THE STATISTICAL ANALYSIS OF GILL NET CATCHES

by

Burwell Gooch
Information Systems Division
Montana Department of Administration
Helena, Montana 59601

in cooperation with
Montana Department of Fish and Game
Helena, Montana 59601

September 1977

The Statistical Analysis of Gill Net Catches[1]

Burwell Gooch

Abstract

The validity of parametric versus nonparametric statistical analysis of gill net catch data is briefly discussed. It is concluded that the nonparametric ranking tests of Mann-Whitney and Kruskal-Wallis are the best methods available for analyzing such data. The theory and conduct of these tests are explained with the aid of simple examples. Also described are computational checks, various ways of handling tied observations, limitations on the use of the methods, and some peculiarities not associated with the more common parametric methods. Numerous examples are provided in order to illustrate pertinent points.

Introduction

Gill net sampling is widely used by fisheries biologists to obtain a variety of information about fish populations. One standard use is to gather information about the relative size of a fish population. Although such data cannot be manipulated so as to produce an estimate of the actual number of fish in a population, it nevertheless has seemed reasonable that, under the appropriate conditions, some kind of positive relationship should exist between the density of fish in a body of water and the number of fish caught in a series

---

[1]This study was part of Montana Federal Aid in Fish Restoration Project F-4-R.

of gill net sets that have been fished in that body of water. Furthermore, even though more refined methods of population size estimation do exist, physical and/or economic limitations often conspire to make gill netting the only practical means available for such work.

Since gill net catches do provide only an index to population size, the primary value of such figures is in reference to similar figures for other populations of fish (where "other populations" may represent fish in different bodies of water, in the same body of water at different locations, or in the same location at different times). That is, can we infer that population A is greater or smaller (or more or less dense) than population B, based on the gill net catches available from each? Clearly, this is the kind of situation in which statistical analysis was designed to play an important role, although a review of recent literature indicates that, with the exception of Moyle's pioneering work (Moyle, 1950; Moyle and Lound, 1960), this role has not received the attention that it should.

Statistical methods of analysis are conventionally divided into two broad categories. First, and far more popular, are the parametric methods. These are related by the fact that they all assume some kind of model and/or frequency distribution associated in some way with the population of interest. As a result, they are intimately concerned with estimating the

parameters that describe the assumed models and distributions. Usually, they are also constrained by several additional assumptions, some of which concern the population, others the sample. Second are the nonparametric methods. These do not specify a particular kind of model or parent frequency distribution, nor do they normally require so many other restrictions.

Parametric Methods

Most of the conventional methods of parametric statistical analysis are based on so-called "normal theory." This theory is predicated partly on the following three assumptions: (1) the attribute of interest (e.g., length, speed, temperature) occurs on a continuous scale of measure, rather than in discrete categories (e.g., as with counts); (2) the essentially infinite population of such attributes has a "normal" frequency distribution (Figure 1); and (3) the standard deviation of the distribution is independent of the mean. Under these conditions, and several others, estimates of population parameters and tests of hypotheses have been developed that are the best possible.

There are several reasons why statistical analysis in terms of normal theory enjoys such great popularity. Foremost among these is the fact that, when all assumptions are satisfied, no nonnormal method is as good. Second, the methods are "robust". This means that the power and efficiency of tests of hypotheses are not significantly reduced by reasonably small departures from the required assumptions. Third, a wide variety of parent

distributions from many different fields of study can be reasonably well approximated by the normal. Fourth, even if a parent distribution is not normal, the distribution of the sample mean approaches the normal with increasing sample size, thus preserving the validity of much of the theory.

The gill netting problems that Moyle attempted to resolve concern assumptions (2) and (3) listed previously. First, in the majority of series of catches, the frequency distribution of catch data is highly skewed, i.e., nonnormal (Figure 2). Second, the standard deviation is positively related to the mean (Figure 3). Moyle did not mention that assumption (1) also is not satisfied by gill net data.

Based on his comprehensive review of the situation, Moyle concluded that statistical analysis of gill net catches in terms of an assumed normal distribution is not legitimate. This means that it would not be valid to use the t test to compare the mean catches of two gill net series, or to use tabulated t values to impose confidence limits on these mean catches. I concur in his evaluation.

Moyle and Lound (1960) proposed two solutions to the problems stated above: one parametric, the other nonparametric. The parametric approach consists of assuming that gill net catches follow a negative binomial distribution. Under this assumption, a suitable transformation may be applied to the individual catches so that the resulting data are distributed more nearly normally. Thus, conventional confidence limits and t tests may be calculated.

There are two drawbacks to this proposal. First, even though many series of gill net catches do satisfactorily approximate a negative binomial distribution, the fact remains that a significant number do not. In other words, the negative binomial distribution is considerably more applicable than the normal distribution, but far from universally so. Second, the estimator of the dispersion index (1/k), which plays a central role in the theory of the negative binomial function, is simple to calculate, but does not provide the best possible estimate. The best estimate can only be obtained as the root of an explicitly insoluble equation. This can, of course, be found iteratively, but without a computer it is not practical to do so.

As a result of the foregoing, I feel that we may justifiably conclude that parametric statistical analysis is inappropriate in general for use on gill net catch data. Rather than use methods that, at best, fit the situation only part of the time, or spend more time searching for other theoretical distributions that might conceivably fit all the data, it seems more logical to investigate the possibility that some standard nonparametric method might be suitable for our needs.

Nonparametric Methods

Nonparametric methods possess a number of features that make them very attractive alternatives to parametric methods (Siegel, 1956). Four of these that are particularly relevant to our problem

of gill net catches are: (1) no particular form of the parent distribution is assumed; (2) both continuous and discrete data are equally well handled; (3) several tests are far superior to normal theory tests under conditions where the latter are inappropriate, whereas they are only slightly inferior where normal tests are appropriate; and (4) exact tests exist for any sample size, no matter how small.

The nonparametric method proposed for gill net catch analysis by Moyle and Lound (1960) is based on use of the median (rather than the mean) catch of a series. In addition to the advantages of a nonparametric method, the median method possesses three useful features: (1) confidence limits may be constructed about the median, although without appropriate tables this may be quite tedious; (2) the test of a difference between sample medians is simple to calculate and understand; and (3) the test may be applied to any number of samples. The primary disadvantage of the method is that it does not provide the most efficient analysis of the data possible. That is, other methods exist that extract more information from the samples, and thus discriminate more effectively between them. The purpose of this paper is to promote the understanding and use of these methods.

Ranking Tests

The methods that appear to be best suited for use on gill net catch data belong to the class of so-called ranking tests. The one that is probably most appropriate is the Kruskal-Wallis H test. This test may be applied to a comparison of any number of samples (i.e., two or more). Another test, known variously as the Mann-Whitney U test (Siegel, 1956) or the Wilcoxon two-sample test (Alder and Roessler, 1964), is a special case of the H test that is applicable to two-sample comparisons only. These tests are designed specifically to discriminate between samples that are drawn from populations with different central tendencies, regardless of the dispersion around the central tendencies.

A basic feature of the methods is the ordering of the sample observations from smallest to largest, and the assignment of ranks according to this ordering. In this procedure, all observations from all samples are combined and treated as one group. The assumption is that if all samples have been selected from the same population (or from populations with the same frequency distribution), then all of the sample mean ranks should be similar. On the contrary, if the samples have been selected from populations with different frequency distributions, then we would expect these differences to be reflected in the sample mean ranks.

Some simple examples will help to clarify the foregoing ideas and provide a basis for discussion of the tests themselves.

Example 1. The purpose of this example is to illustrate the ranking of observations from two samples. The data were chosen to

Sample 1		Sample 2	
Observation	Rank	Observation	Rank
7	2	6	1
10	5	8	3
13	8	9	4
16	11	11	6
19	14	12	7
		14	9
		15	10
		17	12
		18	13
		20	15
Total 65	40	130	80
Mean 13	8	13	8

demonstrate the consequences, on the average, of drawing two samples from the same population: equality of sample mean observations (13), and equality of sample mean ranks (8). Normally, in any given instance, sampling variability would prevent two such samples from turning out precisely this way (Example VI, Illustrative

Examples). Note that the difference between the sums of the sample ranks (40 and 80) is not a reliable indicator of a population difference, any more than is the difference between the sums of the sample observations. This follows from the fact that the two sample sizes are unequal.

Example 2. Assume that two series of gill net sets result in the catches displayed below:

Rank Order Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Catch	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

These data are already combined and ranked, those representing Sample 1 having been underlined. Assume further that the following hypotheses were proposed prior to selecting the samples:

H_0: The two populations are equally abundant (dense)

H_A: Population 1 is less abundant (dense) than Population 2.

In order to test the null hypothesis, we must determine first, the total number of ways that the seven Sample 1 catches can occur among the sixteen rank positions available; and second, the number of ways that the Sample 1 catches can occur in positions as low as or lower than those actually shown.

