Department of Statistics

UNIVERSITY OF WISCONSIN

Madison, Wisconsin

Technical Report Number 7

December, 1961

A FURTHER LOOK AT ROBUSTNESS VIA BAYES' THEOREM[1]

G. E. P. Box and G. C. Tiao*

This research was supported by the United States Navy through the Office of Naval Research, under Contract Nonr-1202(17), Project NR 042 222. Reproduction in whole or in part is permitted for any purpose of the United States Government

* During the period covered by this work Mr. Tiao, of the Social Systems Research Institute, University of Wisconsin, was supported by a grant from the Ford Foundation.

---

[1] Also issued as a System Formulation and Methodology Workshop paper No. 61-04 by the Social Systems Research Institute, University of Wisconsin

A FURTHER LOOK AT ROBUSTNESS VIA BAYES' THEOREM

G. E. P. Box and G. C. Tiao

1. Introduction and summary

In recent years, under the leadership of Savage [13, 14, 15], there has been a great revival of interest in subjective probability and in the interpretation of data via Bayes' theorem. Many statisticians now feel that this provides the most satisfactory basis for a theory of statistical inference. In particular, such an approach seems necessary if one is to give explicit cognisance to the uncertainty in the assumptions which are built into many statistical procedures. Classical statistical arguments lead us to treat such assumptions as if they were in some way axiomatic and yet consideration will show that, in fact, they are conjectures which in practice may be expected to be more or less true. The mathematical expression of more or less true seems to require the explicit injection of subjective probability distributions. For instance, in many problems the particular physical setup is such that the errors involved might behave like a linear aggregate of component errors and, consequently, a central limit effect would operate. In fact, of course, the central limit theorem does not tell us that a linear aggregate of a finite number of component errors would be exactly normal. We are, however, entitled to expect in this physical situation that the distribution we are dealing with will be a member of a "distribution of distributions" in which the normal curve occupies the central place. In this situation we can express our true state of mind by the use of a prior distribution of some parameter or parameters, measuring

the non-normality of the parent distribution.

If we assume normality, we can proceed with an "objective" classical analysis. But by making this normality assumption, however, we act in fact as if the distribution of our non-normality parameter were a delta function. As seems to be inevitably the case in other problems as well as this one, therefore, our "objectivity" is gained by pretending to knowledge we do not have and in so doing we even ignore what the sample has to tell us about the matter in question.

On classical theory, once having assumed the form of the parent distribution, we can derive criterion which is appropriate on this assumption. For example, on the assumption of normality, for the comparison of two means we would derive the t statistic. It is then customary to justify the use of such a normal theory criterion in the practical circumstance in which normality cannot be guaranteed by arguing that the distribution of the criterion is but little affected by non-normality of the parent distribution--that is, it is robust under non-normality. However, this argument ignores the fact that if the parent distribution really differed from the normal, the appropriate criterion would no longer be the normal-theory statistic. It is easy to produce examples in which the distribution of the normal theory criterion is little affected if the parent is assumed to be some distribution other than the normal; and yet, the inference to be drawn when a criterion appropriate to this other distribution is employed is markedly different.

In this paper the anlaysis of Darwin's paired data on the heights of self and cross-fertilized plants quoted by Fisher in "The Design of

Experiments"[6] is reconsidered. In our development the parent distribution is not assumed to be normal, but only a member of a class of symmetric distributions which include the normal, and whose kurtosis is measured by a parameter β. In this example, the physical nature of the experimental environment is certainly such that a central limit effect would be expected. That this expectation justifies us only in supposing that the error distributions will approach the normal is specifically recognized in our formulation by giving a subjective prior probability distribution to β centered at its normal value. The sharpness of this subjective prior distribution can be varied so that we can represent a range of situations in which a greater and greater degree of central limit effect is injected. Finally, when the prior distribution becomes a delta function, we produce the usual formulation in which an exact assumption of normality is made. At the other extreme, we can produce the situation in which all the information about normality or the lack of it is essentially being generated from the sample itself. The extent to which the usual normal theory t test could be approximately justified over this wide range of circumstances is illustrated and discussed.

It is believed that the injection into the model of subjective prior probability distributions to represent tentatively held "assumptions" has general application. Extension of these ideas to other statistical procedures is being carried out.

2. Various approaches to the analysis of Darwin's data

Darwin's experiments were conducted in pairs so that on the normal assumption and on classical theory one may interpret these data by using the paired t test. Figure 1(i) shows the observed differences. On the same scale is shown a t distribution centered about the sample mean $\bar{y}$ with scale factor $s/\sqrt{n}$ where the quantity $s^2 = \Sigma (y_i - \bar{y})^2/n-1$ is the usual sample estimate of the variance of the differences. In what follows for definiteness, we shall call this distribution the reference distribution for the population mean θ. This reference distribution may be variously interpreted. It is the fiducial distribution of θ. It can also be regarded as showing a complete set of confidence intervals for θ for all values of the "confidence coefficient." Finally, if we make suitable assumptions discussed later, concerning the prior distributions of θ and σ, it is the posterior distribution of θ. If we are interested in a significance test appropriate to the hypothesis that $\theta = 0$ against the alternate $\theta > 0$, then the associated significance level for the present example is 2.485%.

Now suppose that instead of assuming normality for the parent distribution, we assumed it to be uniform over some range $\theta - \sigma$ to $\theta + \sigma$, where here and in what follows, σ is used as a general scale parameter and does not necessarily refer to the standard deviation. This assumption would, of course, be quite ridiculous in the present example. First, we know that the many contributing errors arising from genetic differences, soil differences, and so forth, will produce a strong central limit effect so that we may expect with good reason that the heights themselves and, even

more, their differences will be closely normally distributed. Second, the evidence from the sample itself does not support the uniform assumption. However, to illustrate our point, let us make the assumption of a uniform instead of a normal parent. One thing we might then consider is the effect of this extreme degree of non-normality on the distribution of the t statistic. This can be approximately calculated using, for example, the work of Gayen [7] or of Box and Anderson [2]. Following these latter authors, it is readily shown that the null distribution of t^2 is approximated by an F distribution with δ and $\delta(n-1)$ degrees of freedom where

$$\delta = 1 + E(b-3)/n$$

and

$$E(b-3) = \gamma_2' - n^{-1}(2\gamma_4' - 3\gamma_2'^2 + 11\gamma_2') + n^{-2}(3\gamma_6' - 16\gamma_4'\gamma_2' + 15\gamma_2'^3 + 38\gamma_4' + 86\gamma_2')$$

where

$$\gamma_{r-2}' = \kappa_r(y) / \{\kappa_2(y)\}^{\frac{1}{2}r}$$

are the standardized cumulants of the parent distribution of differences. In our present example, δ is found to be 0.913. Thus, t^2 is approximately distributed as F with 0.913 and 12.78 degrees of freedom. In particular, the significance level associated with the hypothesis that $\theta = 0$ against the alternative $\theta > 0$ is now 2.388% as compared with the previous value of 2.485%. The test of the hypothesis that the true difference is zero using the t criterion is thus very little affected by this major departure

from normality. Similarly, confidence intervals based on the t-statistic would be very little affected by this departure. The robustness of the t statistic in this example was, in fact, demonstrated by Fisher who derived the exact randomization distribution in this case and showed that the null probability agreed very closely with that obtained from the t criterion.

However, if we really knew that the parent distribution was uniform, we would not consider the t criterion at all. We would be led instead to consider the function

$$W = | m - \theta | / h$$

where

$$m = \tfrac{1}{2}(y_L + y_S) \qquad h = \tfrac{1}{2}(y_L - y_S)$$

and y_L and y_S are respectively the largest and the smallest of the observations - jointly sufficient statistics for θ and σ on the uniform assumption. On this same assumption, as shown by Carlton [3], the variate $(n-1)W$ is distributed as F with 2, $2(n-1)$ degrees of freedom.

Just as Figure 1(i) exemplifies the inferential situation with the normal assumption, so Figure 1(ii) correspondingly exemplifies this situation with the uniform assumption. As before, we can interpret the distribution in Figure 1(ii) either as the fiducial distribution of θ or as defining a complete set of confidence intervals for θ or, finally, (if we adopt identical assumptions about the prior distributions of θ and σ as those needed before) as the posterior distribution of θ. We notice that this reference distribution is markedly different from that we obtained from the

normal assumption, especailly with regard to its location. In particular, the significance level associated with the hypothesis that $\theta = 0$ against the alternative $\theta > 0$ is not 2.485%, but 23.215%. Thus, whichever form of derivation we favor, we see that the conclusions which we would draw if we assumed a uniform parent distribution are very different from those which would be appropriate if we assumed a normal parent distribution, even though the t criterion itself is very little affected by this large departure from normality. The principal reason for this large difference is that in one case the reference distribution is centered at the sample mean and in the other case it is centered at the sample mid-point. For this particular sample, the mean and the mid-point differ considerably, mainly because of two rather large negative differences.

As we have explained, we are not seriously suggesting that the uniform distribution is a reasonable choice for the parent. We wish only to emphasize that uncertainty in our knowledge of the parent distribution transmits itself rather forcefully into an uncertainty about the conclusion we can draw concerning θ, and the difficulty which this presents in our interpretation of the data is not avoided by our knowledge of the robustness of the t criterion. It seems to us that this difficulty can only be resolved by explicitly including the knowledge that we have about the parent distribution into our formulation. This knowledge is of two kinds, that coming from the sample itself and that coming from knowledge of the physical set-up appropriate to this problem. In the classical formulation of the problem the first kind of information is ignored and the second kind

is misrepresented. We shall see that they are both taken account in an appropriate Bayesian formulation.

At this point, it may be instructive to remind ourselves of the Bayesian justification of the t distribution such as has been essentially given by Jeffreys [9,10].

