# NONPARAMETRIC STATISTICS

(adapted from J. Hurley notes)

Nonparametric statistics are useful when inferences must be made on categorical or ordinal data, when the assumption of normality is not appropriate, or when the sample sizes are small.

Advantages:

1. Easy application (doesn't even need a calculator in many cases).
2. Can serve as a quick check to determine whether or not further analysis is required.
3. Many assumptions concerning the population of the data source can be relaxed.
4. Can be used to test categorical (yes/no) data.
5. Can be used to test ordinal (1, 2, 3) data.

The primary disadvantage of non-parametric methods is that they lack the *power* of parametric methods. That means they produce conclusions that have a higher probability of being incorrect.

The Sign Test:

When dealing with a small sample of size $n$ that is not normally distributed, the Sign Test should be used instead of the z-test or t-test for the mean.

The Sign Test is used to test hypotheses about the median of *any* continuous distribution. Since the median of a symmetrical distribution is equal to the mean, the Sign Test can also be used to test hypotheses about the mean if the underlying distribution is known to be symmetrical.

One-tailed Tests:

$H_0$: $\mu = \mu_0$
$H_a$: $\mu < \mu_0$
*TS:* $R^+$ = Number of observations greater than $\mu_0$.
$p$-value = $P(x \le R^+)$

$H_0$: $\mu = \mu_0$
$H_a$: $\mu > \mu_0$
*TS:* $R^-$ = Number of observations less than $\mu_0$.
$p$-value = $P(x \le R^-)$

Two-tailed Test:

$H_0$: $\mu = \mu_0$
$H_a$: $\mu \ne \mu_0$
*TS:* $R = \min(R^+, R^-)$
$p$-value = $2 \times P(x \le R)$

where $x$ has a binomial distribution with parameters $n$ and $p = ½$ (because, if the median really is $\mu_0$, half of the observations should be less than $\mu_0$ and half will be more than $\mu_0$).

Alternatively, you can use Table X (below) which gives critical values of $R$ for the two-tailed Sign Test based on the number of observations $n$ and the significance level $\alpha$. The critical values are the largest values of $R$ for which $P(x \le R) \le \alpha$. If the calculated test statistic is less than or equal to the critical value, the null hypothesis is rejected. (Note that Table X is for the two-tailed test; the significance level $\alpha$ for one-tailed tests is one-half the value shown because there is only one rejection region instead of two rejection regions.)

TABLE X **Critical Values for the Sign Test**[a]

$R_\alpha^*$

| $n$ \ $\alpha$ | .10 | .05 | .01 |
|---|---|---|---|
| 5 | 0 | | |
| 6 | 0 | 0 | |
| 7 | 0 | 0 | |
| 8 | 1 | 0 | 0 |
| 9 | 1 | 1 | 0 |
| 10 | 1 | 1 | 0 |
| 11 | 2 | 1 | 0 |
| 12 | 2 | 2 | 1 |
| 13 | 3 | 2 | 1 |
| 14 | 3 | 2 | 1 |
| 15 | 3 | 3 | 2 |
| 16 | 4 | 3 | 2 |
| 17 | 4 | 4 | 2 |
| 18 | 5 | 4 | 3 |
| 19 | 5 | 4 | 3 |
| 20 | 5 | 5 | 3 |
| 21 | 6 | 5 | 4 |
| 22 | 6 | 5 | 4 |
| 23 | 7 | 6 | 4 |
| 24 | 7 | 6 | 5 |
| 25 | 7 | 7 | 5 |
| 26 | 8 | 7 | 6 |
| 27 | 8 | 7 | 6 |
| 28 | 9 | 8 | 6 |
| 29 | 9 | 8 | 7 |
| 30 | 10 | 9 | 7 |
| 31 | 10 | 9 | 7 |
| 32 | 10 | 9 | 8 |
| 33 | 11 | 10 | 8 |
| 34 | 11 | 10 | 9 |
| 35 | 12 | 11 | 9 |
| 36 | 12 | 11 | 9 |
| 37 | 13 | 12 | 10 |
| 38 | 13 | 12 | 10 |
| 39 | 13 | 12 | 11 |
| 40 | 14 | 13 | 11 |

[a]For $n > 40$, $R$ is approximately normally distributed with mean $n/2$ and variance $n/4$.

As the note at the bottom of the table implies, if the sample size is large the binomial distribution can be approximated by a normal distribution with a mean of $n/2$ and a variance of $n/4$, so a test of $H_0$: $\mu = \mu_0$ can be based on the statistic

$$Z_0 = \frac{R - 0.5n}{0.5\sqrt{n}} = \frac{2R - n}{\sqrt{n}}$$

and an appropriate critical region can be determined from the cumulative distribution function of the standard normal distribution (Table C4 in the back of Ledolter and Hogg).

*Example – Two-Tailed Sign Test for a Population Median (small samples) Adapted from Hines & Montgomery (1990) p. 561.*

The local ready-mix concrete plant has developed a mix design for concrete with a 28-day compressive strength of 4000 psi. Last month, quality control technicians made 20 concrete cylinders from a batch of the concrete. Yesterday they obtained the following strengths:

| | | | | |
|---|---|---|---|---|
| 4318 | 4330 | 3416 | 4416 | 4106 |
| 3356 | 4800 | 3570 | 3530 | 4828 |
| 4632 | 3560 | 5150 | 4514 | 4402 |
| 4122 | 4674 | 4716 | 3508 | 5308 |

Given this data, test the hypothesis that the median strength of the concrete mix is 4000 psi.

