Advanced Statistics

Dr. Syed Faisal Bukhari
Associate Professor
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

- **Elementary Statistics,** 13th Edition, Mario F. Triola

Reference books

- **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman
- **Probability Demystified**, Allan G. Bluman
- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson
- **Python for Probability, Statistics, and Machine Learning,** José Unpingco

References

- **Elementary Statistics**, 13th Edition, Mario F. Triola

These notes contain material from the above resource.

Kruskal-Wallis Test for Three or More Samples

The ***Kruskal-Wallis test***, which uses ***ranks* of data from three or more independent simple random samples** to test the null hypothesis that the samples come from populations with the same median.

- One-way **analysis of variance (ANOVA)** as a method for testing the **null hypothesis that three or more populations have the same *mean***, but that **ANOVA procedure requires** that all of the involved populations **have normal distributions**.
- The **Kruskal-Wallis test** for **equal *medians* does not require normal distributions**, so it is a **distribution-free or nonparametric test**.

Kruskal-Wallis Test for Three or More Samples

DEFINITION

The **Kruskal-Wallis test** (also called the ***H* test**) is a **nonparametric test that uses ranks** of combined simple random samples from **three or more independent populations** to test the null hypothesis that the **populations have the same median**. (The alternative hypothesis is the claim that the populations have **medians that are not all equal**.)

Kruskal-Wallis Test

Objective

Use the Kruskal-Wallis test with simple random samples from **three or more independent populations** for the following null and alternative hypotheses:

***H_0*:** The samples come from populations with the same median.

***H_1*:** The samples come from populations with medians that are not all equal.

.

Notation

***N* =** total number of observations in all samples combined

***k* =** number of different samples

R_1 **=** sum of ranks for **Sample 1**

n_1 **=** number of observations in **Sample 1**

For Sample 2, the **sum of ranks is** R_2 and the number of observations is n_2, and similar notation is used for the other samples

Requirements

1. We have at **least three independent simple random samples**.

2. Each sample has at **least five observations.**

(If samples have fewer than five observations, refer to special tables of critical values, such as *CRC Standard Probability and Statistics Tables and Formulae,* published by CRC Press.)

Note: There is *no* requirement that the populations have a normal distribution or any other particular distribution**.**

Test Statistic

$$H = \frac{12}{N(N+1)}\left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \ldots + \frac{R_k^2}{n_k}\right) - 3(N + 1)$$

P-Values

P-values are often provided by technology. By using the test statistic *H* and the number of degrees of freedom ($k - 1$), Table A-4 can be used to find a range of values for the *P*-value.

Critical Values

1. The **test is *right-tailed*** and critical values can be found from technology or from the chi-square distribution in Table A-4.

2. df = k - 1 (where df is the number of degrees of freedom and k is the number of different samples)

TABLE A-4 Chi-Square (χ^2) Distribution

	Area to the *Right* of the Critical Value									
Degrees of Freedom	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Table A-4

NOTES:

Degrees of Freedom

$n - 1$ Confidence interval or hypothesis test for a standard deviation σ or variance σ^2.

$k - 1$ Goodness-of-fit test with k different categories

$(r - 1)(c - 1)$ Contingency table test with r rows and c columns

$k - 1$ Kruskal-Wallis test with k different samples

Procedure for Finding the Value of the *H* Test Statistic

Step 1: Temporarily combine all samples into one big sample and assign a rank to each sample value. (Sort the values from lowest to highest, and in cases of ties, assign to each observation the mean of the ranks involved.)

EXAMPLE: In upcoming example, the **numbers in parentheses are the ranks of the combined data set**. The rank of **1 is assigned to the lowest value of 90**, the rank of 2 is assigned to the next lowest value of 114, and so on. In the case of ties, each of the tied values is assigned the mean of the ranks involved in the tie. **(The seventh and eighth values are tied at 178, so they are each assigned a rank of 7.5.)**

Step 2: For each sample, find the sum of the ranks and find the sample size.

EXAMPLE: In Table 13-6, the sum of the ranks from the first sample is 110, the sum of the ranks for the second sample is 47.5, and the sum of the ranks for the third sample is 32.5.

Step 3: Calculate *H* using the results of Step 2 and the notation and test statistic given in the preceding section.