3. Derivation of the t test via Bayes' theorem on the principle of precise measurement

The Bayesian argument requires that we have some prior distributions for θ and σ. We assume, as it seems reasonable, that the local prior distribution of these location and scale parameters are independent. So far as the location parameter is concerned, the situation met in actual circumstances of experimentation would often permit us to assume that the prior distribution of θ was locally uniform, using what Savage calls the principle of precise measurement [14,15]. This principle says, in effect, that we do not need to know exactly what the prior distribution of θ is if we can say only that in the region in which the likelihood is appreciable it does not change very much, and at no other point is it of sufficiently great magnitude as to become appreciable when multiplied by the likelihood. This principle would be applicable in situations like that illustrated in Figure 2(i) in which the likelihood dominates and is inapplicable in the situation illustrated in Figure 2(ii) in which the prior probability density dominates. What makes this principle of particular importance is that most actual experimental situations are

represented by Figure 2(i) rather than by Figure 2(ii). The reason for this is that if the situation is really like that in Figure 2(ii), then there is little point in doing the experiment. For instance, suppose that the value of the gravitational constant in suitable units had been estimated as $32.2 \pm .1$, then there would be little justification for making further measurements with a method whose accuracy was, say, $\pm .2$, but considerable justification for conducting further experiments using a method whose accuracy was $\pm .02$.

The argument that if θ is taken as locally uniform, then $\log \theta$, $\frac{1}{\theta}$, etc. will not be, loses its force if we remember that unless the range of value of θ over which the likelihood is appreciable is large compared with the average magnitude of θ over the same range, then such transformations will make little practical difference in the range considered. In the example considered above, for instance, if the prior distribution of θ were assumed uniform from, say, $\theta = 32.0$ to $\theta = 32.4$, then, to a close approximation, the prior distributions of, for example, $\log \theta$ and $\frac{1}{\theta}$ would be uniform over corresponding ranges.

We can also demonstrate this lack of sensitivity to prior assumption when we consider the scale parameter. Suppose we merely assume that either σ or its logarithm or some power of σ is locally uniform. We have then

$$p(\theta) \propto k, \quad p(\sigma) \propto \begin{cases} \sigma^{q-1} & \text{if } \sigma^q \text{ assumed uniform} \\ \sigma^{-1} & \text{if } \log \sigma \text{ assumed uniform} \end{cases} \tag{1}$$

whence denoting $\ell(\theta, \sigma/\underline{y})$ for the likelihood function given the sample $\underline{y}$

$$p(\theta, \sigma/\underline{y}) = k\, \ell(\theta, \sigma/\underline{y}) \cdot p(\theta) \cdot p(\sigma) \tag{2}$$

where $$k^{-1} = \iint_R \ell(\theta, \sigma/\underline{y}) \cdot p(\theta) \cdot p(\sigma)\, d\theta\, d\sigma \ .$$

On the normal assumption, then

$$p(\theta, \sigma/\underline{y}) = p(\theta/\sigma, \overline{y}) \cdot p(\sigma/s) \tag{3}$$

where

$$p(\theta/\overline{y}, \sigma) = (n/2\pi\sigma^2)^{\frac{1}{2}} \exp\{-(n/2\sigma^2)(\overline{y}-\theta)^2\}$$

$$p(\sigma/s) = \{\Gamma(\tfrac{\nu-q}{2})\}^{-1} (\tfrac{\nu s^2}{2})^{\frac{\nu-q}{2}}\ \sigma^{q-(\nu+1)} \exp\{-\nu s^2/2\sigma^2\}$$

$\nu = n-1$, and $q < \nu$.

On integrating out σ we obtain

$$p(\frac{\theta-y}{s/\sqrt{n}} / \underline{y}) = p[t_{\nu-q}] \tag{4}$$

where $p[t_{\nu-q}]$ is the t distribution with $\nu-q$ degrees of freedom. The only effect of changing the power q of σ supposed uniform, is to change the number of degrees of freedom in the final posterior t distribution. In particular, by assuming log σ to be locally uniform, we obtain a posterior distribution of θ as a t distribution with the traditional $\nu = n-1$ degrees of freedom. If we suppose σ to be locally uniformly distributed, we will have a t distribution with n-2 degrees of freedom, and if we suppose σ^2 to be locally uniform, a t distribution with n-3 degrees of freedom.

If we take log σ as the function of σ to be regarded as locally uniform we are consistent in the sense that log σ is a location parameter for log $(y-\theta)$, just as θ is a location parameter for y. We do not regard this argument as conclusive, but it is comforting to notice that from a moderate sized sample such as that from Darwin's data, rather drastic changes in the nature of the prior distribution of σ do not greatly affect the final conclusion and in what follows we make the assumption of uniform distribution for log σ.

For our later purposes, it is perhaps worth while to consider this well-known result geometrically. The joint posterior distribution of θ and σ

shown by the contour diagram in Figure 3, can be regarded as being built up from a series of normal distributions each centered at y, each of which represents the conditional distribution $p(\theta/\sigma, \bar{y})$ of θ for some given σ, multiplied by $p(\sigma/s)$, the marginal posterior distribution of σ which has the form of an inverted gamma function. If we knew the value of σ, then the posterior distribution of θ would be normal about $\bar{y}$ with this known value of $\sigma = \sigma_0$. When σ is unknown, then we must average all these normal distributions, using for weights the ordinates of the marginal posterior distribution of σ. In so doing, we obtain the t distribution. There is, of course, nothing new in the above; we recall it here only to introduce the more general argument which follows.

4. A wider choice of the parent distribution

If, in the analysis of Darwin's data, we suppose that the parent distribution of self and cross-fertilized plants are of the same form, then the distribution of the differences would certainly be symmetric. Let us, therefore, assume that our parent distribution is a member of a class of symmetric distributions which includes, in particular, the normal, together with other distributions on the one hand more leptokurtic and on the other hand more platykurtic than the normal. A convenient choice is the class of power distributions employed in other contexts, for example, by Diananda [5], Box [1] and Turner [16],

$$p(y) = \omega \exp\left\{-\tfrac{1}{2}\left|\frac{y-\theta}{\sigma}\right|^{2/(1+\beta)}\right\} \qquad (5)$$

where

$$\omega = \left\{\Gamma[1+(1+\beta)/2]\, 2^{[1+(1+\beta)/2]}\, \sigma\right\}^{-1}$$

$$-\infty < y < \infty \qquad 0 < \sigma < \infty$$

$$-\infty < \theta < \infty \qquad -1 < \beta < 1$$

In particular, we see that when $\beta = 0$, we have the normal distribution, when β is 1, we have the double exponential; and when β tends to -1, our distribution tends to the uniform distribution.

5. Derivation of the posterior distribution of θ for a specific symmetric parent

We now derive the posterior distribution of θ, supposing that the parent distribution to be a member of the above class of distributions in which β is assumed to have a fixed value β_0. In so doing, we shall adopt the same assumptions a priori as are necessary to derive the traditional t distribution when β is assumed to be zero.

We have

$$\ell(\theta,\sigma/\underline{y},\beta_0) = \left[\Gamma\left(1+\frac{1+\beta_0}{2}\right) 2^{\left(1+\frac{1+\beta_0}{2}\right)} \sigma\right]^{-n} \exp\left\{-\tfrac{1}{2}\left|\frac{y-\theta}{\sigma}\right|^{2/(1+\beta_0)}\right\}$$

$$p(\theta) \propto k, \quad p(\sigma) \propto \sigma^{-1} \tag{6}$$

so that

$$p(\theta, \sigma/\underline{y},\beta_0) = k\,\sigma^{-(n+1)} \exp\left\{-\tfrac{1}{2}\sum_1 \left|\frac{y_i-\theta}{\sigma}\right|^{2/(1+\beta_0)}\right\}^{*} \tag{7}$$

where

$$k^{-1} = \iint_R \sigma^{-(n+1)} \exp\left\{-\tfrac{1}{2}\sum_1 \left|\frac{y_i-\theta}{\sigma}\right|^{2/(1+\beta_0)}\right\} d\theta\, d\sigma :$$

By integrating our σ as before, we finally obtain for the posterior distribution of θ for any fixed $\beta = \beta_0$ in the permissible range the remarkably simple expression

$$p(\theta/\underline{y}, \beta_0) = k[M(\theta)]^{-\frac{n(\beta_0+1)}{2}} \tag{8}$$

where

$$M(\theta) = \left[\sum_1 |y_i-\theta|^{2/(1+\beta_0)}\right]$$

is the absolute moment of order $\frac{2}{1+\beta_0}$ of the observations about θ.

The integral

$$k^{-1} = \int_{-\infty}^{\infty} [M(\theta)]^{-n(1+\beta_0)/2}\, d\theta$$

is merely a normalizing factor which ensures that the total area under the

---

* It is understood here that at least two of the observations are not equal.

distribution is unity. This integral can not usually be expressed as a simple function; it can, of course, always be computed numerically, and with the availability of electronic computers, this presents no particular difficulty.*

Using equation (8), posterior distributions computed from Darwin's data for various values of β_0 are shown in Figure 4.

6. Properties of the posterior distribution of θ for fixed $\beta = \beta_0$

Since $p(\theta/\underline{y},\beta_0)$ is a monotonic function of $M(\theta)$, we find (see Appendix) the following:

(1) $p(\theta/\underline{y},\beta_0)$ is continuous, differentiable and unimodal, although not necessarily symmetric, the mode being attained in the interval $[y_s, y_L]$.

(2) When $\beta_0=0$, $M(\theta) = \Sigma(y_i-\theta)^2 = (n-1)s^2 + n(\bar{y}-\theta)^2$ and making necessary substitutions we obtain for the posterior distribution of θ

$p\left(\frac{\theta-\bar{y}}{s/\sqrt{n}} / \underline{y}, \beta_0\right) = p(t_{n-1})$ as before.