*Example – One-Tailed Sign Test for a Population Median (small samples)*

Environmental engineers have found that the percentages of active bacteria in sewage specimens collected at a particular sewage treatment plant have a non-normal distribution with a median of 40% when the plant is running properly. If the median is larger than 40%, then adjustments must be made. The percentages of active bacteria in a random sample of 10 specimens are given below.

| | | | | |
|---|---|---|---|---|
| 41 | 33 | 43 | 52 | 46 |
| 37 | 44 | 49 | 53 | 30 |

Do the data provide sufficient evidence (at $\alpha = 0.05$) to indicate that adjustments are needed?

The Sign Test gets its name from that fact that $R^+$ and $R^-$ are actually the number of observations for which the difference $(x_i - \mu_0)$ has either a positive sign or a negative sign. With this in mind, we can also use the Sign Test to determine if there are differences in the medians of populations of *paired* data. In this case, we calculate the differences in value between the paired observations and count how many are positive and how many are negative. If there is no difference, there should be just as many positive differences as negative differences.

*Example – Sign Test for Paired Data (small samples) Adapted from A Blog on Probability and Statistics (http://probabilityandstats.wordpress.com/2010/02/27/the-sign-test/ )*

The latest trend in education is the use of standardized tests to measure whether or not "value" has been added by a teacher. A teacher gave the 15 students in her class one test at the start of the marking period and another test at the end. The scores are shown below. Is there sufficient evidence at $\alpha = 0.05$ to conclude that value has been added?

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Post-Test: | 21 | 26 | 19 | 26 | 30 | 40 | 43 | 15 | 29 | 31 | 46 | 8 | 43 | 31 | 37 |
| Pre-Test: | 17 | 26 | 16 | 28 | 23 | 35 | 41 | 18 | 30 | 29 | 45 | 7 | 38 | 31 | 36 |
| Difference: | +4 | 0 | +3 | –2 | +7 | +5 | +2 | –3 | –1 | +2 | +1 | +1 | +5 | 0 | +1 |

Out of the 15 students, two showed a test difference of zero. Since this is neither positive nor negative, we'll throw them out. Of the 13 remaining students, 10 had a positive difference and 3 had a negative difference. This suggests that value has been added, but is the evidence sufficient (at $\alpha = 0.05$) to conclude that?

NOTE: You must be careful not to reverse your differences! In this case, if the difference

Post-Test – Pre-Test

is positive, it suggests that value has been added, so we will only reject the null hypothesis if the number of *negative* differences ($R^-$) is small. Had we instead calculated

Pre-Test – Post-Test

then a negative difference would suggest that value had been added, so we would only reject the null hypothesis if the number of *positive* differences ($R^+$) were small.

We mentioned at the beginning of this section that nonparametric tests can also be used for categorical data. For these problems, we look directly at the ***proportions*** of the observations (i.e., what fraction fall into Category A as opposed to Category B). If there is no category preference, we would expect the split to be 50:50. Any significant deviation from that would be evidence that there really is a difference between the categories. So for these problems, we are testing the null hypothesis $H_0$: p = ½ against one of the alternatives

$H_a$: $p < ½$ (one-tailed)
$H_a$: $p > ½$ (one-tailed)
$H_a$: $p \neq ½$ (two-tailed)

*Example – Sign Test for Paired Categorical Data Adapted from A Blog on Probability and Statistics (http://probabilityandstats.wordpress.com/2010/02/27/the-sign-test/ )*

The math department at a local college polled its students to determine which of two professors (A or B) was most popular. They surveyed 15 students who had taken classes from both and found that 11 of the 15 preferred Professor B over Professor A. Is Professor B really more popular? Test at $\alpha = 0.05$.

Finally, we mentioned at the beginning of this section that nonparametric tests can also be used for ordinal data, which are numerical scores where the exact quantity has no significance beyond its ability to establish a ranking. Here again, we look directly at the proportions of the observations.

*Example – Sign Test for Paired Ordinal Data*

Fifteen judges were asked to rate leaf samples from two different varieties of tobacco on a scale from 1 to 5. Use the data shown below to test the hypothesis that one variety scores higher than the other. Use $\alpha = 0.05$ for your analysis.

| *Judge* | *Variety 1* | *Variety 2* | *Difference* |
|---|---|---|---|
| 1 | 1 | 2 | |
| 2 | 4 | 3 | |
| 3 | 4 | 3 | |
| 4 | 2 | 1 | |
| 5 | 4 | 3 | |
| 6 | 5 | 4 | |
| 7 | 5 | 3 | |
| 8 | 4 | 2 | |
| 9 | 5 | 3 | |
| 10 | 3 | 1 | |
| 11 | 4 | 4 | |
| 12 | 2 | 3 | |
| 13 | 4 | 2 | |
| 14 | 5 | 3 | |
| 15 | 4 | 3 | |

Wilcoxon's Signed Rank Test:

Assume that the population of interest is both continuous and *symmetric*, though not necessarily normal. In this case, the mean and the median are the same so hypothesis tests on the median are the same as hypothesis tests on the mean. A disadvantage of the sign test for these distributions is that it only considers the *signs* of the deviations and not their *magnitudes*. The Wilcoxon Signed Rank Test overcomes that disadvantage.