The first figure is

$$\binom{16}{7} = \frac{16!}{(16-7)!7!} = 11,440$$

The second is obtained by actually listing each possibility and counting them. This task is simplified somewhat if it is done in terms of the sum of the ranks; that is, we must find each rank ordering whose rank sum is less than or equal to the rank sum in the actual case (32). Since there are 12 possible rankings that satisfy this condition (Table 1), the probability of obtaining the actual or a more extreme result under the null hypothesis is $12/11,440 \simeq 0.00105$. Thus, we have little difficulty in rejecting the null hypothesis.

Example 3. Assume the same conditions as outlined in Example 2 except that prior to sample selection we have no reason to think that Population 1 is necessarily less dense than Population 2. That is, it seems just as likely to be more dense as less dense. Thus, our alternative hypothesis should be changed to read:

H_A: The two populations are not equally abundant (dense).

This situation is an example of two-tailed hypothesis testing, as opposed to one-tailed hypothesis testing which was the case in Example 2. It is called two-tailed because we must be concerned with both low and high rankings that are as extreme or more so than the one actually obtained. For each of the 12 low extremes listed in Table 1, there is an opposite high extreme that is just as likely under the null hypothesis (Table 2).

Since this brings to 24 the total possible rankings that satisfy the stated conditions, the probability of obtaining the actual or a more extreme result under the null hypothesis is $24/11,440 \simeq 0.0021$. Thus, even when the alternative hypothesis is two-tailed, the sample evidence still strongly favors rejection of the null hypothesis.

The results demonstrated in the last two examples are indicative of the general relationship between one- and two-tailed tests: the probability associated with a two-tailed test is double that associated with the corresponding one-tailed test.

Fundamental Properties of the Rank Sum

The notation associated with the ranking tests to be discussed in this paper is as follows:

s = number of samples to be compared

n_i = size of sample i

$n = \sum_{i}^{s} n_i$

R_i = sum of ranks of sample i

$\bar{R}_i = R_i/n_i$ = mean rank of sample i.

R_i is the sample statistic that is the basis of the tests subsequently described. Because it is a sum of count data, it is a discrete variable, i.e., its theoretical frequency distribution has the appearance of a histogram. Under the null hypothesis,

this distribution exhibits the following properties:

1. Minimum: $R_i(\min) = n_i(n_i+1)/2$
2. Maximum: $R_i(\max) = n_i(2n-n_i+1)/2$
3. Mean: $\mu_i = n_i(n+1)/2$
4. Variance: $\sigma_i^2 = n_i(n-n_i)(n+1)/12$
5. Symmetry
6. Approach to the normal distribution as n_i increases in size.

If all n_i are the same, the distributions of all R_i coincide. In the usual case where $s = 2$ and $n_1 \neq n_2$, the ranges (maxima, minima, and means) of the distributions of R_1 and R_2 are different, but the shapes (variances) remain the same. That is,

$$\sigma_1^2 = n_1n_2(n+1)/12 = n_2n_1(n+1)/12 = \sigma_2^2$$

and

$$R_1(\max) - R_1(\min) = n_1(2n-n_1+1)/2 - n_1(n_1+1)/2 = n_1n_2$$

$$R_2(\max) - R_2(\min) = n_2(2n-n_2+1)/2 - n_2(n_2+1)/2 = n_1n_2$$

In fact, these distributions are really mirror images of each other: the occurrence of a particular value of R_1 a distance d above its minimum corresponds to a simultaneous occurrence of R_2 a distance d below its maximum. This follows from the fact that the sum of R_1 and R_2 must always equal $n(n+1)/2$.

Example 4. Suppose that samples of size $n_1 = 5$ and $n_2 = 10$ are to be selected, each from a different population. Under the null hypothesis, the distributions of R_1 and R_2 would be characterized as follows:

Attribute	R_1	R_2
R_i(min)	15	55
R_i(max)	65	105
μ_i	40	80
σ_i^2	$66^2/_3$	$66^2/_3$

Clearly, $\sigma_1^2 = \sigma_2^2$ and $R_1(\max) - R_1(\min) = R_2(\max) - R_2(\min)$.

Suppose that after the samples have been selected, $R_1 = 32$ and $R_2 = 88$. Then,

$$R_1 - R_1(\min) = 32 - 15 = 105 - 88 = R_2(\max) - R_2$$

and

$$R_1(\max) - R_1 = 65 - 32 = 88 - 55 = R_2 - R_2(\min)$$

This demonstrates the mirror-image character of the distributions.

We may note also that the true means μ_1 and μ_2 shown for this case are the same as the sample rank sums shown in Example 1, which was designed to represent the average result of sampling from identical populations. Although, as pointed out in the discussion of that earlier case, the sample mean rank ($\bar{R}_i$) is the relevant statistic for purposes of comparison, it is an interesting fact that neither of the two test methods described below explicitly involves a function of this statistic. The first method avoids this by appealing directly to the theoretical distributions of the R_i, in particular the lower tails where extreme values occur. (The upper tails are just as appropriate, but use of the lower

tails has been established by convention.) The second method actually does involve the mean ranks, but not in a way that requires their calculation.

U Test (s = 2)

In our discussion above we learned that, for the two-sample case, the statistical information provided by R_1 and R_2 is identical, resulting from the fact that the distributions of these two statistics are mirror images of each other. This means that a test of hypothesis based on the appropriate tail of one distribution is identical to a test based on the opposite tail of the other distribution. Since the range of values for R_1 is not, in general, the same as that for R_2, there is a problem of (1) choosing which distribution to tabulate, and (2) doing it in such a way that the lower tail will always be appropriate for a test.

One solution to this problem is provided by the U test. This test consists of calculating the statistic

$$U_i = R_i - n_i(n_i+1)/2 \qquad (1)$$

for each of the two samples to be compared, and then determining which of the two is smaller, called U:

$$U = \text{Min}(U_1, U_2)$$

Calculated U is compared to tabulated U (Table 3, 4, or 5) to determine the significance of the sample comparison. Table 3,

for example, lists the U values for each combination of n_1 and n_2 up to $n_1 = n_2 = 20$ that delimit the lower 10 percent of the distribution of U. For one-tailed tests of hypotheses, this percentage represents the significance level of the table. For two-tailed tests of hypotheses, the correct percentage of the significance level is double this, viz., 20 percent. Tables 4 and 5, respectively, represent the 5 and 2.5 percent values of U for one-tailed tests, and 10 and 5 percent values of U for two-tailed tests.

As an example of their use, suppose that for a two-tailed test of no difference in population distributions, $n_1 = n_2 = 10$, U = 30, and the chosen significance level is 10 percent. According to Table 4, for the sample sizes given, U cannot exceed 27 to be significant at the 10 percent level. As a result, the null hypothesis cannot be rejected.

There are at least two circumstances in which it is desirable or necessary to use U as a basis for a two-sample comparison. First, when both n_i values are less than 10, the distributions of the R_i are not close enough to the normal to justify using the alternative H test to be described subsequently. Since the entries in Tables 3, 4, and 5 are based on the exact distributions of U (hence, R_i), these should be used for small-sample comparisons.

Second, for the purposes of many investigations, there is no need to go through more complicated calculations in order to closely approximate the probability of a particular test; a simple "accept" or "reject" of H_0 based on the U statistic is entirely adequate. For this reason, Tables 3, 4, and 5 include entries for samples up to size 20. Beyond this range it is necessary to rely on the H test.

The rationale behind the U test is quite straightforward. The idea is to operate on the R_i values in such a way that the distributions of the resulting statistics coincide. The function U_i accomplishes this by subtracting from R_i the minimum value that R_i can take, i.e., $R_i(\min)$. Thus, the range of the distribution is changed from

$$\{R_i(\min), R_i(\max)\} = \{n_i(n_i+1)/2, n_i\{2n-n_i+1)/2\}$$

to

$$\{0, n_1n_2\}$$

In this way the values of U_1 and U_2, respectively, that delimit the lower α percent of the distributions of U_1 and U_2 are exactly the same. Therefore, only one table of the lower tail of U values needs to be constructed for a given probability level, simply by ensuring that U is the smaller of the two U_i values.

An alternative method that is equivalent to U is found in some texts. This consists of (1) selecting that R_i value that corresponds to the smaller of n_1 and n_2; (2) calculating $R_i' = n_i(n+1) - R_i$; (3) determining which of R_i and R_i' is smaller

(note the resemblance to the procedure for U); and (4) comparing this smaller value to a table of such values. The only difference between the two methods is that in the case of U the distributions are transformed so that their minimum always occurs at the origin. This means, of course, that the tables for one method are not appropriate for the other.

H Test ($s \geq 2$)

The H test consists of calculating the statistic

$$H = \frac{12}{n(n+1)} \sum_{i}^{s} R_i^2/n_i - 3(n+1) \qquad (2)$$

This test statistic is distributed approximately as chi-square with s-1 degrees of freedom under the following sets of conditions:

1.a. $s = 2$

b. $n_i \geq 10$ for at least one i

2.a. $s = 3$

b. $n_i > 5$ for all i

3.a. $s > 3$

b. $n_i \geq 5$ for all i

If condition 1.a is satisfied, but not 1.b, then the U test described previously should be used.