EXAMPLE: The test statistic is computed in Example 1.

Data Set 35: Car Data

Measurements and crash test results from 48 cars (first five shown here). There are 12 cars in each category of small, midsize, large, and SUV. Crash test results are from cars crashed into a fixed barrier at 35 mi/h with a crash test dummy in the driver's seat. For car measurements, **WEIGHT** is car weight (lb), **LENGTH** is car length (inches), **BRAKING** is braking distance (feet) from 60 mi/h, **CYLINDERS** is the number of cylinders, **DISPLACEMENT** is the engine displacement (liters), **CITY** is the fuel consumption (mi/gal) for city driving conditions, **HWY** is highway fuel consumption (mi/gal) for highway driving conditions, and **GHG** is a measure of greenhouse gas emissions (in tons/year, expressed as CO_2 equivalents). For crash test results, **HIC** is a measurement of a standard "head injury criterion," **CHEST** is chest maximum compression (mm), **LEFT FEMUR** is left leg femur force (in kilonewtons, kN), **RIGHT FEMUR** is right leg femur force (kN), **MAXIMUM NIJ** is a measure of "neck injury criteria." Data are from the National Highway Traffic Safety Administration, the Insurance Institute for Highway Safety, the Environmental Protection Agency, and *Consumer Reports*.

TI-83/84 list names (CARDATA): CWGT, CLNGT, CBRAK, CCYL, CDISPL, CCITY, CHWY, CGHG, CHIC, CCHST, CLFEM, CRFEM, CNIJ (no list for MODEL and SIZE)

MODEL	SIZE	WEIGHT	LENGTH	BRAKING	CYLINDERS	DISPLACEMENT	CITY	HWY	GHG	HIC	CHEST	LEFT FEMUR	RIGHT FEMUR	MAXIMUM NIJ
Toyota Corolla	Small	2844	183	138	4	1.8	29	36	4.6	253	29	1.6	2.8	0.27
Subaru Impreza	Small	3109	176	124	4	2.0	28	38	4.7	143	31	1.4	1.0	0.28
Mazda 3	Small	2870	180	133	4	2.0	30	41	4.4	124	35	0.5	0.3	0.24
Volkswagen Golf	Small	3095	168	130	4	2.0	25	33	5.3	301	33	0.2	0.3	0.33
Honda Civic	Small	2915	182	129	4	1.5	31	42	4.2	422	26	0.4	0.2	0.20

EXAMPLE: Head Injuries in Small, Midsize, and Large Cars

Table given below lists **head injury criterion (HIC)** measurements of **small, midsize,** and **large car crash** tests. Use a 0.05 significance level to test the claim that the three samples of HIC measurements are from **populations with medians that are all equal**

Head Injury Criterion (HIC)

Measurements in Car Crash Tests

Small	Midsize	Large
253	117	249
143	121	90
124	204	178
301	195	114
422	186	183
324	178	
258		
271		

REQUIREMENT CHECK

(1) Each of the three samples is a simple random independent sample.

(2) Each **sample size is at least 5.**

The requirements are satisfied.

Step 1. We state our hypothesis as:

H_0: The populations of **small cars**, midsize cars, and **large cars** all have the same median HIC measurement in crash tests.

H_1: The three populations of **small**, **midsize**, and **large cars** have median HIC measurements that are not all the same.

Step 2. The level of significance is set $\alpha = 0.05$.

Step 3. Test statistic

$$H = \frac{12}{N(N+1)}\left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3}\right) - 3(N + 1)$$

N = total number of observations in all samples combined

$N = n_1 + n_2 + n_3$

R_1 = sum of ranks for **Sample 1**

n_1 = number of observations in **Sample 1**

R_2 = sum of ranks for **Sample 2**

n_2 = number of observations in **Sample 2**

R_3 = sum of ranks for **Sample 3**

n_3 = number of observations in **Sample 3**

Small	Midsize	Large
253 (14)	117 (3)	249 (13)
143 (6)	121 (4)	90 (1)
124 (5)	204 (12)	178 (7.5)
301 (17)	195 (11)	114 (2)
422 (19)	186 (10)	183 (9)
324 (18)	178 (7.5)	
258 (15)		
271 (16)		
$n_1 = 8$ $R_1 = 110$	$n_2 = 6$ $R_2 = 47.5$	$n_3 = 5$ $R_3 = 32.5$

$$H = \frac{12}{19(20)}\left(\frac{110^2}{8} + \frac{(47.5)^2}{6} + \frac{(32.5)^2}{5}\right) - 3(19 + 1)$$

$$H = 6.309$$

Step 5: Critical Value

From Table A-4, with 2 degrees of freedom and α=0.05, the critical value is **5.991**