First proposed by Frank Wilcoxon ("Individual comparisons by ranking methods". *Biometrics Bulletin* 1 (6): 80–83, Dec. 1945), this test is performed by ranking the non-zero deviations in order of increasing magnitude (i.e., the smallest non-zero deviation has a rank of 1 and the largest deviation has a rank of *n*), then summing the *ranks* of those deviations with positive values and those with negative values. These sums are used to determine whether or not the deviations are significantly different from zero:

One-tailed Test:

$H_0$: $\mu = \mu_0$
$H_a$: $\mu < \mu_0$
*TS:* $T^+$ = sum of the positive ranks

One-tailed Test:

$H_0$: $\mu = \mu_0$
$H_a$: $\mu > \mu_0$
*TS:* $T^-$ = absolute value of the sum of the negative ranks

Two-tailed Test

$H_0$: $\mu = \mu_0$
$H_a$: $\mu \neq \mu_0$
*TS:* $T = \min(T^+, T^-)$

Because the underlying population is assumed to be continuous, ties are theoretically impossible, but in practice you can get ties, especially if the data has only a couple of significant digits. If two or more deviations have the same magnitude, they are all given the same average rank so as not to favor one over the other. So, for example, if there are 23 deviations with a magnitude smaller than 72 and the next two deviations are +72 and –72, they would both be assigned a rank of

$$(24 + 25)/2 = 24.5$$

For the same reason, deviations of zero are theoretically impossible but practically possible. Any deviations of exactly zero are simply thrown out and the value of *n* reduced accordingly.

If the sample size is small (which is open to interpretation—I've seen values from $n < 10$ to $n < 50$) you have to refer T to a table of critical values (Table XI below). These tables are developed by enumerating all possibilities for a given sample size $n$ and determining the largest value of T with a probability of occurrence of less than 1%, less than 5%, etc. (Note that the significance level for one-tailed tests is one-half the $\alpha$ value given in the table because there is only one rejection region instead of two rejection regions.)

TABLE XI **Critical Values for the Wilcoxon Signed-Rank Test**[a]

| $n$ \ $\alpha$ | .10 | .05 | .02 | .01 |
|---|---|---|---|---|
| 4 | | | | |
| 5 | 0 | | | |
| 6 | 2 | 0 | | |
| 7 | 3 | 2 | 0 | |
| 8 | 5 | 3 | 1 | 0 |
| 9 | 8 | 5 | 3 | 1 |
| 10 | 10 | 8 | 5 | 3 |
| 11 | 13 | 10 | 7 | 5 |
| 12 | 17 | 13 | 9 | 7 |
| 13 | 21 | 17 | 12 | 9 |
| 14 | 25 | 21 | 15 | 12 |
| 15 | 30 | 25 | 19 | 15 |
| 16 | 35 | 29 | 23 | 19 |
| 17 | 41 | 34 | 27 | 23 |
| 18 | 47 | 40 | 32 | 27 |
| 19 | 53 | 46 | 37 | 32 |
| 20 | 60 | 52 | 43 | 37 |
| 21 | 67 | 58 | 49 | 42 |
| 22 | 75 | 65 | 55 | 48 |
| 23 | 83 | 73 | 62 | 54 |
| 24 | 91 | 81 | 69 | 61 |
| 25 | 100 | 89 | 76 | 68 |
| 26 | 110 | 98 | 84 | 75 |
| 27 | 119 | 107 | 92 | 83 |
| 28 | 130 | 116 | 101 | 91 |
| 29 | 140 | 126 | 110 | 100 |
| 30 | 151 | 137 | 120 | 109 |
| 31 | 163 | 147 | 130 | 118 |
| 32 | 175 | 159 | 140 | 128 |
| 33 | 187 | 170 | 151 | 138 |
| 34 | 200 | 182 | 162 | 148 |
| 35 | 213 | 195 | 173 | 159 |
| 36 | 227 | 208 | 185 | 171 |
| 37 | 241 | 221 | 198 | 182 |
| 38 | 256 | 235 | 211 | 194 |
| 39 | 271 | 249 | 224 | 207 |
| 40 | 286 | 264 | 238 | 220 |
| 41 | 302 | 279 | 252 | 233 |
| 42 | 319 | 294 | 266 | 247 |
| 43 | 336 | 310 | 281 | 261 |
| 44 | 353 | 327 | 296 | 276 |
| 45 | 371 | 343 | 312 | 291 |
| 46 | 389 | 361 | 328 | 307 |
| 47 | 407 | 378 | 345 | 322 |
| 48 | 426 | 396 | 362 | 339 |
| 49 | 446 | 415 | 379 | 355 |
| 50 | 466 | 434 | 397 | 373 |

*Source*: Adapted with permission from "Extended Tables of the Wilcoxon Matched Pair Signed Rank Statistic" by Robert L. McCornack, *Journal of the American Statistical Association*, Vol. 60, September, 1965.

[a]If $n > 50$, $R$ is approximately normally distributed with mean $n(n+1)/4$ and variance $n(n+1)(2n+1)/24$.

For example, if $n = 3$, there are $2^3 = 8$ possible values for T. If the signed ranks are (+1,+2,+3) or (–1,–2,–3) then T = min(6,0) = 0; if they are (–1,+2,+3) or (+1,–2,–3) then T = min (5,1) = 1; if they are (–1,+2,–3) or (+1,–2,+3) then T = min(4,2) = 2, and if they are (–1,–2,+3) or (+1,+2,–3) then T = min(3,3) = 3. So the probability of getting a T value of 0 is 2/8 = 0.25; the probability of getting a T value of 1 or less is 4/8 = 0.50, a value of 2 or less is 6/8 = 0.75, etc. With a little bit of effort, you could do the same thing for any value of *n*.