If condition 2.a is satisfied, but not 2.b, the H test may be used but the test statistic should be compared to the entries in Table O by Siegel (1956). This table lists the true probabilities associated with various values in the

upper tail of H for values of n_1, n_2, and n_3 up to 5. If at least one n_i is as large as 5, and no n_i is less than 2, then tabulated chi-square actually provides very good approximations to the true probabilities above P = 0.10 (Kruskal and Wallis, 1952). Below P = 0.10, or failing the constraints on the n_i, the true probabilities tend to be overestimated. Thus, use of the chi-square table will, in general, provide a conservative test, one that is less likely to reject H_o than when Siegel's Table O is used.

If condition 3.a is satisfied, but not 3.b, I know of no tables of exact probabilities, comparable to Siegel's Table O, that may be resorted to. As an alternative to tables, Kruskal and Wallis propose the use of two functions that are more accurate than H. However, I do not feel that the usual quality of gill net catch data, plus the relatively small increase in accuracy of the functions, justify the additional complications required by their use.

Relying on known facts for the case where s = 3, I think we may reasonably assume that if the chi-square table is used to approximate the probability associated with H, our test will be a conservative one.

Because of the nature of the chi-square distribution, it is tabulated in terms of two-tailed probability levels.

Consequently, if a two-tailed test of hypothesis is being considered, the probability of H is the tabulated value found; under a one-tailed test, the correct probability is half the tabulated value.

In using H for a one-tailed test, the investigator should be careful to ensure that a significant H does in fact support the alternative hypothesis. He can do this by calculating the individual $\bar{R}_i$'s and verifying that their ranking with respect to each other corroborates H_A. As has already been pointed out, the R_i's are not appropriate for this because they represent sums, not means.

It may be noted here that two equivalent formulations of H are given by

$$H = \frac{12}{n(n+1)} \sum_i^s n_i \bar{R}_i^2 - 3(n+1)$$

$$= \frac{12}{n(n+1)} \sum_i^s n_i \{\bar{R}_i - (n+1)/2\}^2$$

These verify the relevance of the $\bar{R}_i$'s. Expression (2) is preferable as the definition formula of H, however, because it requires fewer calculations.

Computational Checks

Because R_i and U_i values are fundamental quantities in the test methods described, the careful investigator would like to be assured that they have been calculated correctly. Fortunately, certain relationships exist that may be taken advantage of in this context. These are as follows:

1. The rank of the largest observation in the combined sample must be equal to n
2. $\sum_{i}^{s} R_i = n(n+1)/2$
3. $U_1 + U_2 = n_1 n_2$

Tied Observations

The characteristics associated with the statistics and test methods described in the foregoing are based on the assumption that each sample observation is unique (different from all other sample observations for all samples combined), and would always be so no matter how many samples are selected. In general, this seems to be a rather unrealistic assumption, and is particularly so in the case of gill net catches.

Of the several methods that have been suggested for dealing with tied observations, only one seems to have gained popularity among text-book writers. This is the mean-rank method. It consists of (1) assigning to each observation in a group of ties the mean of the ranks associated with those ties; and (2) in those cases where the H test is used, increasing the calculated value of H by a factor that is a function of the number of ties in each group of ties. This expansion factor presumably accounts for the fact that the variance of each R_i is smaller as a result of the ties, which in turn produces a more precise test. Although this method is relatively easy to apply and always leads

to a unique test result, it possesses two disadvantages that I feel imperil its validity.

First, it assumes that the configuration of ties obtained for a given set of samples is identical to what would be obtained for all possible sets of samples selected under the same conditions. Obviously, this is not reasonable. The configuration obtained in any given case is just one of a great many possibilities. Statisticians evade this problem by resorting to the argument that the variances and test results are conditional on the configuration obtained. This means that the probability associated with the test result, hence the acceptance or rejection of H_0, is relevant in the context of only that subset of samples, of all possible sets, that exhibit this configuration. I believe that few users of such a method would have much faith in their results if they understood how narrowly these were constrained.

The second objection to the mean-rank technique is that it destroys the comparability of the two test methods U and H. Since the U test does not provide for a reduced variance in the case of tied observations, whereas the H test does, it is obvious that the two methods of analysis will lead to different test results in those cases where either may be used. In extreme cases it is possible to accept H_0 under the U test and reject it under the H test. Although the two tests usually are conceived of as mutually

exclusive on the basis of sample size, this seems to be an unreasonable limitation on the use of U by those investigators who prefer it to H. Moreover, even if this dichotomy of use is adhered to rigidly, there still remains the difficulty of justifying the incompatible philosophies associated with the two methods.

A second technique for dealing with tied observations consists of assigning ranks in such a way as to minimize the difference between the $\bar{R}_i$'s, thus minimizing the probability of rejecting H_o. This represents an extreme example of a conservative test. If H_o can be rejected under these circumstances, the investigator may be even more confident of his results than is indicated by the chosen significance level.

My experience in applying this method to gill net catch data leads me to believe that it is rather too conservative; that many cases of legitimate differences between populations would go undetected simply because of the large number of tied observations that so commonly occur among such data.

A third method that appears to be the most satisfactory for resolving the problem of tied observations is to assign the associated ranks at random. This procedure overcomes all of the objections to the previous two methods because it is completely compatible with all of the requirements imposed by the null hypothesis on the derivations of equations (1) and (2).

If data analysis is computer automated, this technique presents no particular burden. If, on the other hand, data analysis must be performed manually, the method can be somewhat tedious if the volume of data is rather large. In this regard, the following key to a suggested methodology may be helpful:

A. All observations in a group of ties occur as members of one sample. In this case, random assignment of ranks is unnecessary; the results are the same no matter how the ranks are assigned.

B. A group of ties includes observations from two samples.
   1. Determine the range of ranks involved (e.g., 07-11).
   2. Using a table of random numbers, randomly pick values that match those in the appropriate range.
   3. Assign these one at a time to the tied observations associated with the sample with the smaller number of ties in the group. (This minimizes the table search.)
   4. When all ties in this sample have been randomly assigned a rank, assign all remaining ranks to the ties in the other sample. There is no need to do this randomly, the result is the same regardless of how it is done.

C. A group of ties includes observations from three or more samples (multiple comparison situation). In this case,

proceed as in B except that, in step 4, continue assigning ranks at random to the sample with the next larger number of ties in the group, and so on, until all samples have been taken care of except the one with the largest number of ties in the group. Then assign all remaining ranks to the ties in this last sample.

Example 5. Assume that 1 + 3 + 6 = 10 observations from three samples (A, B, and C) occur as ties in one group as shown below.

Rank Order Position	6	7	8	9	10	11	12	13	14	15
Catch	4	4	4	4	4	4	4	4	4	4

Since the range of ranks involved is 06-15, we must select values from the random number table that occur in this range. The first rank selected is assigned to the one observation from Sample A (fewest number of ties in the group). The next three ranks selected are assigned to the three observations associated with Sample B (next larger number of ties). At this point, the random number table is no longer needed because all remaining ranks are assigned to the six observations associated with Sample C.

Using the random number table provided by Hodgman (1959), the following actual results were obtained: 09, 06, 07, 11. Thus, rank 9 is assigned to Sample A; ranks 6, 7, and 11 are assigned to Sample B; and ranks 8, 10, 12, 13, 14, and 15 are assigned to Sample C.

One possible objection to the random-rank method is that the decision to accept or reject H_0 is based on a "flip of a coin", i.e., on the random assignment of the ranks. This is not a rational argument, however. Although it is certainly true that the outcome of the random assignments will influence the test result, its effect in the probabilistic sense is no different from that associated with the outcome of the random selection of population units to be included in the sample. Random assignment of ranks is simply an extension of the sample selection process. An investigator bases his decision to accept or reject H_0 on the "flip of a coin" just as much in the context of sample selection as he does in the context of rank assignment.

Limitations

As with all experimental methods, those described in this paper for comparing the relative abundance of fish populations are constrained by certain limitations. These are associated with (1) the effect of factors other than abundance on gill net catches; (2) the relationship between abundance and gill net catches; and (3) assumptions required by the ranking tests.

Extraneous Factors

Most investigators who use gill nets to study population abundance are aware that numerous factors other than abundance can significantly influence their catches. This results from

the fact that gill nets, as fishing gear, are both passive and selective.

Gill nets are passive fishing gear in the sense that once they have been set in the water, they rely entirely on the movements of the fish to generate a catch. Consequently, any factor that affects the movement of fish (in particular, amount of movement, location of movement, and tendency to congregate) can in turn affect a gill net catch, entirely independent of abundance.

Broadly speaking, such factors belong to two categories: (1) those related to the fish (mostly species specific), and (2) those related to the environment. The latter category can be divided further into those associated with the body of water and those associated with the environment outside the body of water.