H > Critical Value

6.309 > 5.991 (TRUE)

Step 6: Conclusion The test statistic exceeds the critical value, falling within the critical region. Therefore, we reject the null hypothesis of equal population medians.

Or

***P*-Value** With $H = 6.309$ and df = 2, Table A-4 shows that the *P*-value is less than 0.05. Using technology, we get *P*-value = 0.043. Because the *P*-value is less than the significance level of 0.05, we reject the null hypothesis of equal population medians.

TABLE A-4 Chi-Square (χ^2) Distribution

	Area to the *Right* of the Critical Value									
Degrees of Freedom	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Table A-4

NOTES:

Degrees of Freedom

$n - 1$ Confidence interval or hypothesis test for a standard deviation σ or variance σ^2.

$k - 1$ Goodness-of-fit test with k different categories

$(r - 1)(c - 1)$ Contingency table test with r rows and c columns

$k - 1$ Kruskal-Wallis test with k different samples

Rank Correlation

The **rank correlation test** (or **Spearman's rank correlation test**) is a nonparametric test that **uses ranks of sample data consisting of matched pairs**. It is used to test for an association between two variables.

Rank Correlation

Objective

Compute the **rank correlation coefficient r_s** and use it to test for an association between two variables. The null and alternative hypotheses are as follows:

H_0: $\rho_s = 0$ (There is no correlation.)

H_1: $\rho_s \neq 0$ (There is a correlation.)

Notation

r_s = rank correlation coefficient for sample paired data

(r_s is a sample statistic)

ρ_s = rank correlation coefficient for all the population data (ρ_s is a population parameter)

n = number of pairs of sample data

d = difference between ranks for the two values within an individual pair

Requirements

1. The paired data are a simple random sample.

2. The data are ranks or can be converted to ranks.

Note: Unlike the parametric methods, there is *no* requirement that the sample pairs of data have **a bivariate normal distribution.**

There is ***no*** **requirement of a normal distribution** for any population

Test Statistic

Within each sample, **first convert the data to *ranks*,** then find the exact value of the rank correlation coefficient r_s by using

Formula 10-1

$$r_s = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

Simpler Test Statistic if There Are No Ties: After converting the data in each sample to ranks, if there are no ties among ranks for the first variable and there are no ties among ranks for the second variable, the exact value of the test statistic can be calculated using the Formula 10-1 or with the following relatively simple formula, but it is probably easier to use Formula 10-1 with technology:

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

P-Values

P-values are sometimes provided by technology, but use them only if they result from Spearman's rank correlation.

(*Caution: Do not use P-values from linear correlation for methods of rank correlation*. When working with data having ties among ranks, the rank correlation coefficient *rs* can be calculated using Formula 10-1. Technology can be used instead of manual calculations with Formula 10-1, but the displayed *P*-values from *linear* correlation do not apply to the methods of *rank* correlation.)

Critical Values

1. If $n \leq 30$, critical values are found in **Table A-9**.

2. If $n > 30$, critical values of r_s are found using the formula:

$$r_s = \frac{\pm z}{\sqrt{n-1}} \text{ (critical values for } n > 30\text{)}$$

where the value of z corresponds to the significance level. (For example, if $\alpha = 0.05$, $z = 1.96$.)

Advantages of Rank Correlation

Rank correlation has these advantages over the parametric methods:

1. Rank correlation can be used with paired data that are ranks or can be **converted to ranks**. Unlike the parametric methods, the method of rank correlation does ***not*** **require a normal distribution for any population**.

2. Rank correlation can be used to **detect some (not all) relationships** that are **not linear.**

Efficiency of Rank Correlation

- **Efficiency Rating**: Rank correlation has an efficiency rating of approximately **0.91.**

- **Sample Size Requirement:**
  - The nonparametric rank correlation approach if requires **100 pairs of sample data.**

  - This is equivalent to the results achieved with **91 pairs of sample observations** using a parametric method.