If the sample size is large ($n > 50$ according to Table XI) the sampling distribution of T is reasonably approximated by a normal distribution with mean

$$\mu_T = \frac{n(n+1)}{4}$$

and variance

$$\sigma_T^2 = \frac{n(n+1)(2n+1)}{24}$$

In this case, a test of $H_0$: $\mu = \mu_0$ can be based on the statistic

$$z = \frac{T - \mu_T}{\sigma_T}$$

and an appropriate critical region can be determined from the cumulative distribution function of the standard normal distribution (e.g., Table C4 in the back of Ledolter and Hogg).

Let's redo the concrete strength example using both methods to illustrate the two procedures. It is well-established that the distribution of concrete strengths is symmetrical about the mean, so that data should be a good candidate for this test.

*Example – Wilcoxon Signed-Rank Test for a Population Mean Adapted from Hines & Montgomery (1990) p. 561.*

The local ready-mix concrete plant has developed a mix design for concrete with a 28-day compressive strength of 4000 psi. Last month, quality control technicians made 20 concrete cylinders from a batch of the concrete. Yesterday they obtained the following strengths:

| | | | | |
|---|---|---|---|---|
| 4318 | 4330 | 3416 | 4416 | 4106 |
| 3356 | 4800 | 3570 | 3530 | 4828 |
| 4632 | 3560 | 5150 | 4514 | 4402 |
| 4122 | 4674 | 4716 | 3508 | 5308 |

Given this data, test the hypothesis that the median strength of the concrete mix is 4000 psi.

Begin by calculating the differences between each strength value and 4000 psi, then rank those differences in order of increasing magnitude:

| Strength | Difference | Signed Rank |
|---|---|---|
| 4106 | +106 | +1 |
| 4122 | +122 | +2 |
| 4318 | +318 | +3 |
| 4330 | +330 | +4 |
| 4402 | +402 | +5 |
| 4416 | +416 | +6 |
| 3570 | –430 | –7 |
| 3560 | –440 | –8 |
| 3530 | –470 | –9 |
| 3508 | –492 | –10 |

| Strength | Difference | Signed Rank |
|---|---|---|
| 4514 | +514 | +11 |
| 3416 | –584 | –12 |
| 4632 | +632 | +13 |
| 3356 | –644 | –14 |
| 4674 | +674 | +15 |
| 4716 | +716 | +16 |
| 4800 | +800 | +17 |
| 4828 | +828 | +18 |
| 5150 | 1150 | +19 |
| 5308 | 1308 | +20 |

*Example – Wilcoxon Signed-Rank Test for Paired Data*

Two different brands of fertilizer (A and B) were compared on each of 10 different two-acre plots of barley. Each plot was subdivided into two one-acre subplots, with Brand A randomly assigned to one subplot and Brand B to the other. Fertilizers were then applied to the subplots at the rate of 60 pounds per acre. The data, barley yields in bushels per acre, are listed below:

| *Plot* | *Fertilizer A* | *Fertilizer B* | *Difference (B – A)* | *Ranked Differences* |
|---|---|---|---|---|
| 1 | 312 | 346 | 34 | |
| 2 | 333 | 372 | 39 | |
| 3 | 356 | 392 | 36 | |
| 4 | 316 | 351 | 35 | |
| 5 | 310 | 330 | 20 | |
| 6 | 352 | 364 | 12 | |
| 7 | 389 | 375 | –14 | |
| 8 | 313 | 315 | 2 | |
| 9 | 316 | 327 | 11 | |
| 10 | 346 | 378 | 32 | |

Use the Wilcoxon Signed-Rank Test to test the hypothesis (at $\alpha = 0.05$) that the yields for the two brands of fertilizer are identical against the alternative that Fertilizer B is better.

Wilcoxon's Rank Sum Test

This test can be used to compare the means of two populations with the same shape and spread based on samples obtained from each population. The shapes do not have to be symmetrical but they do have to be the same. The sample sizes do not have to be the same.

In this test, also proposed by Frank Wilcoxon ("Individual comparisons by ranking methods". *Biometrics Bulletin* 1 (6): 80–83, Dec. 1945), the similarity between the two samples is measured by jointly ranking (from lowest to highest) the measurements from the combined samples then examining the sum of the ranks for each individual sample. If the underlying populations have the same mean, shape and spread, the summed ranks should be nearly identical.

This procedure is equivalent to the Mann-Whitney test, but the Mann-Whitney test statistic is usually expressed in a different way. We won't go into the differences here.

Definitions:

$n_1$ = the number of observations in the first (smaller) sample
$n_2$ = the number of observations in the other (larger) sample
$R_1$ = the sum of the ranks of the observations in the first (smaller) sample
$R_2$ = the sum of the ranks of the observations in the other (larger) sample
$R$ = min ($R_1$, $R_2$)

We know that $R_1$ and $R_2$ are related because

$$R_1 + R_2 = \sum_{i=1}^{n_1+n_2} i = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$$

So once we've calculated $R_1$ (for the smaller sample) we can calculate $R_2$ from that value as

$$R_2 = \frac{(n_1+n_2)(n_1+n_2+1)}{2} - R_1$$

As with the Ranked Sum Test, if two or more observations have the same value, they are all given the same average rank so as not to favor one sample over the other.

If the means of the two distributions are the same, then $R_1$ and $R_2$ should be equal to each other and equal to half the sum of $R_1$ and $R_2$ above:

$$R_1 = R_2 = \frac{1}{2}\sum_{i=1}^{n_1+n_2} i = \frac{(n_1+n_2)(n_1+n_2+1)}{4}$$

The further $R_1$ and $R_2$ are from this ideal, the greater the distance between the population means.