(3) When β approaches -1, $\lim_{\beta_0 \to -1} [M(\theta)]^{\frac{n(\beta_0+1)}{2}} = (h+|m-\theta|)$

and making the necessary substitutions,

$$\lim_{\beta_0 \to -1} p(\theta/\underline{y},\beta_0) = k[h+|m-\theta|]^{-n} \qquad (9)$$

where $$k^{-1} = \int_{-\infty}^{\infty} (h+|m-\theta|)^{-n}\, d\theta$$

---

* It is interesting to notice here that when $\beta \leq 0$ and the quantity $n\left(\frac{\beta+1}{2}\right)$ is large, we can approximate the posterior distribution by

$$p(\theta/\underline{y},\beta) \sim k \exp\{-\tfrac{1}{2}[\tfrac{n(1+\beta)}{2}]h''(\theta_0)(\theta-\theta_0)^2\}$$

where

$$k = \{\tfrac{n(1+\beta)}{2} h''(\theta_0)/2\pi\}^{\frac{1}{2}} \qquad h(\theta) = \log[M(\theta)]$$

and θ_0 is the value of θ at which $h(\theta)$ attains its minimum. This is a special case of a powerful method, known as "Saddle Point Approximation," developed by Jeffreys, [11] and Daniels [4]. In our case, it is equivalent to the normal approximation of the distribution $p(\theta/\underline{y},\beta)$ around the maximum likelihood estimate of θ.

so that

$$\lim_{\beta_0 \to -1} p\left(\frac{|\theta - m|}{h/(n-1)} \Big/ \underline{y} \ , \ \beta_0 \right) = p[F_{2,\,2(n-1)}]$$

This is then the reference distribution shown in Figure 1(ii) but now derived as a posterior distribution.

Thus, we see that, when the parent is normal $(\beta_0 = 0)$, our expression (8) yields the t distribution as expected, and when the parent approaches the uniform $(\beta_0 \to -1)$, again as expected, our expression (8) gives the double F distribution with 2 and $2(n - 1)$ degrees of freedom. In each of these cases, the posterior distribution can be expressed in terms of simple functions of the observations which provide then, of course, minimal sufficient statistics for θ and σ .

(4) When β_0 approaches 1, the distribution is not expressible in simple function of the observations, but in the limit the mode of the posterior distribution is the median of the observations if the latter is uniquely defined; and, if not, it is some unique value between the values of the middle two observations.

(5) In certain other cases, it is possible to express the posterior distribution of θ in terms of a fixed number of functions of the observations. For instance, when

$$\beta = (1-q)/q \ , \qquad q = 1 \ , \ 2 \ , \ 3 \ , \ \ldots$$

we have

$$p(\theta \ , \sigma / \underline{y} \ , \ \beta_0) \propto \sigma^{-(n+1)} \exp\left\{ -\tfrac{1}{2}\sigma^{-2q} \sum_{r=0}^{2q} (-1)^r \binom{2q}{r} \theta^r \, S_{2q-r} \right. \tag{10}$$

and

$$p(\theta / \underline{y}, \ \beta_0) \propto \left[\sum_{r=0}^{2q} (-1)^r \binom{2q}{r} \theta^r \, S_{2q-r} \right]^{-\frac{n}{2q}} \tag{11}$$

where

$$S_r = \sum_1 y_i^r$$

and it is readily seen that the set of $2q$ functions, S_1 , S_2 , ... S_{2q} of the observations are jointly sufficient for θ and σ.

In general, however, the posterior distribution can not be expressed in terms of a few simple functions of the observations, the minimal sufficient statistics are the observations themselves. If we wish to think in terms of sufficiency and information as defined by Fisher, our posterior distribution always of course employs a complete set of sufficient statistics, and, consequently, no matter what is the value of β, no information is lost. The posterior distribution of θ always has as its central value, the maximum likelihood estimate of θ, but it should be noticed that we are not concerned with the distribution of this maximum likelihood estimate; rather, we are considering the distribution of θ given each one of the observations.

From the family of distributions for various values of β_0 as shown in Figure 4, we see that very different inference will be drawn concerning θ, depending upon which value of β_0 is assumed. The chief reason for this wide discrepancy is the fact that in Darwin's data, the center of the posterior distribution changes markedly as β is changed. In particular, for this sample, the median, mean and the mid-point are respectively 24.0, 20.9, 4.0, and these are the modes of the posterior distributions for the double exponential, normal and uniform parent respectively.

7. Posterior distribution of θ and β when β is regarded as a variable parameter

Because of the wide differences which occur in the posterior distribution of θ depending on which parent distribution (that is, which value of β_0) we employ, in practice it might be thought there would be considerable uncertainty as to the nature of the valid inference that could be drawn from this data. We now show in this section that this is not the case, when we use appropriate evidence concerning the value of β. We have, in fact, two sources of information about the value of β, one from the data itself and the other from our knowledge a priori that a central limit effect would operate in the circumstances of the experiment. Both types of evidence can be injected into our analysis by allowing β itself to be a variable parameter associated with a prior distribution.

We can represent the central limit tendency of the errors by choosing a prior distribution for β which has a maximum value at $\beta = 0$, and which extends from -1 to $+1$. A convenient distribution for this purpose is the beta distribution having mean zero and extending from -1 to $+1$ and, consequently, possessing only one adjustable parameter which we call a. We assume then:

$$p(\beta) = w\,(1 - \beta^2)^{a-1} \qquad -1 < \beta < 1$$

where

$$w = \Gamma(2a)\,[\Gamma(a)]^{-2}\,2^{-(2a-1)} \qquad a \geq 1 \tag{12}$$

When $a = 1$, this distribution is uniform. With $a > 1$, it is a symmetric distribution having its mode at the normal theory value $\beta = 0$. If we wished to represent a situation in which some value other than $\beta = 0$ occupied the central position, then this could be done in a similar way by using a

beta function having two adjustable parameters.

After eliminating the scale parameter σ, we now obtain for the joint posterior distribution of θ and β:

$$p(\theta, \beta/\underline{y}) = k_1(1-\beta^2)^{a-1} \Gamma\left[1+\frac{n(1+\beta)}{2}\right] \left[\Gamma\left(1+\frac{1+\beta}{2}\right) \right]^{-n} [M(\theta)]^{-\frac{n(1+\beta)}{2}} \quad (13)$$

$$= k_2 \cdot f(\theta, \beta/\underline{y})\, p(\beta)$$

where $p(\beta)$ is given by equation (12), $f(\theta, \beta/\underline{y})$ is the function represented by the surface shown in Figure 5 and k_1 and k_2 are the appropriate normalizing constants. We can write

$$f(\theta, \beta/\underline{y}) = p(\theta/\beta, \underline{y})\, \phi(\beta) \quad (14)$$

where

$$\phi(\beta) = k_3 \Gamma\left[1 + \frac{n(1+\beta)}{2}\right] \left[\Gamma\left(1 + \frac{1+\beta}{2}\right) \right]^{-n} \int_{\infty}^{\infty} [M(\theta)]^{-\frac{n(1+\beta)}{2}} d\theta \quad (15)$$

and $p(\theta/\beta, \underline{y})$ is given by equation (8). The conditional distributions $p(\theta/\beta, \underline{y})$ are the t-like distributions which we have already plotted in Figure 4 and which represent the posterior distributions of θ for different specific choices of β. The function $\phi(\beta)$ which is sketched in Figure 6(i) can be regarded as representing <u>information coming from the sample concerning</u> β. We see that the function $f(\theta, \beta/\underline{y})$ is in fact built up of these t-like distributions suitably weighed with the weight function $\phi(\beta)$.

We can interpret $f(\theta, \beta/\underline{y})$ as the joint posterior distribution of θ and for which the prior distribution of β is uniform. In this case, $\phi(\beta)$ is then the posterior distribution of β. Of course, it would usually be quite unrealistic to suppose that the distribution of β were uniform a priori. Since

$$p(\theta, \beta/\underline{y}) = k\, p(\theta/\beta, \underline{y})\ \phi(\beta)\ p(\beta)$$

we see that the joint posterior distribution $p(\theta, \beta/\underline{y})$ will in general be obtained by weighing $p(\theta/\beta, \underline{y})$ not with $\phi(\beta)$ but with the function $\phi(\beta) \cdot p(\beta)$. The parameter a in $p(\beta)$ [equation(12)] can be adjusted to allow for any desired strength of central limit effect. The case $a = 1$ giving a uniform distribution for $p(\beta)$ corresponds to no central limit effect. When a tends to infinity, $p(\beta)$ becomes a delta function and represents an overwhelmingly strong central effect. This corresponds to the assumption of exact normality for the parent distribution.

8. Posterior distribution of θ

From the joint posterior distribution of θ and β

$$p(\theta, \beta/\underline{y}) = p(\theta/\beta, \underline{y}) \quad \phi(\beta) \quad p(\beta)$$

the posterior distribution of θ is obtained by integrating out β yielding

$$\begin{aligned} p(\theta/\underline{y}) &= \int_{-1}^{1} p(\theta, \beta/\underline{y}) \, d\beta \\ &= \int_{-1}^{1} p(\theta/\beta, \underline{y}) \quad \phi(\beta) \quad p(\beta) \, d\beta \end{aligned} \tag{16}$$

In obtaining this integral, we are averaging the t-like distributions $p(\theta/\beta, \underline{y})$ with a weight function $\phi(\beta) \cdot p(\beta)$ which is in fact $p(\beta/\underline{y})$, the posterior distribution of β. The value of this weight function is seen to depend partly upon information from the sample through $\phi(\beta)$ and partly from prior information characterized by $p(\beta)$. The way in which this weight function $\phi(\beta) \cdot p(\beta)$ changes as the assumed central limit effect is increased is shown in Figures 6(i) - 6(iv). In these diagrams, the dotted curve is, in each case, the prior distribution $p(\beta)$. When $a = 1$ (Fig. 6(i),

$p(\beta)$ is uniform and $p(\beta/\underline{y})$ equals $\phi(\beta)$. This represents the situation where the information concerning β is essentially coming from the sample itself. The value of the parameter a is 3 in Fig. 6(ii), 6 in Fig. 6(iii) and 10 in Fig. 6(iv). These three diagrams show how increasing certainty of a central limit effect tends to override the information from the sample. Finally, when a tends to infinity, both $p(\beta)$ and $p(\beta/\underline{y})$ would approach a delta function at $\beta = 0$.