Some of the more important factors that may affect fish movements on a seasonal basis are: spawning activity, water level, physico-chemical properties of the water, water temperature, barometric pressure, light, food supply, and predators. The last six of these can also play a significant role on a much smaller time scale, even daily. Two others that are particularly important, but not necessarily seasonally, are underwater topography and sources of inflow.

Although an investigator cannot prevent these factors from having an influence on his gill net catches, that is not important. What is important is that he can and should select

his samples so that as many of these factors as possible have an equal effect on each of the samples, thus minimizing irrevalent variation. A species difference, of course, is rarely if ever a source of variability because there generally is no meaning to be derived from comparing the abundance of one species to another, either in the same body of water or in different waters. Studies of individual species in one body of water over time, with constancy maintained in as many outside factors as possible, represent the ideal situation. Gill net studies in Montana are of this nature. It is apparent, however, that there are needs elsewhere to compare populations from different bodies of water (Moyle, 1950). In such cases, consideration must be given to the possibility that differences in gill net catches result in part or in whole from differences in fish behavior that are caused by habitat differences.

Gill nets are selective fishing gear because a given mesh size is best at catching fish with a certain girth size; the greater the deviation from this girth size the less the probability of capturing the fish (ignoring capture by entanglement). Thus, there are two ways that selectivity can influence the catch of a gill net: (1) on the basis of the mesh size(s) used, and (2) on the basis of the size distribution of fish in the population. Normally, one would want to use gill nets with the same mesh sizes when comparing catches from different populations of fish. This means, however, that a significant difference in catch between the populations may

be the result of a difference in size distribution rather than in abundance. Investigators who are interested only in the latter, therefore, may wish to conduct further studies in order to distinguish between the two possibilities.

Abundance

It seems likely that abundance itself may have an effect on gill net catch that is independent of its effect in terms of numbers. For example, it would not be unreasonable to expect changes in behavior to accompany reasonably large changes in abundance of a fish population. As a result, everything else being equal, a given change in abundance might lead to a greater or lesser change in gill net catch, depending on the behavioral changes in the fish. This compound effect of abundance on gill net catches creates difficulties in interpreting differences between them.

One can imagine, for example, two populations of fish where a difference in abundance of, say, 20 percent is reflected on average as a difference in gill net catch of 25 percent. If the investigator has chosen a 25 percent difference in abundance as representing the minimum level of biological significance, and uses the gill net difference as a measure of this criterion, then he would incorrectly conclude that these two populations are significantly different (assuming that statistical significance is also attained).

To my knowledge, the relationship between population abundance and gill net catch has never been studied. In lieu of such information, therefore, the investigator should choose a level of biological significance that is large enough to compensate for discrepancies of the kind and magnitude described above.

Assumptions

The test methods described in this paper are limited by very few assumptions. Those mentioned by Kruskal and Wallis (1952) are:

1. All observations are randomly selected
2. All observations in a particular sample are selected from one population
3. For the various populations sampled, the distributions of the attribute being measured are "approximately the same" in form or dispersion.

The last assumption is included only because conventional tests of hypothesis are structured in terms of population means (e.g., H_o: $\mu_1 = \mu_2$ vs H_A: $\mu_1 \neq \mu_2$). Tests of this nature are rather meaningless unless such an assumption is true. Thus, assumption (3) is a requirement of the hypothesis, not the test method. Comparisons of population abundance in terms of gill net catches belong to this class of tests because the level of biological significance is normally conceived of in terms of some

minimum difference in mean catch. This means that, in order for such comparisons to be valid, the sample distributions of gill net catches should be reasonably similar in form. If there is any doubt about their similarity, the distributions should at least be plotted for visual comparison.

U and H can also be used to test a more general class of hypothesis, where interest may lie in any kind of population difference involving central tendency or dispersion. Under these circumstances, assumption (3) would not be included.

Many text-book writers mention an additional requirement of these tests, viz., continuous population distributions. Kruskal (1952), however, has shown that continuity is not necessary. Thus, U and H may be applied to gill net catch data without any additional qualifications required by their discrete nature.

Unusual Features

Although the ranking tests discussed in this paper are designed, and admirably suited, to discriminate between samples drawn from populations with different central tendencies, it is a remarkable fact that measures of central tendency (e.g., mean, median, mode) play no role in the conduct of these tests. Further, the individual sample statistics that are used (R_i and U_i) obviously cannot be calculated without reference to some other sample or samples; that is, they are not unique

functions of the sample they represent, independent of any other sample; which is the case with measures of central tendency. There are two consequences of these facts that have important implications for users who are accustomed to parametric statistical methods.

Interpretation of Results

Since no measure of central tendency, or any other unique sample statistic, is generated in the process of conducting a ranking test, it may be difficult to precisely and explicitly explain how samples do or do not differ as a result of applying one of these tests. Although conventional measures of central tendency may be used as descriptive statistics in the analysis of the data, these cannot always be relied on to explain the results of a ranking test. For example, as we shall see in a later section, it is quite within the realm of possibility to find significant differences between samples whose means and/or medians are identical. Contrariwise, we can also find samples between which the ranking tests are incapable of discriminating in spite of astronomical differences between these statistics. Most amazing of all, cases can be found where measures of central tendency imply that Population A is more dense than Population B, say, whereas the ranking tests indicate the opposite. All this does not mean that ranking tests are unreliable or inefficient, but simply that assumption (3) of the previous section, viz.,

similar population distributions, has not been satisfied. Since these tests are also capable of discriminating on the basis of dispersion differences, independent of central tendency differences, such results may occasionally occur in the analysis of gill net catch data. Unless the investigator is willing to admit population differences in terms other than of mean (or median) catches, he will be unable to rely on these results.

A more conventional problem in interpretation involves the case where $s \geq 3$. A significant test result in this situation means that some samples differ from other samples. However, it does not indicate where the differences lie. Such determination is left to the judgment of the investigator. In this respect ranking tests are no different from other multiple-comparison tests such as F, χ^2, or the median test that Moyle and Lound recommend.

Determination of Sample Size

Investigators need to know how large a sample to select in order to detect a given population difference with a predetermined degree of confidence. Since ranking tests do not operate on the principle of differences between measures of central tendency, we cannot develop a sample size formula in terms of such differences (as we can for the *t* test, for example).

On the other hand it can be shown that, under the proper circumstances, the following relation is satisfactorily close:

$$H_\varepsilon/H \simeq n_\varepsilon/n = \varepsilon, \text{ say} \qquad (3)$$

where

H = test statistic obtained with a sample of size n

H_ε = test statistic obtained with a sample of size n_ε.

Thus, if values for H, H_ε, and n are known, we can determine n_ε as

$$n_\varepsilon = \varepsilon n. \qquad (4)$$

The technique required consists of three steps as follows:

(1) select samples from the populations involved, using sizes that appear reasonable (normally, $10 \leq n_i \leq 20$ should be reasonable).

(2) calculate the test statistic H and determine if it is significant at the chosen probability level, say α; if it is, the original sample size was sufficient.

(3) if H is not significant, calculate the required sample size expansion factor ε from relation (3), where H_ε is replaced by the chi-square value needed for statistical significance at the chosen probability level; then calculate n_ε from relation (4). Since n units have already been selected, only $n_\varepsilon - n$ additional units need be drawn.

Ideally, any additional units that must be selected should be spread equally among the various populations being compared. For a population that is being studied over time, this obviously

is impossible, and all additional units must be selected from the current version of the population.

The conditions under which this sample-size procedure is valid are simply that H_A is true and the sample data are representative of the populations from which they have been selected. Obviously, it cannot work if H_0 is true. In that case, regardless of how large a sample is selected, we will reject H_0 only with probability α. Consequently, Step (3) above should be resorted to only if there is good reason to believe that there is, in fact, a biologically important difference between the populations involved.

Illustrative Examples

The purpose of this section is to use examples to illustrate various ideas presented in prior discussion. These examples are all in terms of gill net catches, either real or hypothetical, and demonstrate the recommended method of tabulating and summarizing data. The origin of the data is shown at the top of each table. Unless otherwise stated, the following conditions are assumed for each example:

H_0: populations do not differ in density

H_A: populations do differ in density

Significance level: 20 percent.

Example I. (Significance level = 5 percent)

Source: Moyle and Lound (1960)

	Sample 1		Sample 2	
	Catch	Rank	Catch	Rank
	0	1	2	4
	0	2	4	8
	1	3	4	9
	2	5[a]	5	11
	3	6	6	12
	3	7	6	13
	4	10[a]	8	14
			9	15
			11	16
TOTAL	13	34	55	102
MEAN	1.86	4.86	6.11	11.33
MEDIAN	2		6	

[a] Assigned Randomly

$s = 2$ $\quad n = 16$ $\quad U_1 = 34 - 7(8)/2 = 6$

$n_1 = 7$ $\quad R_1 = 34$ $\quad U_2 = 102 - 9(10)/2 = 57$

$n_2 = 9$ $\quad R_2 = 102$ $\quad U = \text{Min}(U_1, U_2) = 6$

A. As a check on the calculations, we note that:

1. The rank associated with the highest catch in the sample (11) is 16 = n
2. $R_1 + R_2 = 34 + 102 = 136 = 16(17)/2 = n(n+1)/2$
3. $U_1 + U_2 = 6 + 57 = 63 = 7(9) = n_1 n_2$

B. Since both n_i are less than 10, the U test is preferable to the H test. Using Table 5 we find that, for $n_1 = 7$ and $n_2 = 9$, a value of U = 12 is significant at the 5 percent level for a two-tailed test. Since our U is only 6, this means that we easily reject the null hypothesis. Clearly, U = 6 is significant at a much lower level than 5 percent.