If the sample sizes are small, you have to refer to a table of critical values (Table IX on the next page). These tables are developed by enumerating all of the possibilities for specified sample sizes $n_1$ and $n_2$ and determining the largest value of $R$ with a probability of occurrence of 1% or less and 5% or less, respectively.

If the sample sizes are moderately large (typically greater than 8) the distribution of $R$ is well approximated by a normal distribution with a mean of

$$\mu_R = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{4}$$

and a variance of

$$\sigma_R^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

In this case, a test of $H_0$: $\mu_1 = \mu_2$ can be based on the statistic

$$z = \frac{R - \mu_R}{\sigma_R}$$

and an appropriate critical region can be determined from the cumulative distribution function of the standard normal distribution (e.g., Table C4 in the back of Ledolter and Hogg).

TABLE IX **Critical Values for the Wilcoxon Two-Sample Test**

$R^*_{.01}$

| $n_2$ \ $n_1$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | | | | 15 | | | | | | | | | | |
| 6 | | | 10 | 16 | 23 | | | | | | | | | |
| 7 | | | 10 | 17 | 24 | 32 | | | | | | | | |
| 8 | | | 11 | 17 | 25 | 34 | 43 | | | | | | | |
| 9 | | 6 | 11 | 18 | 26 | 35 | 45 | 56 | | | | | | |
| 10 | | 6 | 12 | 19 | 27 | 37 | 47 | 58 | 71 | | | | | |
| 11 | | 6 | 12 | 20 | 28 | 38 | 49 | 61 | 74 | 87 | | | | |
| 12 | | 7 | 13 | 21 | 30 | 40 | 51 | 63 | 76 | 90 | 106 | | | |
| 13 | | 7 | 14 | 22 | 31 | 41 | 53 | 65 | 79 | 93 | 109 | 125 | | |
| 14 | | 7 | 14 | 22 | 32 | 43 | 54 | 67 | 81 | 96 | 112 | 129 | 147 | |
| 15 | | 8 | 15 | 23 | 33 | 44 | 56 | 70 | 84 | 99 | 115 | 133 | 151 | 171 |
| 16 | | 8 | 15 | 24 | 34 | 46 | 58 | 72 | 86 | 102 | 119 | 137 | 155 | |
| 17 | | 8 | 16 | 25 | 36 | 47 | 60 | 74 | 89 | 105 | 122 | 140 | | |
| 18 | | 8 | 16 | 26 | 37 | 49 | 62 | 76 | 92 | 108 | 125 | | | |
| 19 | 3 | 9 | 17 | 27 | 38 | 50 | 64 | 78 | 94 | 111 | | | | |
| 20 | 3 | 9 | 18 | 28 | 39 | 52 | 66 | 81 | 97 | | | | | |
| 21 | 3 | 9 | 18 | 29 | 40 | 53 | 68 | 83 | | | | | | |
| 22 | 3 | 10 | 19 | 29 | 42 | 55 | 70 | | | | | | | |
| 23 | 3 | 10 | 19 | 30 | 43 | 57 | | | | | | | | |
| 24 | 3 | 10 | 20 | 31 | 44 | | | | | | | | | |
| 25 | 3 | 11 | 20 | 32 | | | | | | | | | | |
| 26 | 3 | 11 | 21 | | | | | | | | | | | |
| 27 | 4 | 11 | | | | | | | | | | | | |
| 28 | 4 | | | | | | | | | | | | | |

$R^*_{.05}$

| $n_2$ \ $n_1$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | | | 10 | | | | | | | | | | | |
| 5 | | 6 | 11 | 17 | | | | | | | | | | |
| 6 | | 7 | 12 | 18 | 26 | | | | | | | | | |
| 7 | | 7 | 13 | 20 | 27 | 36 | | | | | | | | |
| 8 | 3 | 8 | 14 | 21 | 29 | 38 | 49 | | | | | | | |
| 9 | 3 | 8 | 15 | 22 | 31 | 40 | 51 | 63 | | | | | | |
| 10 | 3 | 9 | 15 | 23 | 32 | 42 | 53 | 65 | 78 | | | | | |
| 11 | 4 | 9 | 16 | 24 | 34 | 44 | 55 | 68 | 81 | 96 | | | | |
| 12 | 4 | 10 | 17 | 26 | 35 | 46 | 58 | 71 | 85 | 99 | 115 | | | |
| 13 | 4 | 10 | 18 | 27 | 37 | 48 | 60 | 73 | 88 | 103 | 119 | 137 | | |
| 14 | 4 | 11 | 19 | 28 | 38 | 50 | 63 | 76 | 91 | 106 | 123 | 141 | 160 | |
| 15 | 4 | 11 | 20 | 29 | 40 | 52 | 65 | 79 | 94 | 110 | 127 | 145 | 164 | 185 |
| 16 | 4 | 12 | 21 | 31 | 42 | 54 | 67 | 82 | 97 | 114 | 131 | 150 | 169 | |
| 17 | 5 | 12 | 21 | 32 | 43 | 56 | 70 | 84 | 100 | 117 | 135 | 154 | | |
| 18 | 5 | 13 | 22 | 33 | 45 | 58 | 72 | 87 | 103 | 121 | 139 | | | |
| 19 | 5 | 13 | 23 | 34 | 46 | 60 | 74 | 90 | 107 | 124 | | | | |
| 20 | 5 | 14 | 24 | 35 | 48 | 62 | 77 | 93 | 110 | | | | | |
| 21 | 6 | 14 | 25 | 37 | 50 | 64 | 79 | 95 | | | | | | |
| 22 | 6 | 15 | 26 | 38 | 51 | 66 | 82 | | | | | | | |
| 23 | 6 | 15 | 27 | 39 | 53 | 68 | | | | | | | | |
| 24 | 6 | 16 | 28 | 40 | 55 | | | | | | | | | |
| 25 | 6 | 16 | 28 | 42 | | | | | | | | | | |
| 26 | 7 | 17 | 29 | | | | | | | | | | | |
| 27 | 7 | 17 | | | | | | | | | | | | |
| 28 | 7 | | | | | | | | | | | | | |