The integration

$$p(\theta/\underline{y}) = \int_{-1}^{1} p(\theta/\beta, \underline{y}) \ \phi(\beta) \ p(\beta) \ d\beta$$

has been actually carried out for each of these weight functions and the results are shown in Figure 7 together with the t distribution which would be appropriate for the case $a \to \infty$ corresponding to an assumption of exact normality. In the diagram, the case $a = 10$ is not shown since this curve is almost indistinguishable from the t distribution. This is interesting because as will be seen from Fig. 6(iv), the central limit effect implied by $a = 10$ is not an overwhelmingly strong one. For instance, it certainly leaves as acceptable a priori the possibility that $\beta = +.3$ or $\beta = -.3$. If we call the normal distribution a "second power" distribution, this is equivalent to supposing a priori that "third power" distributions and "$1\frac{1}{2}$ power" distributions are possible.

Figure 7 then represents the final inference we could draw for θ depending on how strong a central limit effect would be appropriate in the physical situation. In view of the very large differences exhibited by the

t-like distributions in Figure 4, it seems remarkable how alike these distributions are. In particular, it will be seen that the tail areas which have been traditionally regarded as the most important part of the distribution are very little affected even with no "central limit effect." The main reason for this is that those widely discrepant t-like distributions generated by parents which approach the uniform are almost ruled out by information coming from the sample itself. (See Fig. 6(i).)

We may remark here that the precise form of the posterior distribution of θ would of course depend to some extent upon the way we parameterize the constant measuring normality. The measure β which makes the double exponential and the rectangular distributions equally discrepant from the normal seems not unreasonable. However, we might have used for example the familiar kurtosis measure $\lambda_4 = \kappa_4/\kappa_2^2$ for the class of distributions we have considered. It is easily shown, in fact, that

$\lambda_4 = \dfrac{\Gamma[5(1+\beta)/2]\,\Gamma[(1+\beta)/2]}{\Gamma[3(1+\beta)/2]^2} - 3$. On this scale, the double exponential distribution would appear as 3 and the rectangular distribution as -1.2. It seems that whether β, λ_4 or any other reasonable measure of non-normality is adopted, the overall conclusions are very similar.

9. <u>Information concerning the nature of the parent distribution coming from the sample</u>

In the past the normality or otherwise of a sample has usually been decided either by employing certain preliminary tests, for example, the χ^2 goodness of fit test, and the Kolmogoroff-Smirnoff test [12] or by calculating statistics of skewness or kurtosis such as $\lambda_3 = \kappa_3 / (\kappa_2)^{\frac{3}{2}}$,

$\lambda_4 = \kappa_4 / k_2^2$. In the present instance, since we are dealing with differences in heights, it is reasonable to assume the distribution is symmetric. It would then seem that the calculation of $\phi(\beta)$, i.e. $p(\beta/\underline{y})$ for uniform $p(\beta)$ as shown in Figure 6(i) would provide a much more satisfactory way of summarizing what the data has to tell us concerning the nature of the parent distribution from which the sample is drawn. It will be noted that in our approach, we have done more than merely "test" the assumption of normality and then, in the absence of "a significant" result, assume it. The information concerning β coming from the sample is included in the formulation itself and as we have seen in the case of Darwin's data it plays an important role in virtually eliminating the influence of unlikely parent distributions.

FIGURE 1(i)
REFERENCE DISTRIBUTION FOR θ WHEN
THE PARENT IS ASSUMED NORMAL.

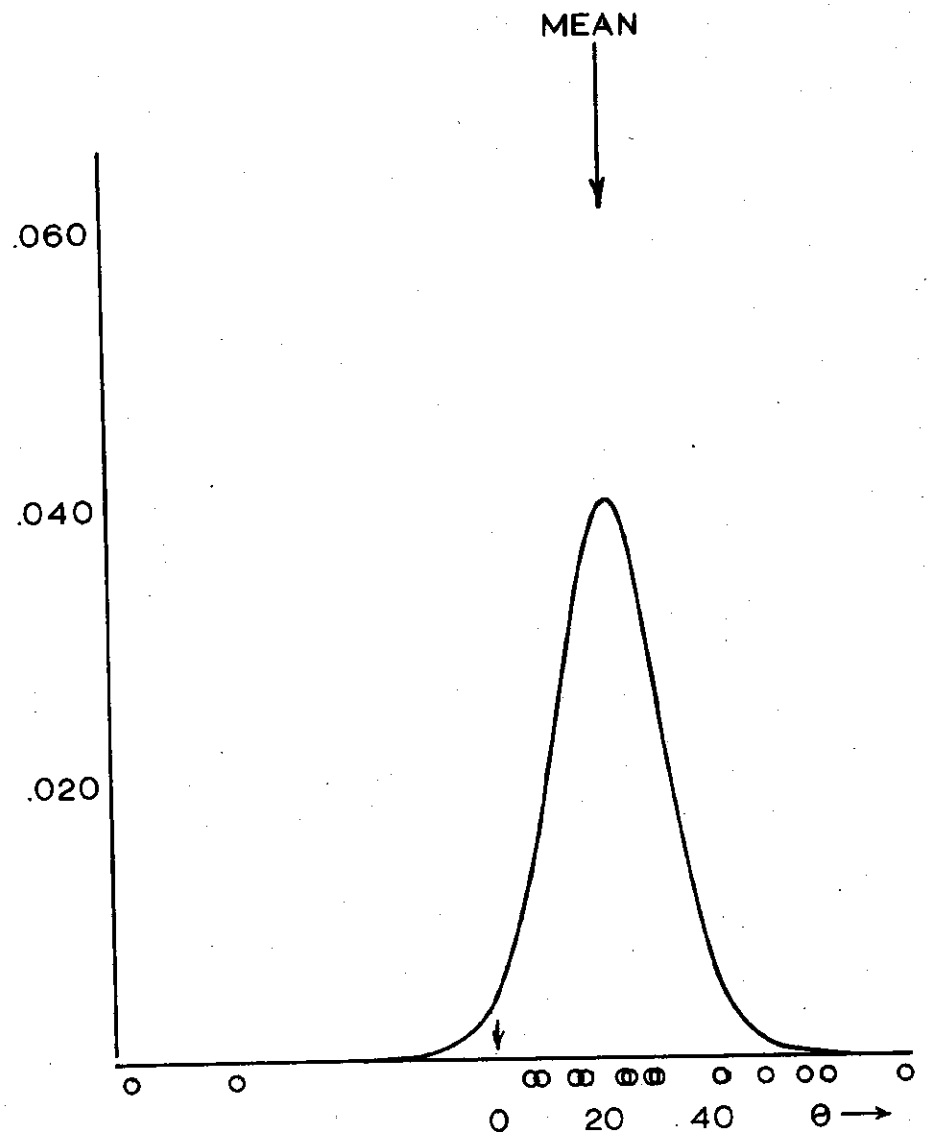


FIGURE 1(ii)
REFERENCE DISTRIBUTION FOR θ WHEN
THE PARENT IS ASSUMED RECTANGULAR.

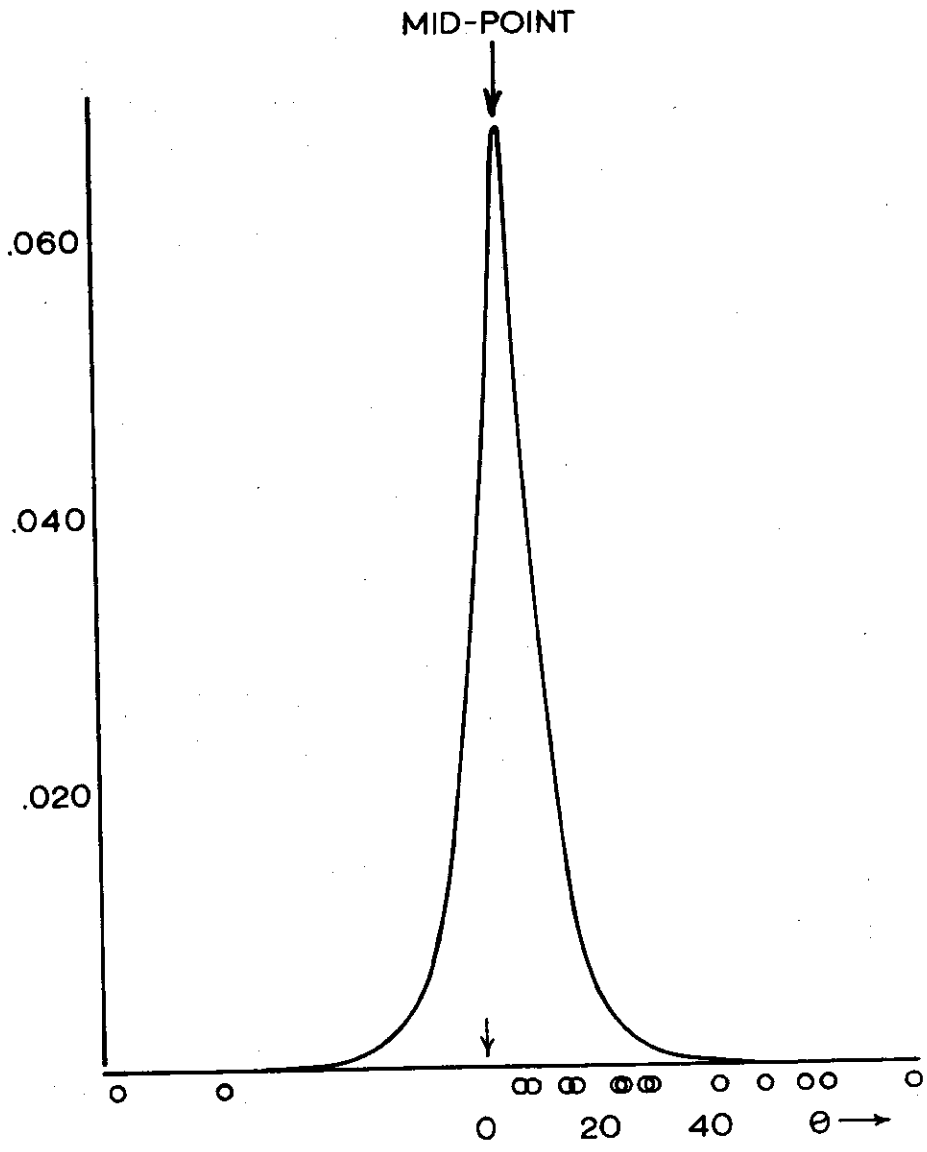


FIGURE 2(i)