Efficiency of Rank Correlation cont.

- **Assumption Dependence**:
  - The parametric method assumes stricter conditions **(e.g., normality)** for this efficiency comparison to hold true.

- **Key Implication:** Rank correlation is **slightly less efficient** than parametric methods when the parametric assumptions are met.

TABLE A-4 Chi-Square (χ^2) Distribution

	Area to the *Right* of the Critical Value									
Degrees of Freedom	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Chi-Square (χ^2) Critical Values

Degrees of Freedom	α = 0.10	α = 0.05	α = 0.025	α = 0.01	α = 0.005
1	2.705	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.144	32.852	36.191	38.582
20	28.412	31.410	34.170	37.566	39.997

Table A-4

NOTES:

Degrees of Freedom

$n - 1$ Confidence interval or hypothesis test for a standard deviation σ or variance σ^2.

$k - 1$ Goodness-of-fit test with k different categories

$(r - 1)(c - 1)$ Contingency table test with r rows and c columns

$k - 1$ Kruskal-Wallis test with k different samples

EXAMPLE 1 Do Better Smartphones Cost More?

Table given below lists **ranks and costs (dollars) of smartphones** (based on data from *Consumer Reports*). **Lower ranks correspond to better smartphones**.

Find the value of the **rank correlation coefficient** and use it to determine whether there is sufficient evidence to support the claim of a **correlation between quality and price**. Use a 0.05 significance level. Based on the result, **does it appear that you get a better quality smartphone by spending more**?

Quality Rank	1	2	3	4	5	6	7	8	9	10
Cost (dollars)	1000	1100	900	1000	750	1000	900	700	750	600

REQUIREMENT CHECK

- The sample data are a **simple random sample** from the smartphones that were tested.

- The **data are ranks** or can be **converted to ranks**.

The quality ranks are consecutive integers and are not from a population that is normally distributed, so we use the rank **correlation coefficient instead of the linear correlation** coefficient to test for a relationship between quality and price.

Step 1. We state our hypothesis as:

H_0: $\rho_s = 0$ **(**There is *no* correlation between quality and price.)

H_1: $\rho_s \neq 0$ (There is a correlation between quality and price.)

Step 2. The level of significance is set $\alpha = 0.05$.

Step 3. Test statistic to be used is

$$r_s = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

The **data (i.e, x and y) are ranks** or can be **converted to ranks**.

Step 4. Calculations:

Convert cost in cost rank

Rank	1	2	3	4	5	6	7	8	9	10
Quality Rank	1	2	3	4	5	6	7	8	9	10
Cost Rank	8	10	5.5	8	3.5	8	5.5	2	3.5	1

x	y	xy	x^2	y^2
1	8	8	1	64
2	10	20	4	100
3	5.5	16.5	9	30.25
4	8	32	16	64
5	3.5	17.5	25	12.25
6	8	48	36	64
7	5.5	38.5	49	30.25
8	2	16	64	4
9	3.5	31.5	81	12.25
10	1	10	100	1
$\sum x$ =55	$\sum y$ = 55	$\sum xy$ = 238	$\sum x^2$= 285	$\sum y^2$ =382

$$r_s = \frac{\mathrm{n}(\sum xy) - (\sum x)(\sum y)}{\sqrt{\mathrm{n}(\sum x^2) - (\sum x)^2}\sqrt{\mathrm{n}(\sum y^2) - (\sum y)^2}}$$

$$r_s = \frac{10(238) - (55)(55)}{\sqrt{10(285) - (55)^2}\sqrt{\mathrm{n}(382) - (55)^2}}$$

$$= -0.796$$

Step 5: Critical Value

Referring to **Table A-9**, we identify the critical values as **±0.648**, determined using $\alpha = 0.05$ and $n = 10$.

The computed test statistic, $r_s = -0.796$, lies outside the critical range of **-0.648 and 0.648**.

Step 6: Conclusion

We **reject the null hypothesis**.

There is sufficient evidence to support a claim **of a correlation between quality and cost**. It appears that you do get better quality by paying more, but this conclusion incorrectly implies causation.

