

By integration of equation (A), we find that the cumulative distribution function of ${}_nZ_j$ is

$$G({}_nZ_j) = 1 - (1-{}_nZ_j)^n - n{}_nZ_j(1-{}_nZ_j)^{n-1} - \frac{n^{(2)}}{2!}{}_nZ_j^2(1-{}_nZ_j)^{n-2} - \ldots - \frac{n^{(j-1)}}{(j-1)!}{}_nZ_j^{j-1}(1-{}_nZ_j)^{n-j+1} \qquad (B)$$

(Note: $n^{(2)}=n(n-1)$ $n^{(3)}=n(n-1)(n-2)$, etc., .. $n^{(j-1)}= n(n-1)(n-2)\ldots(n-j+2)$.)

The median value of ${}_nZ_j$ is found by putting $G({}_nZ_j) = 1/2$ in equation (B) and then solving for ${}_nZ_j$, i.e., by solving the following equation for ${}_nZ_j$:

$$1 - (1-{}_nZ_j)^n - n{}_nZ_j(1-{}_nZ_j)^{n-1} - \frac{n^{(2)}}{2!}{}_nZ_j^2(1-{}_nZ_j)^{n-2} - \ldots - \frac{n^{(j-1)}}{(j-1)!}{}_nZ_j^{j-1}(1-{}_nZ_j)^{n-j+1} = \frac{1}{2} \qquad (C)$$

Such an equation contains exactly one real root between 0 and 1, and Table 1 consists of such roots for n and j up to 20. The table is constructed by solving equation (C) exactly for j=1 and for j=n, for each value of n up to 20. The intermediate values in each column are then filled in by forming an arithmetic progression. When this is done, we find that in every case $.495 \leq G({}_nZ_j) \leq .505$, which is good enough for practical purposes.

For n>20 the following convenient formula may be used:

$\lambda_n(j) = \dfrac{j - (1-\ln 2) - (2\ln 2 - 1)\left(\frac{j-1}{n-1}\right)}{n}$, where $\lambda_n(j)$ denotes the median rank of the jth value in n.

APPENDIX IV

THE WEIBULL DISTRIBUTION

If $P(X \leq x) = F(x)$

then $F(x) = 1-e^{-\phi(x)}$

The probability of the occurrance of some event x from a set of events $x_1, x_2, \ldots x_n$ is defined by:

$$P_n = 1-e^{-n\phi(x)}$$

The function $\phi(x)$ must be specified with the necessary conditions that it be a positive, non-decreasing function, vanishing at some point $\geq$ zero.

$\phi(x) = -\left(\frac{x-\alpha}{\theta-\alpha}\right)^{\beta}$ satisfies these requirements.

Then $F(x) = 1-e^{-\left(\frac{x-\alpha}{\theta-\alpha}\right)^{\beta}}$ is a three parameter Weibull distribution function.

If alpha is assumed to be zero then $F(x) = 1-e^{-\left(\frac{x}{\theta}\right)^{\beta}}$ is a two parameter Weibull distribution function.

For the two parameter Weibull distribution we may show that the shape parameter beta is the slope of the linear function Y = BX+A when plotted on Weibull probability paper (ordinate-log log and abscissa-log) as follows:

$$F(x) = 1-e^{-\left(\frac{x}{\theta}\right)^{\beta}}$$

$$1-F(x) = e^{-\left(\frac{x}{\theta}\right)^{\beta}}$$

$$\frac{1}{1-F(x)} = e^{+\left(\frac{x}{\theta}\right)^{\beta}}$$

$$\ln \frac{1}{1-F(x)} = \left(\frac{x}{\theta}\right)^{\beta}$$

$$\ln \ln \frac{1}{1-F(x)} = \beta \ln x - \beta \ln \theta$$

$$\text{let } Y = \ln \ln \frac{1}{1-F(x)}$$

$$\text{let } X = \ln x$$

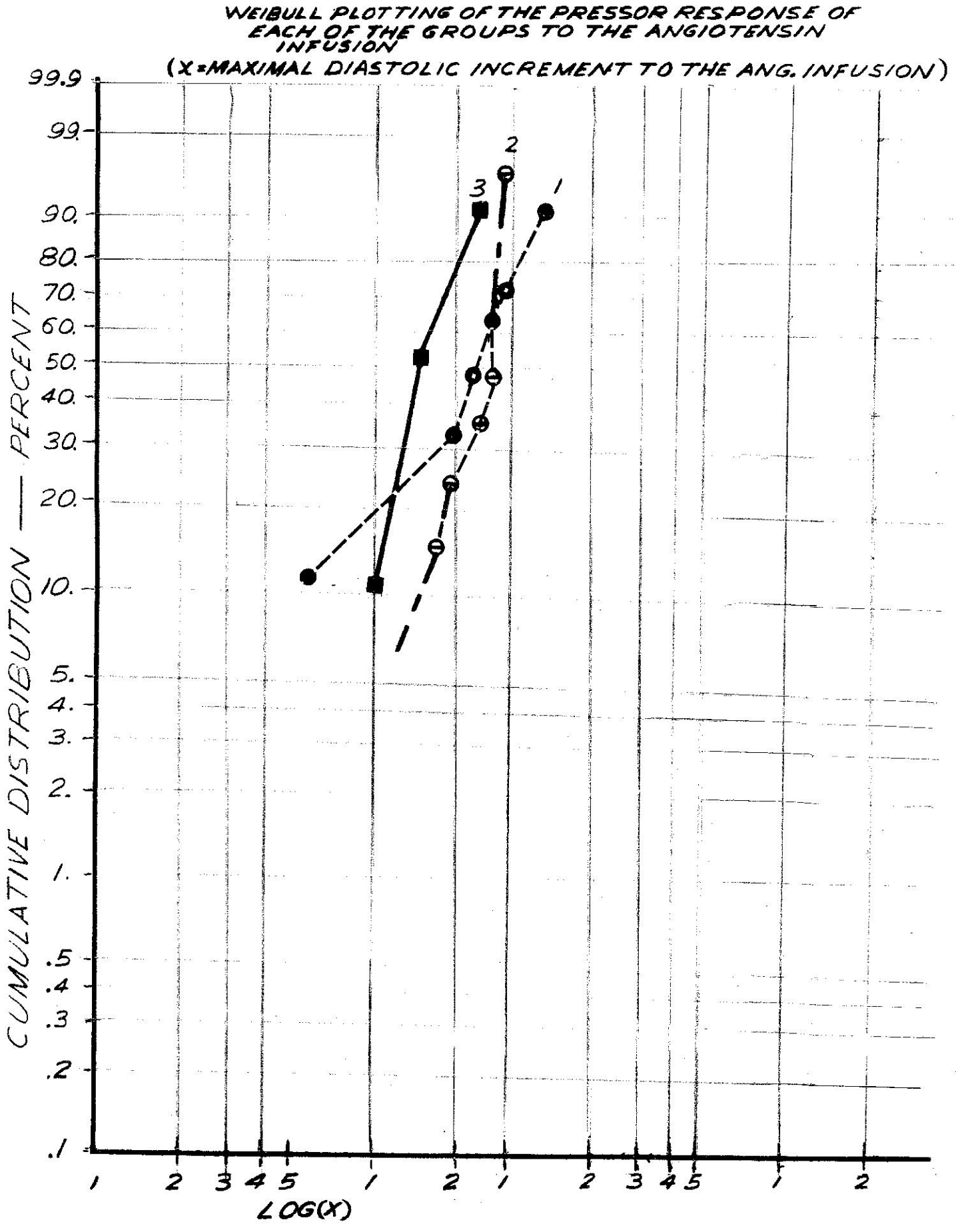
$$\text{let } A = -\beta \ln \theta$$

$$\text{let } B = \beta$$

$$\text{then } Y = BX + A$$

The probability density function of the two parameter Weibull Function is:

$$f(x) = \frac{\beta x^{\beta-1}}{\theta^{\beta}} \; e^{-\left(\frac{x}{\theta}\right)^{\beta}}$$


WEIBULL PLOTTING OF THE PRESSOR RESPONSE OF EACH OF THE GROUPS TO THE ANGIOTENSIN INFUSION
(X = MAXIMAL DIASTOLIC INCREMENT TO THE ANG. INFUSION)
CUMULATIVE DISTRIBUTION — PERCENT
99.9
99.
90.
80.
70.
60.
50.
40.
30.
20.
10.
5.
4.
3.
2.
1.
.5
.4
.3
.2
.1
1
2
3
4
5
1
2
1
2
3
4
5
1
2
LOG(X)