*Source*: Reproduced with permission from "The Use of Ranks in a Test of Significance for Comparing Two Treatments," by C. White, *Biometrics*, 1952, Vol. 8, p. 37.

[a]For large $n_1$ and $n_2$, $R$ is approximately normally distributed with mean $\frac{1}{2}n_1(n_1+n_2+1)$ and variance $\frac{1}{12}n_1n_2(n_1+n_2+1)$.

*Example: Wilcoxon's Rank Sum Test*

Environmental engineers were interested in determining whether a cleanup project on a nearby lake was effective. One indicator of effectiveness would be a decrease in dissolved oxygen over a period of time. Before starting the project, 12 samples of water were obtained at random from the lake and analyzed for the amount of dissolved oxygen (in ppm). Due to diurnal fluctuations in the dissolved oxygen, all measurements were obtained at the 2 P.M. peak period. Similar data were taken six months after the initiation of the cleanup project. The before and after data are presented below. Do the data show a statistically significant difference at $\alpha = 0.05$?

| *Before Cleanup* | *Rank* | *After Cleanup* | *Rank* |
|---|---|---|---|
| 11.0 | 10 | 10.2 | 1 |
| 11.2 | 14 | 10.3 | 2 |
| 11.2 | 14 | 10.4 | 3 |
| 11.2 | 14 | 10.6 | 4.5 |
| 11.4 | 17 | 10.6 | 4.5 |
| 11.5 | 18 | 10.7 | 6 |
| 11.6 | 19 | 10.8 | 7.5 |
| 11.7 | 20 | 10.8 | 7.5 |
| 11.8 | 21 | 10.9 | 9 |
| 11.9 | 22.5 | 11.1 | 11.5 |
| 11.9 | 22.5 | 11.1 | 11.5 |
| 12.1 | 24 | 11.3 | 16 |

The Runs (Wald-Wolfowitz) Test

This test can be used to determine whether or not events occur in random order. With a bit of creative manipulation, it can also be used to determine whether or not two samples were obtained from the same or different distributions.

In the runs test, the events in a sequence are typically classified as successes (S) or failures (F). Suppose that the following sequence of successes and failures occurred:

S S F F S S S S F F F S S S S

We want to answer the question, "Is there evidence to indicate non-randomness in the sequence?"

A run is defined as a series of like events, with the first and last elements being preceded and succeeded, respectively, by unlike events. For the series above, the runs are as follows:

$$\underset{1}{S\,S}\quad \underset{2}{F\,F}\quad \underset{3}{S\,S\,S\,S}\quad \underset{4}{F\,F\,F}\quad \underset{5}{S\,S\,S\,S}$$

Thus, there are 5 runs in the sequence. You may expect non-randomness if you find either a very large number of runs or a very small number of runs. If the number of runs is very small, the data is non-random due to *clustering*. In other words, all of the data is grouped into a few clusters.

Why would a very large number of runs suggest that the data is not random? If you toss a coin 100 times, do you really expect every odd-numbered toss will be "heads" and every even-numbered toss will be "tails"? Probability doesn't work that way!

If the number of runs is very large, the data are non-random due to *uniformity*. The distribution of heads and tails is too regular (head-tail-head-tail).

*Notation*:

$r$ = the number of runs in a sequence
$n_1$, $n_2$ = the number of successes or failures, $n_1 \leq n_2$

For small sample sizes where $n_1 \leq n_2$ and both are no more than 10-20, the attached table at the end of this handout gives the probability that $r$ is less than or equal to a specified value $\ell$ that ranges from 1 (all of the events are the same) to $n_1 + n_2$ (every odd-numbered toss is "heads" and every even-numbered toss is "tails").

*Example:*

a.) For $n_1 = 3$ successes and $n_2 = 9$ failures, the probability that $r \leq 2$ is 0.009.
b.) For $n_1 = n_2 = 8$, the probability that $r \geq 8$ is: $1 - P(r \leq 7) = 1 - 0.214 = 0.786$

This table was developed by calculating the number of different ways of getting runs of size $r$ and dividing by the total number of ways to arrange $n_1$ successes and $n_2$ failures (or $n_1$ failures and $n_2$ successes, since the definitions of "success" and "failure" are completely arbitrary).