C. Although the H test tends to be conservative in this case, it is instructive to see how it performs. From equation (2) we calculate

$$H = \frac{12}{n(n+1)} \sum_i^2 R_i^2/n_i - 3(n+1)$$

$$= \frac{12}{16(17)} \{(34)^2/7 + (102)^2/9\} - 3(17)$$

$$= 7.29, \text{ 1 df}$$

For the same degrees of freedom, $\chi^2_{.01} = 6.63$ and $\chi^2_{.005} = 7.88$. Since our H value is a little larger than the half-way point between these two values, we may conclude that the associated probability is somewhat less than 0.0075. Thus, even with the H test, we have little difficulty rejecting H_0.

D. When Moyle and Lound (1960) used these same data to demonstrate the median test method of comparing samples of gill net catches, they reported that the probability associated with a one-tailed test is P = 0.0104 (correct value is P = 0.0105). Doubling this value for the comparable two-tailed test, we find P = 0.021. Since this is considerably higher than the probability associated with the H test, the greater power of the latter, even under less than ideal conditions, is evident.

Example II.

Source: Lake Mary Ronan kokanee (Oncorhynchus nerka) catches

	Sample 1		Sample 2	
	Catch	Rank	Catch	Rank
	0	1	0	3[a]
	0	2	0	8[a]
	0	4	7	11
	0	5	9	14
	0	6	42	15
	0	7		
	0	9		
	6	10		
	8	12		
	9	13[a]		
TOTAL	23	69	58	51
MEAN	2.3	6.9	11.6	10.2
MEDIAN	0		7	

[a] Assigned Randomly

$s = 2$ $n = 15$ $U_1 = 69 - 10(11)/2 = 14$

$n_1 = 10$ $R_1 = 69$ $U_2 = 51 - 5(6)/2 = 36$

$n_2 = 5$ $R_2 = 51$ $U = 14$

A. Computational checks:

1. The rank of the highest catch (42) is 15 = n

2. $R_1 + R_2 = 120 = n(n+1)/2$

3. $U_1 + U_2 = 50 = n_1 n_2$

B. Table 3 shows that for samples of size $n_2 = 10$, $n_1 = 5$ (equivalent to $n_1 = 10$, $n_2 = 5$), U cannot exceed 14 for significance at the 20 percent level. Since calculated U equals 14, we may reject the null hypothesis.

C. Since $n_1 = 10$, it is also permissable to use the H test:

$$H = \frac{12}{15(16)} \{(69)^2/10 + (51)^2/5\} - 3(16)$$

$$= 1.815, 1 \text{ df}$$

$\chi^2_{0.2} = 1.64$, so the null hypothesis may be rejected, consistent with our result in B.

D. If the median test is used to compare these two samples, the two-tailed probability would be found to be P = 0.57, again demonstrating the superior power of the ranking tests.

E. There is a total of $\binom{9}{2}\binom{2}{1} = 72$ different ways that ranks 1-9 and 13-14 can be randomly assigned to the tied observations in this example. Of this total, 28 (39 percent) result in $U \leq 14$ ($R_1 \leq 69$); i.e., favor rejection of the null hypothesis at the 20 percent probability level. Thus, before the ranks were actually assigned in this example, the probability was less than 2 in 5 that we would select them in such a way that the null hypothesis would be rejected. In spite of these odds, however, that is precisely what we did.

If we had chosen our level of significance at 10 percent instead of 20 percent, there would have been only 10 chances in 72 (less than 14 percent) of randomly choosing ranks such that the null hypothesis could be rejected ($U \leq 11$). This demonstrates the role that chance plays in sampling experiments. In conventional situations chance enters into the sample selection phase only. In our case it plays a role in both the sample selection and rank assignment phases.

Example III. (Significance level = 10 percent)

Source: Lake Mary Ronan pumpkinseed (*Lepomis gibbosus*) catches

	Sample 1		Sample 2	
	Catch	Rank	Catch	Rank
	0	1	1	5
	0	2	23	8
	0	3	82	13
	0	4	105	14
	4	6	147	15
	4	7		
	24	9		
	35	10		
	42	11		
	75	12		
TOTAL	184	65	358	55
MEAN	18.4	6.5	71.6	11
MEDIAN	4		82	

$s = 2$	$n = 15$	$U_1 = 65 - 10(11)/2 = 10$
$n_1 = 10$	$R_1 = 65$	$U_2 = 55 - 5(6)/2 = 40$
$n_2 = 5$	$R_2 = 55$	$U = 10$

A. Computational checks:

1. The rank of the highest catch (147) is 15 = n

2. $R_1 + R_2 = 120 = n(n+1)/2$

3. $U_1 + U_2 = 50 = n_1 n_2$

B. Table 4 shows that for samples of size $n_2 = 10$, $n_1 = 5$, U cannot exceed 11 for significance at the 10 percent level. Since calculated U equals 10, we may reject the null hypothesis.

C. Since $n_1 = 10$, we can use the H test:

$$H = \frac{12}{15(16)} \{(65)^2/19 + (55)^2/5\} - 3(16)$$

$$= 3.375, 1 \text{ df}$$

$\chi^2_{0.1} = 2.71$, so the null hypothesis is rejected, consistent with our result in B.

Example IV.

Source: Lake Mary Ronan pumpkinseed catches

	Sample 1		Sample 2[a]	
	Catch	rank	Catch	rank
	0	1	1	2
	7	3	23	4
	30	5	82	8
	32	6	105	9
	48	7	147	10
TOTAL	117	22	358	33
MEAN	23.4	4.4	71.6	6.6
MEDIAN	30		82	

[a] Same as Sample 2 data in Example III.

$s = 2$ $n = 10$ $U_1 = 22 - 5(6)/2 = 7$

$n_1 = 5$ $R_1 = 22$ $U_2 = 33 - 5(6)/2 = 18$

$n_2 = 5$ $R_2 = 33$ $U = 7$

A. Computational checks:

1. The rank of the highest catch (147) is 10 = n
2. $R_1 + R_2 = 55 = n(n+1)/2$
3. $U_1 + U_2 = 25 = n_1 n_2$

B. Using Table 3 for samples of size $n_1 = n_2 = 5$, we find that U cannot exceed 5 for significance at the 20 percent level. Since calculated U equals 7, we accept the null hypothesis.

C. This example demonstrates how the R_i value for a given sample changes relative to the sample with which it is being compared.

In this example, the Sample 2 data have an R_i value of 33 (relative to the Sample 1 data), whereas in Example III the same data have an R_i value of 55 (relative to the Sample 1 data).

D. It is interesting to note that, in spite of the large differences between the two sample means and between the two sample medians, the ranking test was unable to detect a significant difference between the samples. Assuming that we are certain of a real difference between the populations sampled, we can use the procedure for determining a sample size that is large enough to provide statistical significance. Since this requires calculating H, which normally would not be used for comparing samples this small, we realize that our calculations may not be entirely accurate. If there is any doubt about this, however, we can always inflate the figure somewhat to be on the safe side.

Using relation (2),

$$H = \frac{12}{10(11)} \{(33)^2/5 + (22)^2/5\} - 3(11)$$

$$= 1.32$$

Using relation (3),

$$\varepsilon = H_\varepsilon/H = 1.64/1.32 = 1.24$$

where $H_\varepsilon = 1.64$ represents the value that delimits the upper 20 percent of the 1-df chi-square distribution.

Finally, using relation (4),

$$n_\varepsilon = \varepsilon n = 1.24(10) = 12.4 \text{ or } \underline{13}.$$

This says that, with a sample size of 13, selected under the same conditions as the original sample of 10, we should be able to show a statistically significant difference between the individual population samples at the 20 percent probability level. Since the initial sample already contains 10 units, it is only necessary to select an additional 3 to complete the requirement.

Example V.