SITUATION WHERE INFORMATION FROM THE SAMPLE REPRESENTED BY THE LIKELIHOOD $L(\theta)$ DOMINATES THE PRIOR PROBABILITY.

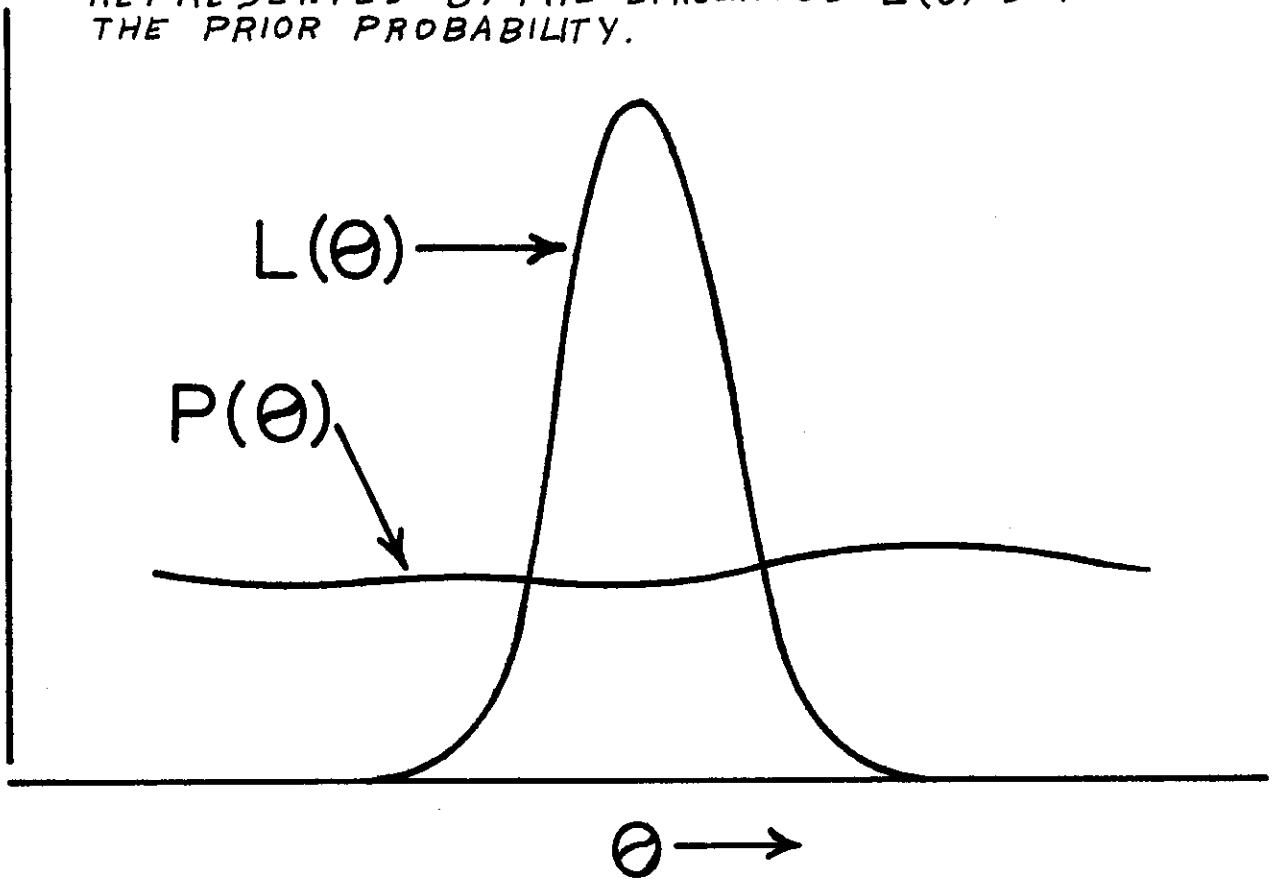
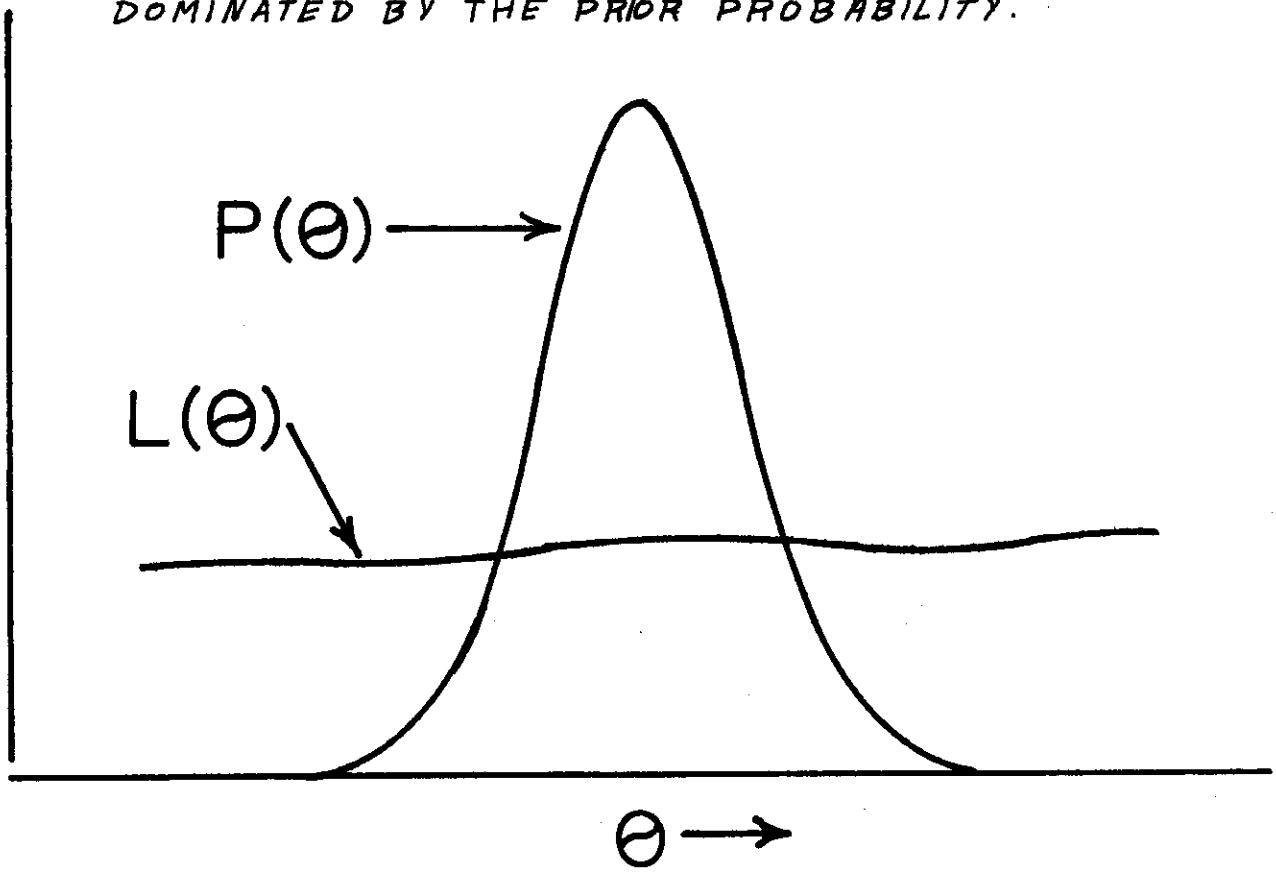


FIGURE 2(ii)

SITUATION WHERE INFORMATION FROM THE SAMPLE REPRESENTED BY THE LIKELIHOOD $L(\theta)$ IS DOMINATED BY THE PRIOR PROBABILITY.