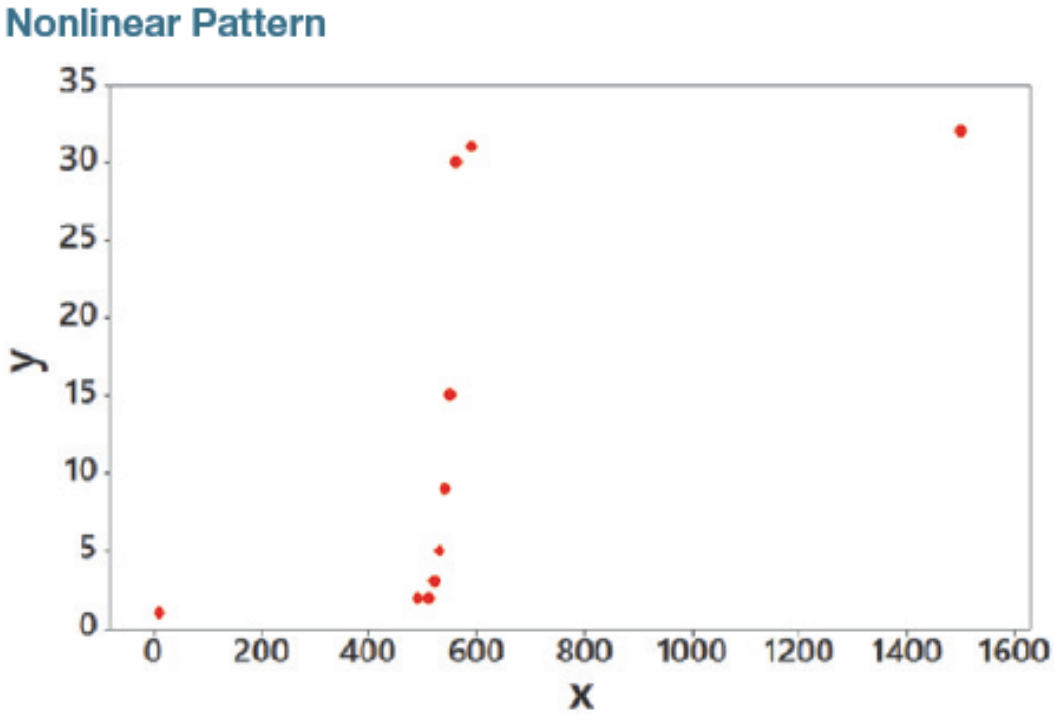
Detecting Nonlinear Patterns

***Rank correlation* methods sometimes** allow us to detect **relationships that we cannot detect with the *linear correlation* methods.**

See scatterplot on the following page, which shows an **S-shaped pattern of points suggesting that there is a correlation between *x* and *y*.** The methods in the **linear correlation coefficient of *r* = 0.627** and critical values of $\pm$ 0.632, suggesting that there is not sufficient evidence to support the claim of a linear correlation between *x* and *y*.

If we use rank correlation and the methods of this section, we get ***r* = 0.997 and critical values of $\pm$ 0.648**, suggesting that there is sufficient evidence to support the claim of a correlation between *x* and *y*. **Linear correlation missed it, but rank correlation recognized it.**

With rank correlation, we can sometimes detect relationships that are not linear.



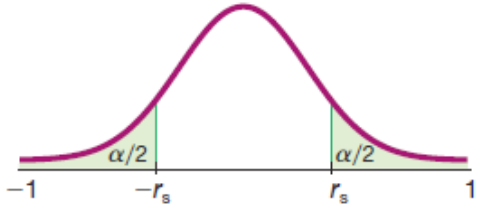


TABLE A-9 Critical Values of Spearman's Rank Correlation Coefficient r_s

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.02$	$\alpha = 0.01$
5	.900	—	—	—
6	.829	.886	.943	—
7	.714	.786	.893	.929
8	.643	.738	.833	.881
9	.600	.700	.783	.833
10	.564	.648	.745	.794
11	.536	.618	.709	.755
12	.503	.587	.678	.727
13	.484	.560	.648	.703
14	.464	.538	.626	.679
15	.446	.521	.604	.654
16	.429	.503	.582	.635
17	.414	.485	.566	.615
18	.401	.472	.550	.600
19	.391	.460	.535	.584
20	.380	.447	.520	.570
21	.370	.435	.508	.556
22	.361	.425	.496	.544
23	.353	.415	.486	.532
24	.344	.406	.476	.521
25	.337	.398	.466	.511
26	.331	.390	.457	.501
27	.324	.382	.448	.491
28	.317	.375	.440	.483
29	.312	.368	.433	.475
30	.306	.362	.425	.467