Source: Middle Bowman Lake rainbow (Salmo gairdneri) catches

	Sample 1		Sample 2		Sample 3		Sample 4		Sample 5	
	Catch	Rank	Catch	Rank	Catch	Rank	Catch	Rank	Catch	Rank
	0	4[a]	0	10[a]	0	1[a]	1	12	0	2
	0	5[a]	1	14[a]	0	8[a]	1	13	0	3
	0	11[a]	2	20	0	9[a]	1	16	0	6
	1	15[a]	2	21	1	18[a]	3	25[a]	0	7
	1	17[a]	6	30	1	19[a]	3	27[a]	2	23[a]
	2	22[a]								
	3	24								
	3	26								
	3	28								
	4	29								
TOTAL	17	181	11	95	2	55	9	93	2	41
MEAN	1.7	18.1	2.2	19	0.4	11	1.8	18.6	0.4	8.2
MEDIAN	1.5		2		0		1		0	

[a] Assigned randomly

$s = 5$ $n = 30$

$n_1 = 10$ $R_1 = 181$

$n_2 = 5$ $R_2 = 95$

$n_3 = 5$ $R_3 = 55$

$n_4 = 5$ $R_4 = 93$

$n_5 = 5$ $R_5 = 41$

A. Computational checks:

1. The rank of the highest catch (6) is 30 = n
2. $\sum_{i}^{5} R_i = 465 = n(n+1)/2$

B. Since s > 3, and all $n_i \geq 5$, we can use the H test in conjunction with the tabulated chi-square distribution:

$$H = \frac{12}{30(31)} \{(181)^2/10 + (95)^2/5 + (55)^2/5 + (93)^2/5 + (41)^2/5\} - 3(31)$$

$$= 7.03, \text{ 4 df}$$

$\chi^2_{0.2} = 5.99$, so the null hypothesis is rejected.

C. Since this is a multiple-comparison test, we cannot be certain where the sample differences occur. Looking at the five mean ranks, however, we may reasonably speculate that Samples 1, 2, and 4 differ from Samples 3 and 5. Hopefully, this division of the data would be compatible with other information about the five populations.

Example VI.

Source: Random number table

	Sample 1		Sample 2	
	Catch	Rank	Catch	Rank
	11	1	13	2
	27	5	18	3
	37	7	25	4
	43	8	36	6
	44	9	55	10
	59	11	66	14
	63	12	77	15
	64	13	94	18
	85	16	96	19
	88	17	99	20
TOTAL	521	99	579	111
MEAN	52.1	9.9	57.9	11.1
MEDIAN	51.5		60.5	

$s = 2$ $n = 20$ $U_1 = 99 - 10(11)/2 = 44$

$n_1 = 10$ $R_1 = 99$ $U_2 = 111 - 10(11)/2 = 56$

$n_2 = 10$ $R_2 = 111$ $U = 44$

A. Computational checks:

1. The rank of the highest catch (99) is 20 = n

2. $R_1 + R_2 = 210 = n(n+1)/2$

3. $U_1 + U_2 = 100 = n_1 n_2$

B. Using Table 3, we find that for sample sizes of $n_i = n_2 = 10$, U cannot exceed 33 for significance at the 20 percent probability

level. Since calculated U equals 44, we accept the null hypothesis.

C. This example provides a measure of the amount of sample variability to expect when sampling from a uniform distribution (all values of the attribute occur with equal frequency in the population). Since both samples are drawn from the same population, we expect our 20 percent test to indicate no significant difference four times out of five, on the average. The two samples actually drawn meet our expectation in this regard.

Example VII.

Source: Hypothetical

	Sample 1		Sample 2	
	Catch	Rank	Catch	Rank
	0	1	2	7
	0	2	2	8
	0	3	2	9
	0	4	2	10
	0	5	2	11
	1	6	2	12
	3	14	2	13
	3	15	4	19
	3	16	5	20
	3	17	6	21
	3	18	7	22
	32	24	12	23
TOTAL	48	125	48	175
MEAN	4	10.42	4	14.58
MEDIAN	2		2	

$s = 2$ $n = 24$ $U_1 = 125 - 12(13)/2 = 47$

$n_1 = 12$ $R_1 = 125$ $U_2 = 175 - 12(13)/2 = 97$

$n_2 = 12$ $R_2 = 175$ $U = 47$

A. Computational checks:

1. The rank of the highest catch (32) is 24 = n
2. $R_1 + R_2 = 300 = n(n+1)/2$
3. $U_1 + U_2 = 144 = n_1 n_2$

B. Using Table 3, we find that for samples of size $n_1 = n_2 = 12$, U cannot exceed 49 for significance at the 20 percent level. Since calculated U equals 47, we reject the null hypothesis.

C. The interesting point of this example is that both the mean catches and median catches of the two samples are identically the same (4 fish/net set and 2 fish/net set, respectively), yet our ranking test indicates that there is a significant difference (somewhere) between the two. This shows that ranking tests are capable of discriminating in terms of differences other than in central tendency. As we learned in a previous section, these include differences in dispersion. Looking at the distribution of catches in the two samples, we can see that there is a fairly obvious difference in this regard. One possible interpretation of this difference is that Population-1 fish tend to congregate more than Population-2 fish. Thus, a larger percentage of their habitat is devoid of fish, but they are more abundant in the areas in which they occur. If the investigator is interested only in the total number (or density) of fish in each population, then he would have to reject the results of

the statistical test (more appropriately, he would not even make the test because the mean and/or median catches tell him there is no need to). If, on the other hand, he is also interested in other kinds of differences between the populations (e.g., behavioral), then he would conclude from his statistical test that there is indeed a difference.

Example VIII.

Source: Hypothetical

	Sample 1		Sample 2	
	Catch	Rank	Catch	Rank
	0	1	2	7
	0	2	2	8
	0	3	2	9
	0	4	2	10
	0	5	2	11
	1	6	2	12
	3	14	2	13
	3	15	4	19
	3	16	5	20
	3	17	6	21
	3	18	7	22
	80	24	12	23
TOTAL	96	125	48	175
MEAN	8	10.42	4	14.58
MEDIAN	2		2	

$s = 2$ $\quad n = 24$ $\quad U_1 = 125 - 12(13)/2 = 47$

$n_1 = 12$ $\quad R_1 = 125$ $\quad U_2 = 175 - 12(13)/2 = 97$

$n_2 = 12$ $\quad R_2 = 175$ $\quad U = 47$

A. Computational checks:

1. The rank of the highest catch (80) is 24 = n
2. $R_1 + R_2 = 300 = n(n+1)/2$
3. $U_1 + U_2 = 144 = n_1 n_2$

B. Since calculated U equals 47 compared to tabulated U of 49 (Table 3), we reject the null hypothesis.

C. This example is identical to the previous one, except that the largest catch in Sample 1 is now 80 instead of 32, resulting in a 2-fold difference between the sample means. There are two observations to make about this case. First, in spite of the change in the one catch (and its effect on the mean catch), there is no change in the ranking test: U = 47 in both cases. Clearly, this is because the change in catch does not cause a change in the individual ranks. Furthermore, increasing the mean catch in the same way by a factor of 10, 100, etc., will have no effect either. This emphasizes the concept that means of sample observations are not necessarily relevant statistics in the context of ranking tests.

Second, and perhaps more important, there is a very real danger of misinterpreting the results of this test. For the average investigator who is looking only for a difference in population abundance, it would be extremely easy to conclude that

Population 1 is more abundant than Population 2. After all, isn't this justified by both the 2-fold difference in mean catch plus the significant U value? Unfortunately, it is not. If one takes the trouble to look at the mean ranks, he will see that $\bar{R}_1$ is smaller than $\bar{R}_2$. Thus, the U test cannot possibly be telling us that Population 1 is more abundant than Population 2. What it is telling us is exactly the same thing it told us in Example VII, viz., that there is a dispersion difference between the catch distributions. It just so happens that there appears to be (and may in fact be) an abundance difference, but this has no bearing on the statistical test.

Summary

On the basis of the results presented in this paper, the following recommendations are proposed:

1. Nonparametric ranking tests offer a statistically valid and efficient way of analyzing gill net catch data.
2. For samples that are large enough for the normal approximation to apply, the Kruskal-Wallis H test is appropriate for any comparison involving two or more samples.
3. For two-sample cases where the samples are not large enough for the normal approximation to apply, or in those cases where the investigator prefers, tables of the U statistic may be used as an alternative to the H test.

4. For three-sample cases where the samples are not large enough, Siegel's Table O of the H statistic may be used as an alternative to the chi-square approximation of the H test.
5. In order to satisfy all assumptions of the method, in addition to maintaining the broadest possible scope for statistical inference, the ranks of tied observations should be assigned randomly whenever they involve two or more samples.
6. Since ranking tests are capable of discriminating between samples on the basis of differences in central tendency as well as differences in dispersion, the investigator should be alert for instances of the latter; otherwise, misinterpretations of data analyses may result. As an aid to avoiding such problems, he should make it a habit to do two things. First, prior to sample selection, he should choose a minimum difference in mean catch that he considers biologically significant. If the difference actually found falls below this value, then no statistical test would be made. Second, assuming that a statistical test is called for, he should rank the set of mean ranks and the set of mean catches. If the two sets do not fall in the same order, then again no test should be made.

Acknowledgment

I wish to thank Mr. George D. Holton, Montana Department of Fish and Game, who provided the opportunity for this undertaking.