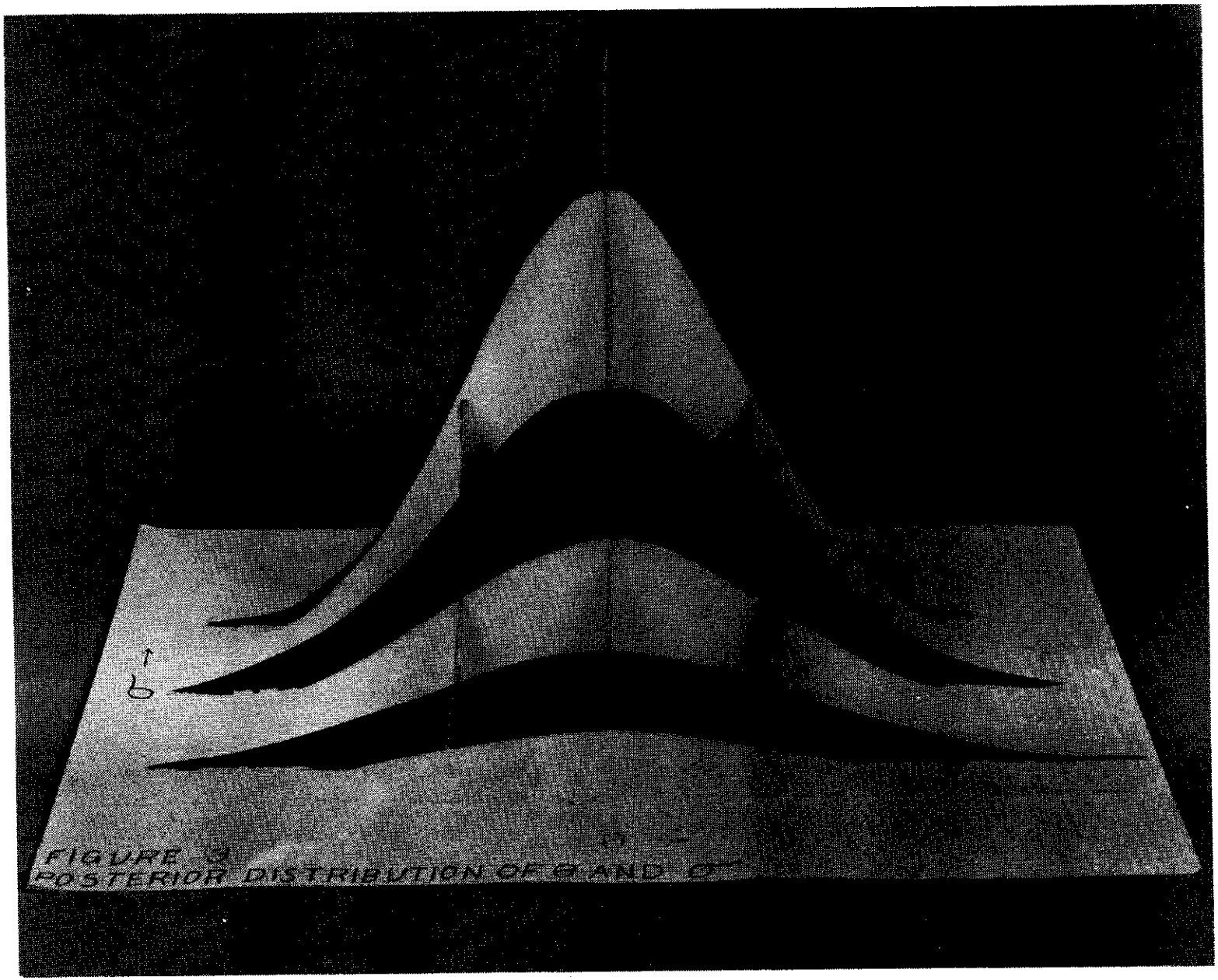
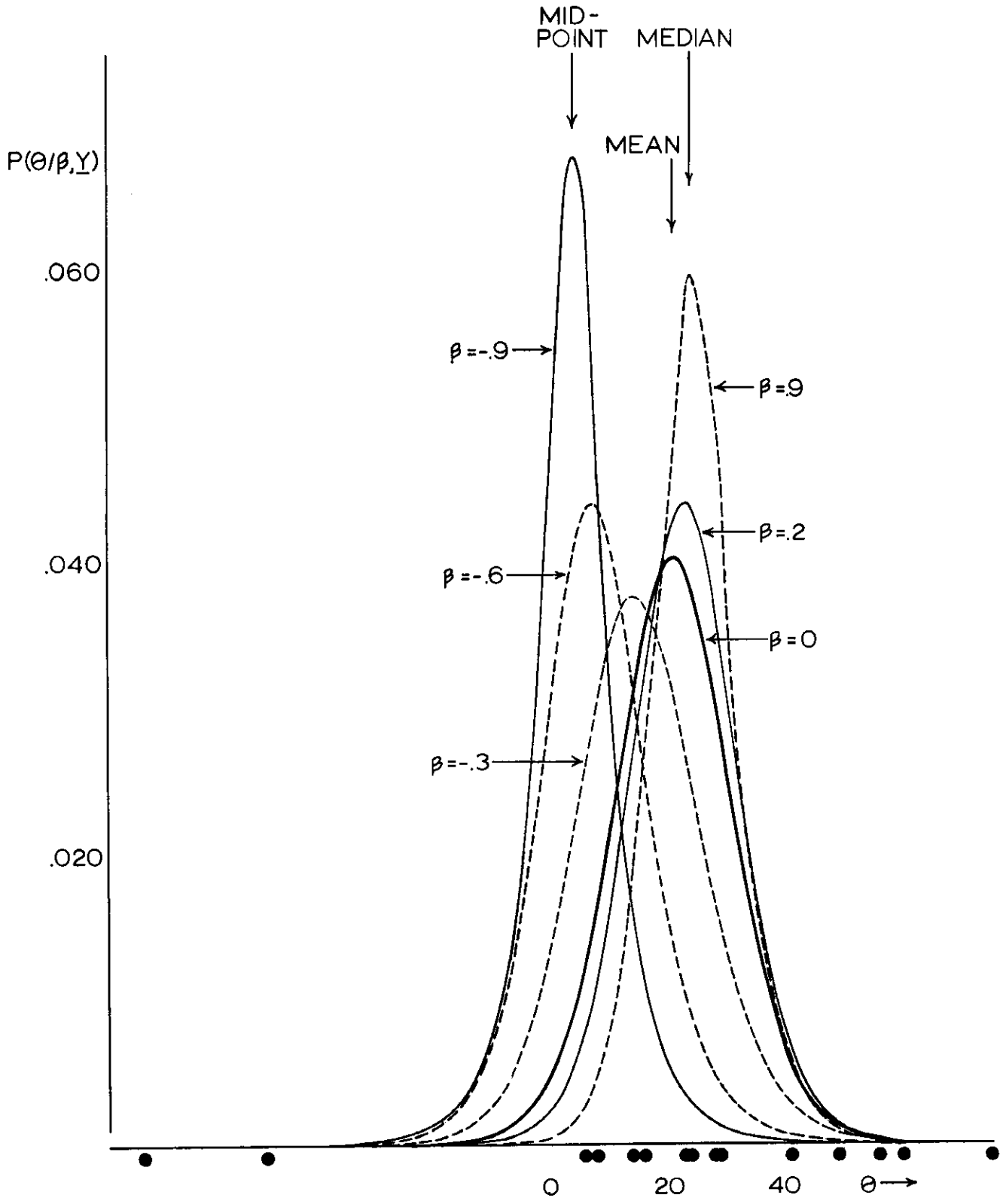
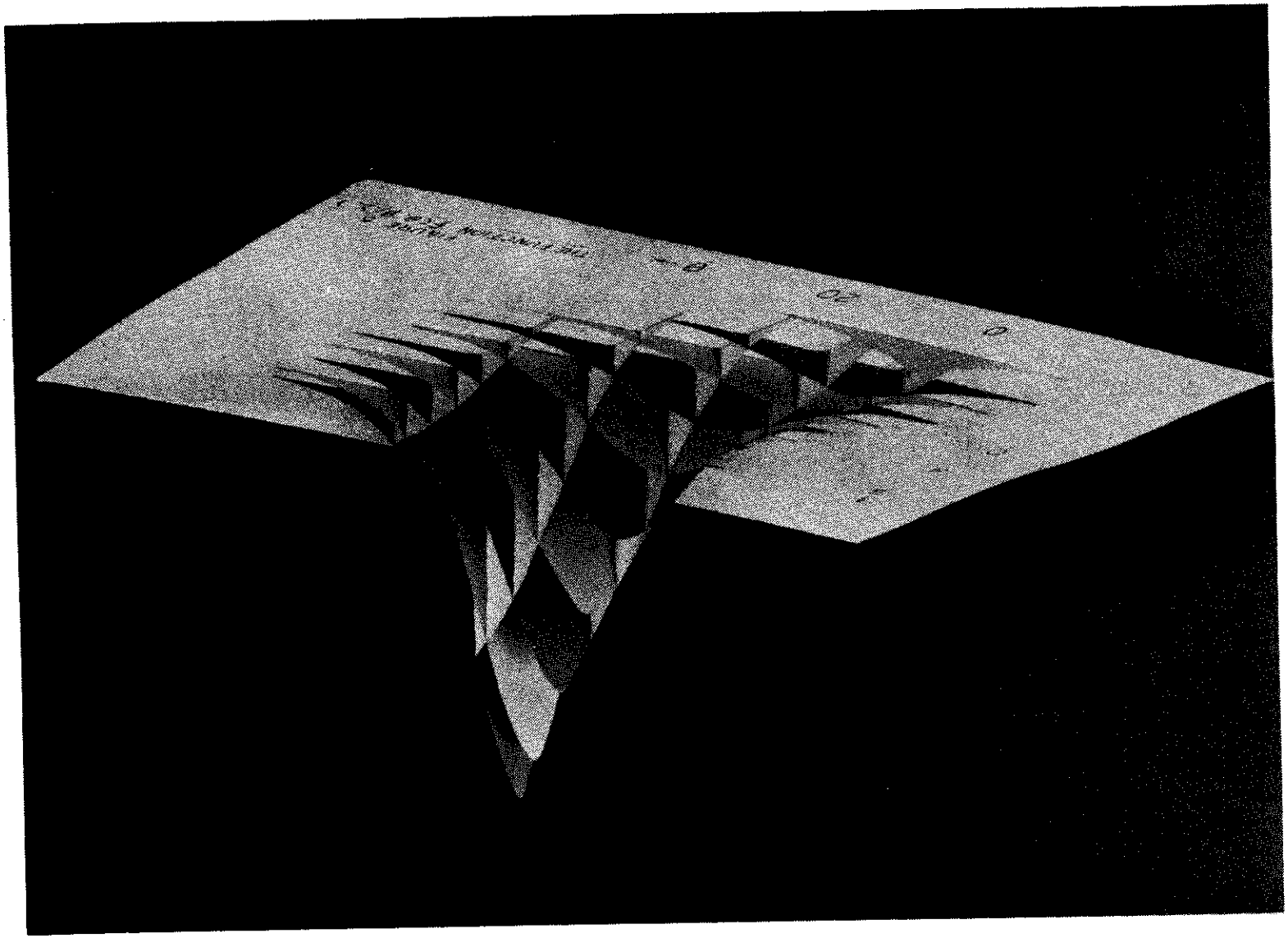