Right-tailed Test:

$H_0$: the sequence is a random arrangement of successes and failures
$H_a$: the sequence is not a random arrangement due to uniformity (too many runs)
*TS:* $\ell$ = the observed number of runs
*RR:* $P(r \geq \ell) = 1 - P(r \leq \ell - 1) \leq \alpha$

Left-tailed Test:

$H_0$: the sequence is a random arrangement of successes and failures
$H_a$: the sequence is not a random arrangement due to clustering (too few runs)
*TS:* $\ell$ = the observed number of runs
*RR:* $P(r \leq \ell) \leq \alpha$

Two-tailed Test

$H_0$: the sequence is a random arrangement of successes and failures
$H_a$: the sequence is not a random arrangement of successes and failures
*TS:* $\ell$ = the observed number of runs
*RR:* $P(r \leq \ell) \leq \alpha/2$ OR $P(r \geq \ell) \leq \alpha/2$

Where $n_1$ and $n_2$ are both equal to 10 or more, $r$ is approximately normally distributed with

$$\mu_T = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

$$\sigma_r = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

This means that for large sample sizes we can use a z test with

$$z = \frac{r - \mu_r}{\sigma_r}$$

instead of using the table at the end of this handout (which only covers $n_1 \leq 10$ and $n_2 \leq 10$).

*Example: Small Sample Runs Test for Randomness*

The data below is a sequence of successes and failures for car emissions inspections. Is there sufficient reason to believe that the sequence is not random? Use $\alpha = 0.05$.

S S F F S S S S F F F S S S S

The sequence has $n_2 = 10$ successes, $n_1 = 5$ failures, and $\ell = 5$ runs:

$$\underset{1}{S\,S}\quad \underset{2}{F\,F}\quad \underset{3}{S\,S\,S\,S}\quad \underset{4}{F\,F\,F}\quad \underset{5}{S\,S\,S\,S}$$

Assume we wish to perform a ***right-tailed*** test to determine if there are too many runs. The rejection region is the upper tail of the distribution of $r$, which means we would reject the null hypothesis if $P(r \geq \ell) \leq 0.05$. From the attached table we find that for $(n_1, n_2) = (5, 10)$:

$$P(r \geq 5) = 1 - P(r \leq 4) = 1.0 - 0.029 = 0.971$$

So there is a 97% chance of having 5 or more runs in a group consisting of 10 successes and 5 failures (or 10 failures and 5 successes). This is certainly not less than $\alpha = 0.05$, so we fail to reject the null hypothesis and conclude that there is insufficient evidence to indicate a lack of randomness due to uniformity.

If we wish to perform a ***left-tailed*** test to determine if there are too few runs, the rejection region is the lower tail of the distribution of $r$, which means we would reject the null hypothesis if $P(r \leq \ell) \leq 0.05$. From the attached table we find that for $(n_1, n_2) = (5, 10)$:

$$P(r \leq 5) = 0.095$$

So there is a 9.5% chance of getting 5 or fewer runs in a group consisting of 10 successes and 5 failures (or 10 failures and 5 successes). This is also not less than $\alpha = 0.05$, so we fail to reject the null hypothesis and conclude that there is insufficient evidence to indicate a lack of randomness due to clustering.

*Runs Test for Different Distributions*

We mentioned earlier that the runs test can also be used to determine whether or not two samples were obtained from the same or different distributions. The two samples (let's call them A and B) don't have to have the same number of observations. Arrange the combined observations from the two samples in numerical order, then replace each observation with the designation (A or B) of the sample from which it came. If the resulting sequence has too few runs, the two samples are probably not from the same distribution.

*Example: Small Sample Runs Test for Different Distributions*

Concrete field technicians made 10 concrete cylinders on the first day of a two-day concrete pour and 10 more cylinders on the second day. After 28 days, each batch of cylinders was tested and the strength data below was reported. You suspect that the concrete plant did not provide the same concrete mix both days. Test your hypothesis by performing a runs test on the data.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Day 1: | 4887 | 4437 | 6040 | 4826 | 3311 | 4570 | 4553 | 4357 | 4612 | 4531 |
| *Day 2:* | *4404* | *4172* | *2931* | *3596* | *3159* | *3596* | *3549* | *2915* | *2962* | *2505* |

We'll start by arranging the strengths in numerical order, making sure to preserve the batch from which each value came:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2505 | 2915 | 2931 | 2962 | 3159 | 3311 | 3549 | 3596 | 3596 | 4172 |
| 4357 | 4404 | 4437 | 4531 | 4553 | 4570 | 4612 | 4826 | 4887 | 6040 |

*Example: Large Sample Runs Test for Different Distributions*

In the previous example $n_1 = n_2 = 10$, so we could also use the large-sample test. Let's repeat the previous analysis using the z test statistic.