References

1. Alder, H.L., and E.B. Roessler. 1964. Introduction to probability and statistics. W.H. Freeman, San Francisco. 313 pp.
2. Hodgman, C.D.(Ed.). 1959. C.R.C. Standard mathematical tables. Chem. Rub. Pub. Co., Cleveland. 525 pp.
3. Kruskal, W.H. 1952. A nonparametric test for the several sample problem. Ann. Math. Stat. 23:525-540.
4. Kruskal, W.H., and W.A. Wallis. 1952. Use of ranks in one-criterion variance analysis. J. Amer. Stat. Assoc. 47:583-621.
5. Moyle, J.B. 1950. Gill nets for sampling fish populations in Minnesota waters. Trans. Am. Fish. Soc. (1949):195-204.
6. Moyle, J.B., and R. Lound. 1960. Confidence limits associated with means and medians of series of net catches. Trans. Am. Fish. Soc. 89(1):53-58.
7. Siegel, S. 1956. Nonparametric statistics. McGraw-Hill, New York. 312 pp.

Table 1. Sample-1 rankings as low as or lower than that actually obtained.

Rank Order Position																Sum of
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Ranks
X	X	X	X	X	X	X										28
X	X	X	X	X	X		X									29
X	X	X	X	X	X			X								30
X	X	X	X	X	X				X							31
X	X	X	X	X	X					X						32
X	X	X	X	X		X	X									30
X	X	X	X	X		X		X								31
X	X	X	X	X		X			X							32
X	X	X	X	X			X	X								32
X	X	X	X		X	X	X									31
X	X	X	X		X	X		X								32
X	X	X		X	X	X	X									32

Table 2. High sample-1 rankings as extreme as or more extreme than that actually obtained.

Rank Order Position																Sum of
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Ranks
									X	X	X	X	X	X	X	91
								X		X	X	X	X	X	X	90
							X			X	X	X	X	X	X	89
						X				X	X	X	X	X	X	88
					X					X	X	X	X	X	X	87
								X	X		X	X	X	X	X	89
							X		X		X	X	X	X	X	88
						X			X		X	X	X	X	X	87
							X	X			X	X	X	X	X	87
								X	X	X		X	X	X	X	88
							X		X	X		X	X	X	X	87
								X	X	X	X		X	X	X	87

Table 3. 20% Probability values of U for a two-tailed test[a,b]

n_1 \ n_2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	-	-	-	-	-	-	0	0	1	1	1	1	1	1	2	2	2	2
2	0	0	1	1	1	2	3	4	4	5	5	6	6	7	7	8	8	8
3	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16
4		3	4	5	6	7	9	11	12	13	14	16	17	18	19	21	22	23
5			5	7	9	10	13	14	16	18	19	21	23	24	26	27	29	31
6				9	11	13	16	18	20	22	24	26	28	30	32	34	37	39
7					13	16	19	22	24	27	29	32	34	37	39	42	44	47
8						19	22	25	28	31	34	37	40	43	46	49	52	54
9							26	29	32	36	39	42	46	49	52	56	59	63
10								33	36	40	44	48	52	55	59	63	67	71
11									41	45	49	53	57	62	66	70	74	79
12										49	54	59	63	68	73	77	82	87
13											59	64	69	74	80	85	90	95
14												70	75	81	86	92	98	103
15													81	87	93	99	105	111
16														94	100	107	113	119
17															107	114	121	128
18																121	129	136
19																	136	144
20																		152

[a]Each entry corresponds to the value of U that delimits the lower 10% of the distribution of U. For one-tailed tests, the entries represent 10% probability values.

[b] Since the table is symmetric about the main diagonal $\{U(n_1,n_2) = U(n_2,n_1)\}$, those entries below the main diagonal are omitted.

Table 4. 10% probability values of U for a two-tailed test[a,b]

n_1 \ n_2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
2	-	-	0	0	0	1	1	1	1	2	2	2	3	3	3	4	4	4
3	0	0	1	2	2	3	3	4	5	5	6	7	7	8	9	9	10	11
4		1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5			4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6				7	9	10	12	14	16	17	19	21	23	25	26	28	30	32
7					11	13	15	17	19	21	24	26	28	30	33	35	37	39
8						16	18	20	23	26	28	31	33	36	39	41	44	47
9							21	24	27	30	33	36	39	42	45	48	51	54
10								27	31	34	37	41	44	48	51	55	58	62
11									34	38	42	46	50	54	57	61	65	69
12										42	47	51	55	60	64	68	72	77
13											51	56	61	65	70	75	80	84
14												61	66	71	77	82	87	92
15													72	77	83	88	94	100
16														83	89	95	101	107
17															96	102	109	115
18																109	116	123
19																	123	130
20																		138

[a] Each entry corresponds to the value of U that delimits the lower 5% of the distribution of U. For one-tailed tests, the entries represent 5% probability values.

[b] Since the table is symmetric about the main diagonal $\{U(n_1,n_2) = U(n_2,n_1)\}$, those entries below the main diagonal are omitted.

Table 5. 5% probability values of U for a two-tailed test[a,b]

n_1 \ n_2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	-	-	-	-	-	0	0	0	0	1	1	1	1	1	2	2	2	2
3	-	-	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4		1	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5			2	4	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6				5	7	8	10	11	13	14	16	17	19	21	22	24	25	27
7					8	10	12	14	16	18	20	22	24	26	28	30	32	34
8						13	15	17	19	22	24	26	29	31	34	36	38	41
9							17	20	23	26	28	31	34	37	39	42	45	48
10								23	26	29	33	36	39	42	45	48	52	55
11									30	33	37	40	44	47	51	55	58	62
12										37	41	45	49	53	57	61	65	69
13											45	50	54	59	63	67	72	76
14												55	59	64	67	74	78	83
15													64	70	75	80	85	90
16														75	81	86	92	98
17															87	93	99	105
18																99	106	112
19																	113	119
20																		127

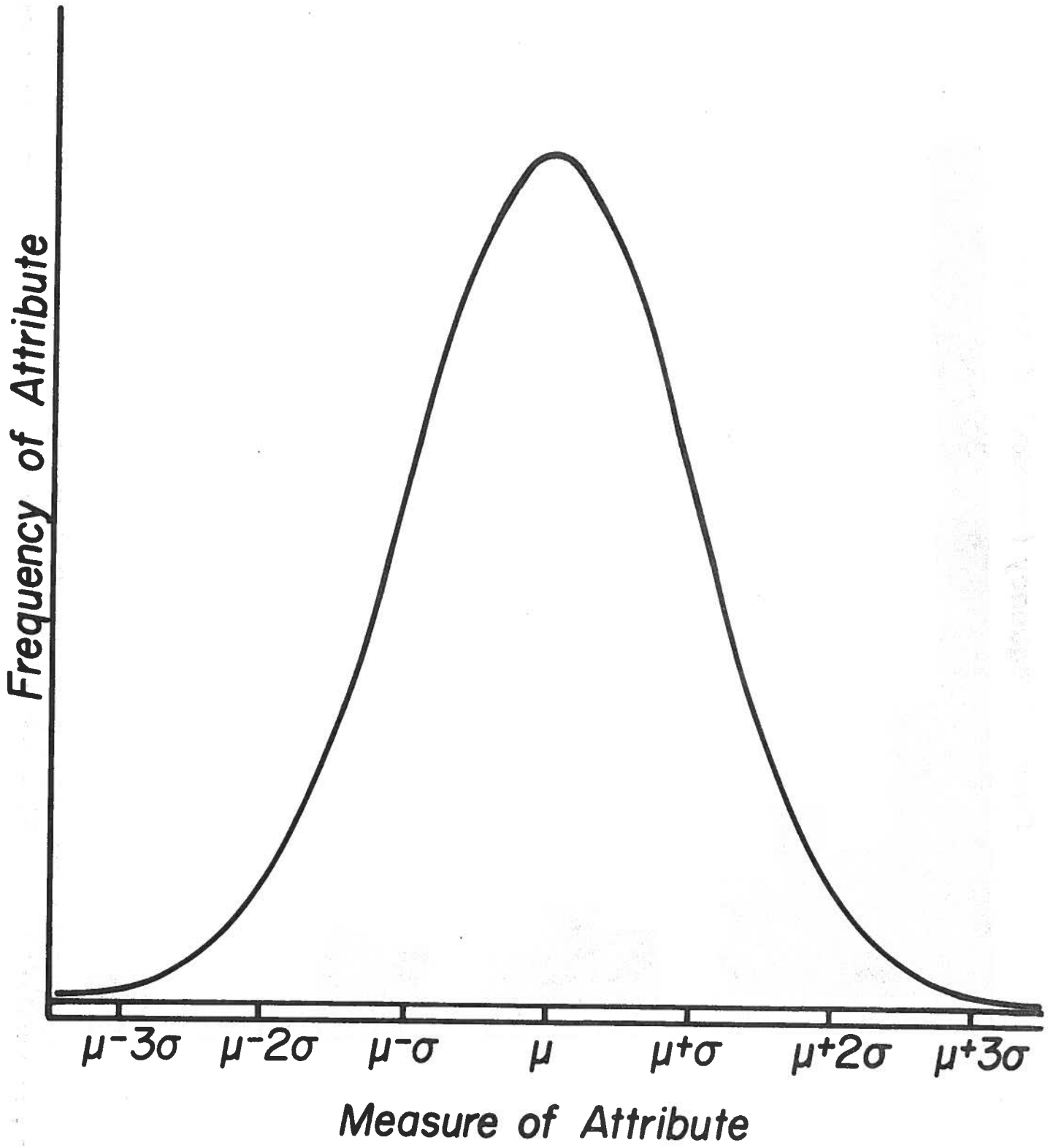
[a] Each entry corresponds to the value of U that delimits the lower 2.5% of the distribution of U. For one-tailed tests, the entries represent 2.5% probability values.