FIGURE 3
POSTERIOR DISTRIBUTION OF θ AND σ

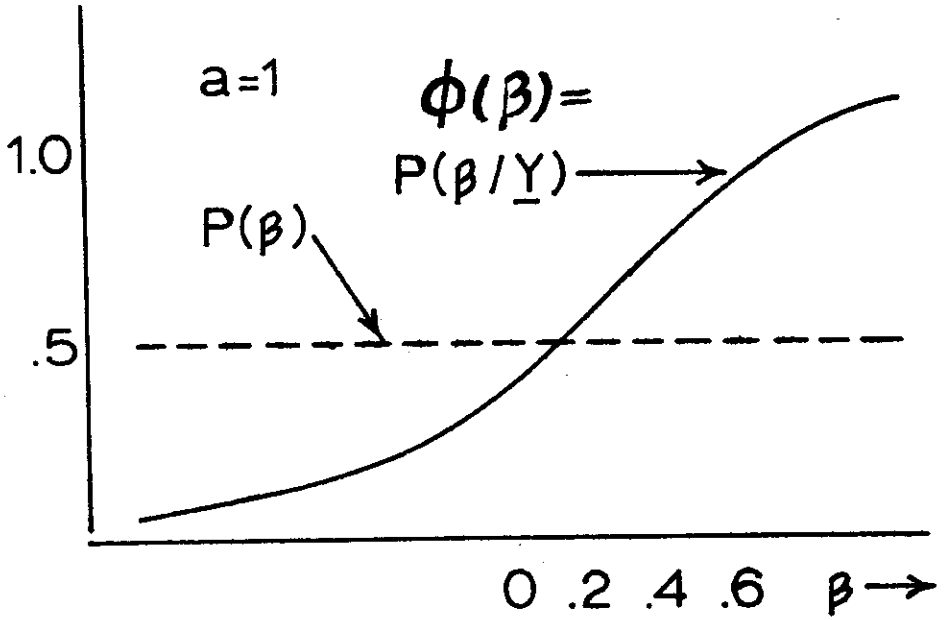
FIGURE 4
POSTERIOR DISTRIBUTIONS OF
θ FOR VARIOUS CHOICES OF β.



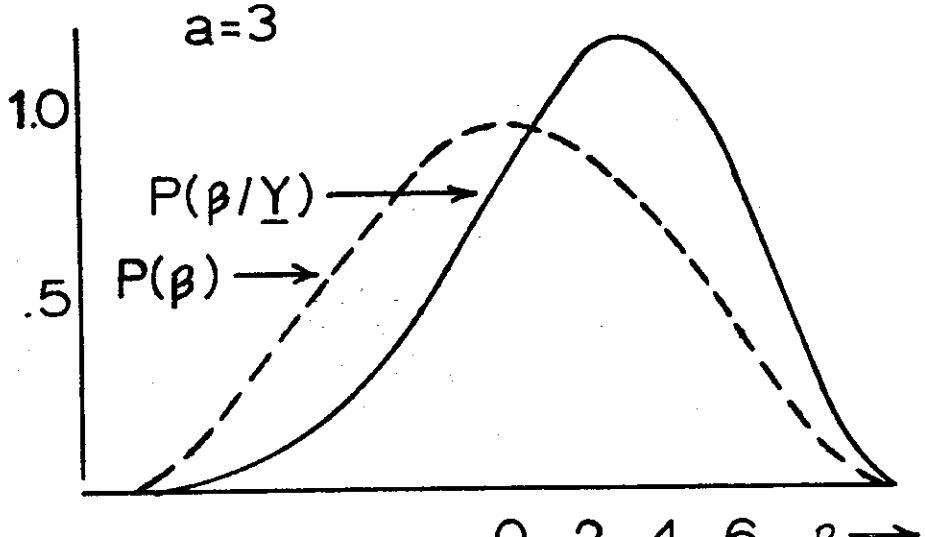


FIGURES 6(i)-6(iv)
PRIOR AND POSTERIOR DISTRIBUTIONS OF β FOR VARIOUS CHOICES OF a.

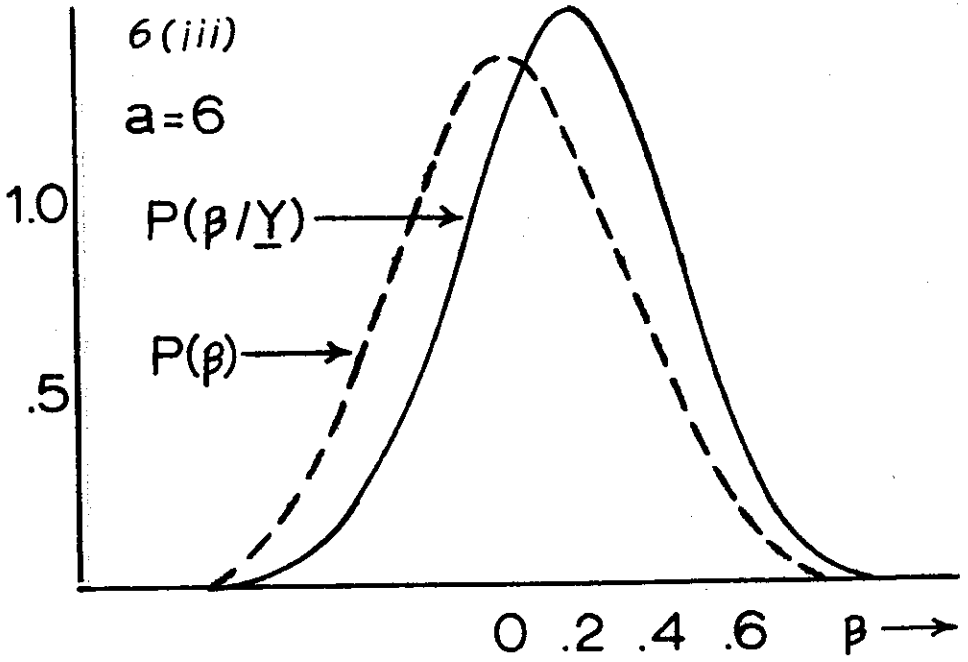
6(i)



6(ii)



6(iii)



6(iv)

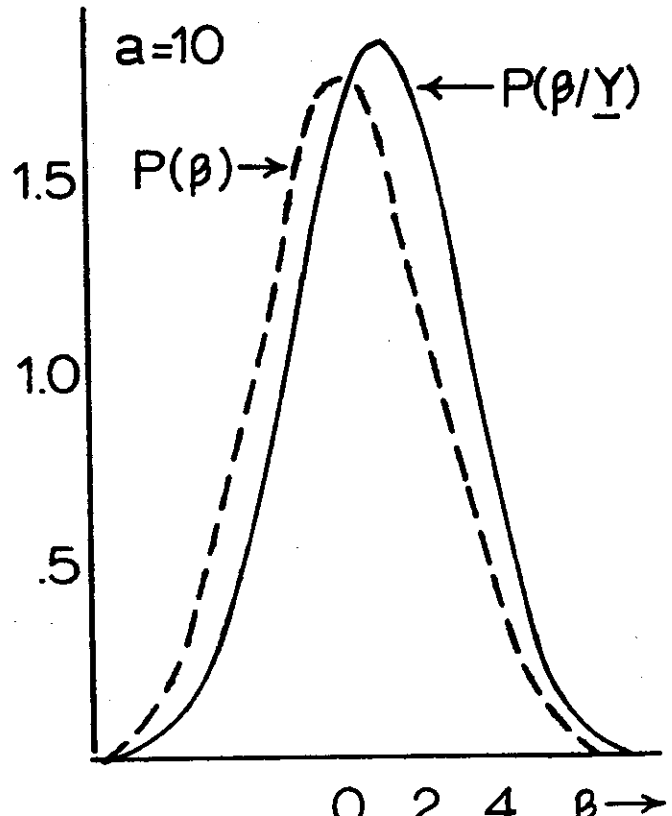
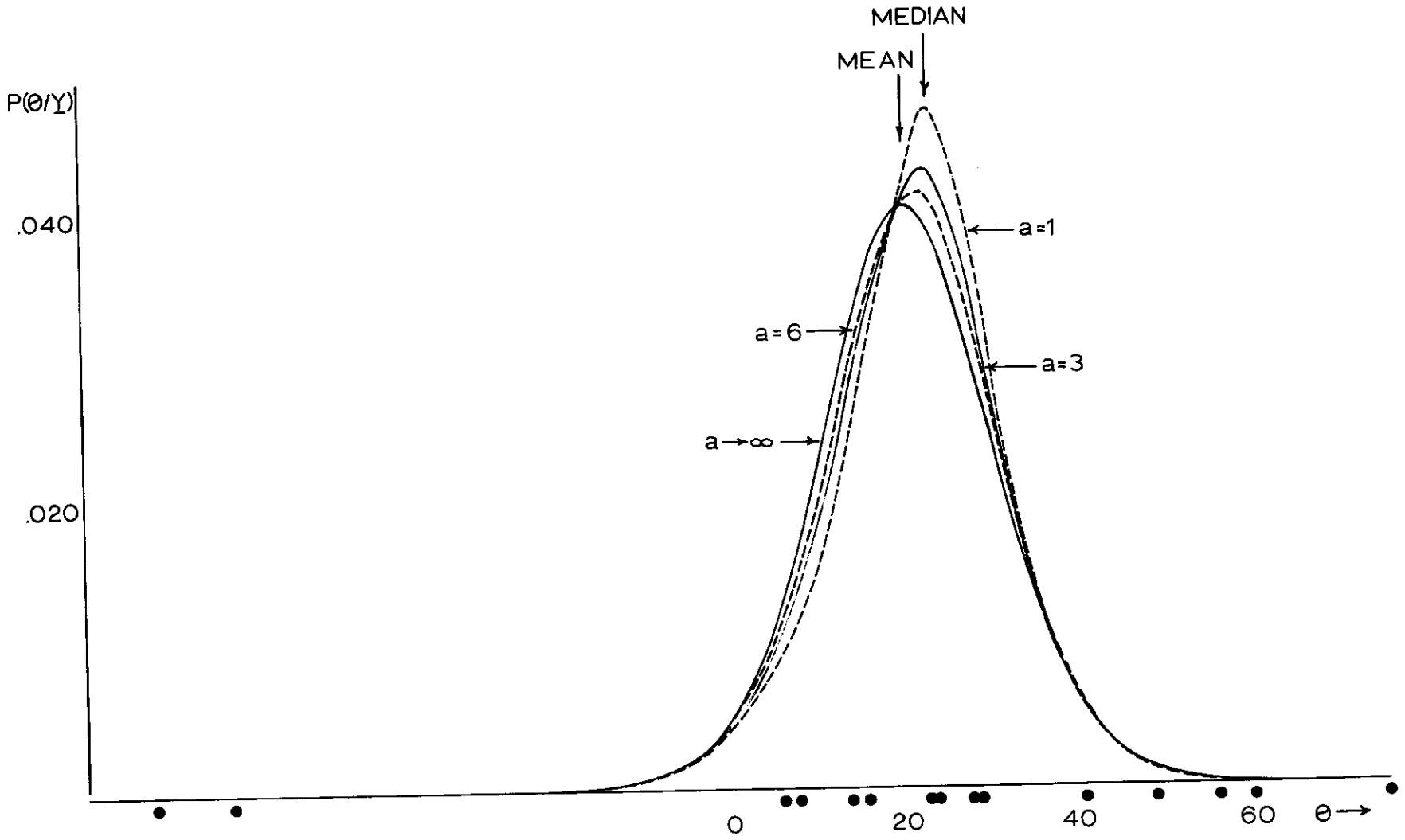