**Table 10** Distribution of the total number of runs $r$ in samples of size $(n_1, n_2)$; $P[r \leq \ell]$

| | $\ell$ | | | | | | | | | $\ell$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $(n_1, n_2)$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| (2,3) | .200 | .500 | .900 | 1.000 | | | | | | | | | | | | | | | |
| (2,4) | .133 | .400 | .800 | 1.000 | | | | | | | | | | | | | | | |
| (2,5) | .095 | .333 | .714 | 1.000 | | | | | | | | | | | | | | | |
| (2,6) | .071 | .286 | .643 | 1.000 | | | | | | | | | | | | | | | |
| (2,7) | .056 | .250 | .583 | 1.000 | | | | | | | | | | | | | | | |
| (2,8) | .044 | .222 | .533 | 1.000 | | | | | | | | | | | | | | | |
| (2,9) | .036 | .200 | .491 | 1.000 | | | | | | | | | | | | | | | |
| (2,10) | .030 | .182 | .455 | 1.000 | | | | | | | | | | | | | | | |
| (3,3) | .100 | .300 | .700 | .900 | 1.000 | | | | | | | | | | | | | | |
| (3,4) | .057 | .200 | .543 | .800 | .971 | 1.000 | | | | | | | | | | | | | |
| (3,5) | .036 | .143 | .429 | .714 | .929 | 1.000 | | | | | | | | | | | | | |
| (3,6) | .024 | .107 | .345 | .643 | .881 | 1.000 | | | | | | | | | | | | | |
| (3,7) | .017 | .083 | .283 | .583 | .833 | 1.000 | | | | | | | | | | | | | |
| (3,8) | .012 | .067 | .236 | .533 | .788 | 1.000 | | | | | | | | | | | | | |
| (3,9) | .009 | .055 | .200 | .491 | .745 | 1.000 | | | | | | | | | | | | | |
| (3,10) | .007 | .045 | .171 | .455 | .706 | 1.000 | | | | | | | | | | | | | |
| (4,4) | .029 | .114 | .371 | .629 | .886 | .971 | 1.000 | | | | | | | | | | | | |
| (4,5) | .016 | .071 | .262 | .500 | .786 | .929 | .992 | 1.000 | | | | | | | | | | | |
| (4,6) | .010 | .048 | .190 | .405 | .690 | .881 | .976 | 1.000 | | | | | | | | | | | |
| (4,7) | .006 | .033 | .142 | .333 | .606 | .833 | .954 | 1.000 | | | | | | | | | | | |
| (4,8) | .004 | .024 | .109 | .279 | .533 | .788 | .929 | 1.000 | | | | | | | | | | | |
| (4,9) | .003 | .018 | .085 | .236 | .471 | .745 | .902 | 1.000 | | | | | | | | | | | |
| (4,10) | .002 | .014 | .068 | .203 | .419 | .706 | .874 | 1.000 | | | | | | | | | | | |
| (5,5) | .008 | .040 | .167 | .357 | .643 | .833 | .960 | .992 | 1.000 | | | | | | | | | | |
| (5,6) | .004 | .024 | .110 | .262 | .522 | .738 | .911 | .976 | .998 | 1.000 | | | | | | | | | |
| (5,7) | .003 | .015 | .076 | .197 | .424 | .652 | .854 | .955 | .992 | 1.000 | | | | | | | | | |
| (5,8) | .002 | .010 | .054 | .152 | .347 | .576 | .793 | .929 | .984 | 1.000 | | | | | | | | | |
| (5,9) | .001 | .007 | .039 | .119 | .287 | .510 | .734 | .902 | .972 | 1.000 | | | | | | | | | |
| (5,10) | .001 | .005 | .029 | .095 | .239 | .455 | .678 | .874 | .958 | 1.000 | | | | | | | | | |
| (6,6) | .002 | .013 | .067 | .175 | .392 | .608 | .825 | .933 | .987 | .998 | 1.000 | | | | | | | | |
| (6,7) | .001 | .008 | .043 | .121 | .296 | .500 | .733 | .879 | .966 | .992 | .999 | 1.000 | | | | | | | |
| (6,8) | .001 | .005 | .028 | .086 | .226 | .413 | .646 | .821 | .937 | .984 | .998 | 1.000 | | | | | | | |
| (6,9) | .000 | .003 | .019 | .063 | .175 | .343 | .566 | .762 | .902 | .972 | .994 | 1.000 | | | | | | | |
| (6,10) | .000 | .002 | .013 | .047 | .137 | .288 | .497 | .706 | .864 | .958 | .990 | 1.000 | | | | | | | |
| (7,7) | .001 | .004 | .025 | .078 | .209 | .383 | .617 | .791 | .922 | .975 | .996 | .999 | 1.000 | | | | | | |
| (7,8) | .000 | .002 | .015 | .051 | .149 | .296 | .514 | .704 | .867 | .949 | .988 | .998 | 1.000 | 1.000 | | | | | |
| (7,9) | .000 | .001 | .010 | .035 | .108 | .231 | .427 | .622 | .806 | .916 | .975 | .994 | .999 | 1.000 | | | | | |
| (7,10) | .000 | .001 | .006 | .024 | .080 | .182 | .355 | .549 | .743 | .879 | .957 | .990 | .998 | 1.000 | | | | | |
| (8,8) | .000 | .001 | .009 | .032 | .100 | .214 | .405 | .595 | .786 | .900 | .968 | .991 | .999 | 1.000 | 1.000 | | | | |
| (8,9) | 000 | .001 | .005 | .020 | .069 | .157 | .319 | .500 | .702 | .843 | .939 | .980 | .996 | .999 | 1.000 | 1.000 | | | |
| (8,10) | .000 | .000 | .003 | .013 | .048 | .117 | .251 | .419 | .621 | .782 | .903 | .964 | .990 | .998 | 1.000 | 1.000 | | | |
| (9,9) | .000 | .000 | .003 | .012 | .044 | .109 | .238 | .399 | .601 | .762 | .891 | .956 | .988 | .997 | 1.000 | 1.000 | 1.000 | | |
| (9,10) | .000 | .000 | .002 | .008 | .029 | .077 | .179 | .319 | .510 | .681 | .834 | .923 | .974 | .992 | .999 | 1.000 | 1.000 | 1.000 | |
| (10,10) | .000 | .000 | .001 | .004 | .019 | .051 | .128 | .242 | .414 | .586 | .758 | .872 | .949 | .981 | .996 | .999 | 1.000 | 1.000 | 1.000 |

From "Tables for Testing Randomness of Grouping in a Sequence of Alternatives," C. Eisenhart and F. Swed, *Annals of Mathematical Statistics*, Volume 14 (1943). Reproduced with the kind permission of the Editor, *Annals of Mathematical Statistics*.