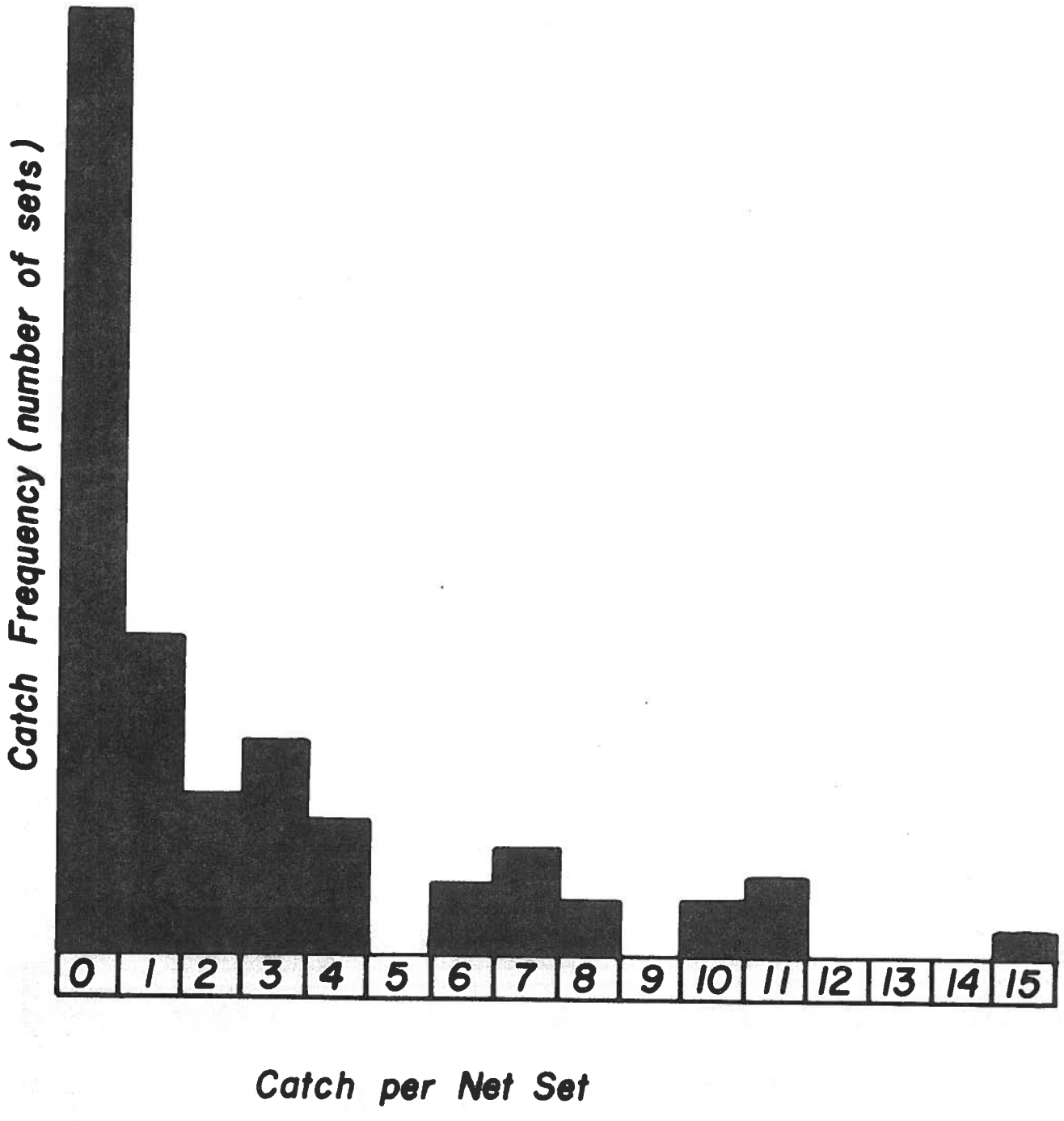
[b] Since the table is symmetric about the main diagonal $\{U(n_1,n_2) = U(n_2,n_1)\}$, those entries below the main diagonal are omitted.

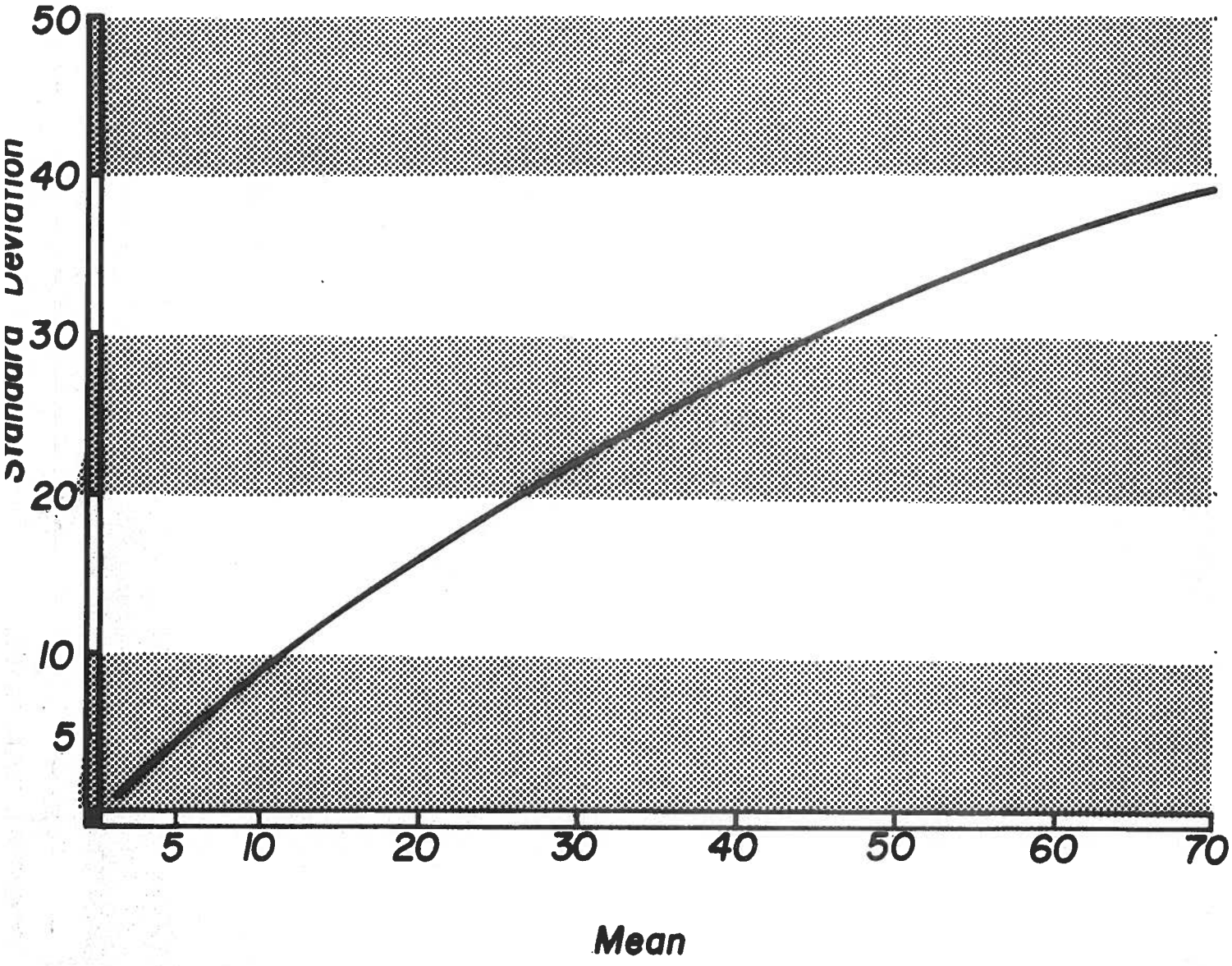
Figure 1. Graph of normal frequency function with mean μ and standard deviation σ.

Figure 2. Typical frequency distribution of gill net catches.

Figure 3. Typical relationship between mean and standard deviation of gill net catches. (Data after Moyle, 1950).


Frequency of Attribute
μ-3σ
μ-2σ
μ-σ
μ
μ+σ
μ+2σ
μ+3σ
Measure of Attribute


Catch Frequency (number of sets)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Catch per Net Set


50
40
30
20
10
5
Standard Deviation
5
10
20
30
40
50
60
70
Mean