FIGURE 7
POSTERIOR DISTRIBUTIONS OF
θ FOR VARIOUS CHOICES OF β.



REFERENCES

1. Box, G. E. P., (1953), "A note on regions for test of kurtosis," Biometrika, Vol. 40.

2. Box, G. E. P. and Anderson, S. L., (1955), "Permutation theory in the Derivation of robust criteria and the study of departures from assumptions," J. R. S. S., Series B, Vol. 17.

3. Carlton, G. A., (1946), "Estimating the parameters of a rectangular distribution," Ann. Math. Stats., Vol. 17.

4. Daniels, H. E., (1954), "Saddle point approximation in statistics," Ann. Math. Stats., Vol. 25.

5. Diananda, P. H., (1949), "Note on some properties of maximum likelihood estimates," Proceedings of the Cambridge Philosophical Society, Vol. 45.

6. Fisher, R. A., (1935), "The Design of Experiments," Edinburgh: Oliver and Boyd.

7. Gayen, A. K., (1950), "The distribution of the variance ratio in random sample of any size drawn from non-normal universe," Biometrika, Vol. 37.

8. Jackson, D., (1921), "Note on the median of a set of numbers," Bull. Am. Math. Soc., Vol. 27.

9. Jeffreys, H., (1948), "Theory of Probability," Clarendon Press, Oxford, (2nd Edition.)

10. Jeffreys, H., (1957), "Scientific Inference," Cambridge (2nd Edition).

11. Jeffreys, H., and Jeffreys, B. S., (1956), "Methods of Mathematical Physics," Cambridge, (3rd Edition).

12. Kolmogoroff, A., (1941), "Confidence limits for an unknown distribution function," Ann. Math. Stats., Vol. 12.

13. Savage, L.J., (1954), "The Foundation of Statistics," Wiley.

14. Savage, L.J., (1959), "Subjective probability and statistical practice," Technical Note 59-1161, Air Force Office of Scientific Research.

15. Savage, L.J., (1960), "The foundation of statistics reconsidered," Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press.

16. Turner, M.C., (1960), "On heuristic estimation method," Biometrics, Vol. 16.

APPENDIX

In section 6 we have asserted certain properties of the posterior distribution $p(\theta/\underline{y}, \beta_0)$ in the permissible range of β. That they are so follows essentially from work on the median of a set of numbers by Jackson [8] some 40 years ago. For the class of parent distributions given by equation (5), the maximum likelihood estimates of θ for fixed β_0, which is the mode of $p(\theta/\underline{y}, \beta_0)$, was considered for certain specific choices of β_0 by Turner [16], who seems to have been unaware of the much more general result obtained by Jackson. In our notation, consider the function:

$$M(\theta) = \sum_{i=1}^{n} |y_i - \theta|^{\frac{2}{1+\beta}} \qquad \begin{matrix} -\infty < \theta < \infty \\ -1 < \beta < 1 \end{matrix}$$

For convenience, let us denote $q = \frac{2}{1+\beta}$ so that

$$M(\theta) = \sum_{i=1}^{n} |y_i - \theta|^{q} \qquad q > 1$$

(1) We first show that

(a) $M(\theta)$ is continuous and has continuous first derivative, and

(b) $M(\theta)$ has a unique minimum which is attained in $[y_s, y_L]$.

To see (a), consider

$$g_i(\theta) = |\theta - y_i|^{q} \qquad i = 1, 2, \ldots, n$$

Clearly $g_i(\theta)$ is continuous everywhere. Now,

for $\theta < y_i$

$$g_i'(\theta) = -q(y_i - \theta)^{q-1} ,$$

for $\theta > y_i$

$$g_i'(\theta) = q(\theta - y_i)^{q-1}$$

and as θ approaches y_i from both directions,

$$\lim_{\theta \uparrow y_i} g_i'(\theta) = \lim_{\theta \downarrow y_i} g_i'(\theta) = 0$$

which implies that $g_i'(y_i) = 0$.

Since $q - 1 > 0$, so that $g_i'(\theta)$ exists and is continuous everywhere. Our assertion (a) is proved since $M(\theta)$ is the sum of all $g_i(\theta)$.

Let us now consider $M'(\theta)$. We see that

when $$\theta < y_S$$

$$M'(\theta) = -q \sum_{i=1}^{n} (y_i - \theta)^{q-1} < 0$$

and when $$\theta > y_L$$

$$M'(\theta) = q \sum_{i=1}^{n} (\theta - y_i)^{q-1} > 0$$

Thus, by properties of continuous function, there exist at least a θ_0 , $y_S \leq \theta_0 \leq y_L$, such that

$$M'(\theta_0) = 0 .$$

Further, since $M'(\theta)$ is a monotonically increasing function of θ, we conclude that $M'(\theta)$ can vanish once and only once and that the extreme value of $M(\theta)$ must be a minimum. This completes the proof of assertion (b).

(2) It has been shown by Jackson that when q approaches 1, in the limit the value of θ which minimizes $M(\theta)$ is the median of the y_i's , if the latter is uniquely defined; and, if not, is some unique value between the middle two of the y_i's .

(3) We now show that, when q is arbitrarily large

$$\lim_{q \to \infty} [M(\theta)]^{\frac{1}{q}} = (h + |m - \theta|)$$

where $$m = \tfrac{1}{2}(y_L + y_S) \qquad h = \tfrac{1}{2}(y_L - y_S) .$$

Proof: Consider a finite sequence of monotone increasing positive numbers $\{a_n\}$ and a number S such that

$$S = \{ \sum_{i=1}^{n} a_i^q \}^{\frac{1}{q}}$$

we can write

$$S = a_n \{ \sum_{i=1}^{n} (\frac{a_i}{a_n})^q \}^{\frac{1}{q}}$$

where $\frac{a_i}{a_n} \leq 1$ for all i.

Hence $$(\frac{S}{a_n}) = \{ \sum_{i=1}^{n} (\frac{a_i}{a_n})^q \}^{\frac{1}{q}}$$

so that $$\log (\frac{S}{a_n}) = \frac{1}{q} \log \{ \sum_{i=1}^{n} (\frac{a_i}{a_n})^q \} .$$

When $q \to \infty$, $\lim_{q \to \infty} \log \{ \sum_{i=1}^{n} (\frac{a_i}{a_n})^q \} = \log r$ where $1 \leq r \leq n$

But this implies that

$$\lim_{q \to \infty} \log (\frac{S}{a_n}) = 0$$

whence $$\lim_{q \to \infty} S = a_n .$$

Thus, for any given value of θ, when q is arbitrarily large,

$$\lim_{q \to \infty} [M(\theta)]^{\frac{1}{q}} = \text{Sup} | y_i - \theta |$$

$$= \text{Sup} [| \theta - y_s | , | \theta - y_L |]$$

$$= \begin{cases} h + (m - \theta) & \theta < m \\ h & \text{for } \theta = m \\ h + (\theta - m) & \theta > m . \end{cases}$$

Hence, $$\lim_{q \to \infty} [M(\theta)]^{\frac{1}{q}} = (h + | m - \theta |)$$ and the assertion is proved.