NOTES:

1. For $n > 30$ use $\boldsymbol{r_s} = \frac{\pm z}{\sqrt{n-1}}$, where z corresponds to the level of significance. For example, if $\alpha = 0.05$, then $z = 1.96$.

2. If the **absolute value of the test statistic $\boldsymbol{r_s}$** is greater than or equal to the positive critical value, then reject H_0: $\boldsymbol{\rho_s} = 0$ and conclude that there is sufficient evidence to support the claim of a correlation.

Runs Test For Randomness

The ***runs test for randomness,*** which is used to **determine whether a sequence of sample data has a random order**. This test requires a criterion for categorizing each **data value into one of two separate categories**, and it analyzes *runs* of those two categories to determine whether the runs appear to result from a random process, or whether the runs suggest that the order of the data is not random.

TABLE A-10 Runs Test for Randomness: Critical Values for Number of Runs G

Value of n_1	Value of n_2 = 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
3	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
	6	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
4	1	1	1	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
	6	8	9	9	9	10	10	10	10	10	10	10	10	10	10	10	10	10	10
5	1	1	2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
	6	8	9	10	10	11	11	12	12	12	12	12	12	12	12	12	12	12	12
6	1	2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
	6	8	9	10	11	12	12	13	13	13	13	14	14	14	14	14	14	14	14
7	1	2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
	6	8	10	11	12	13	13	14	14	14	14	15	15	15	16	16	16	16	16
8	1	2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
	6	8	10	11	12	13	14	14	15	15	16	16	16	16	17	17	17	17	17
9	1	2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
	6	8	10	12	13	14	14	15	16	16	16	17	17	18	18	18	18	18	18
10	1	2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
	6	8	10	12	13	14	15	16	16	17	17	18	18	18	19	19	19	20	20
11	1	2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
	6	8	10	12	13	14	15	16	17	17	18	19	19	19	20	20	20	21	21
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10
	6	8	10	12	13	14	16	16	17	18	19	19	20	20	21	21	21	22	22
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
	6	8	10	12	14	15	16	17	18	19	19	20	20	21	21	22	22	23	23
14	2	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	11
	6	8	10	12	14	15	16	17	18	19	20	20	21	22	22	23	23	23	24
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
	6	8	10	12	14	15	16	18	18	19	20	21	22	22	23	23	24	24	25
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
	6	8	10	12	14	16	17	18	19	20	21	21	22	23	23	24	25	25	25
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13
	6	8	10	12	14	16	17	18	19	20	21	22	23	23	24	25	25	26	26
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
	6	8	10	12	14	16	17	18	19	20	21	22	23	24	25	25	26	26	27
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
	6	8	10	12	14	16	17	18	20	21	22	23	23	24	25	26	26	27	27
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14
	6	8	10	12	14	16	17	18	20	21	22	23	24	25	25	26	27	27	28

A-10 Table

NOTES:

1. The entries in this table are the critical *G* values, assuming a two-tailed test with a significance level of α = 0.05.

2. Reject the null hypothesis of randomness if either of these conditions is satisfied:

- The number of runs *G* is less than or equal to the smaller entry in the table.
- The number of runs *G* is greater than or equal to the larger entry in the table.

Runs Test For Randomness

DEFINITIONS

After characterizing each data value as one of two separate categories, **a run is a sequence of data having the same characteristic**; the sequence is preceded and followed by data with a different characteristic or by no data at all.

The **runs test** uses the number of runs in a sequence of sample data to test for randomness in the order of the data.

Fundamental Principle of the Runs Test

Here is the key idea underlying the runs test:

Reject randomness if the number of runs is very *low* or very *high*.

Example: The sequence of genders **FFFFFMMMMM** is not random because it has only 2 runs, so the number of runs is very *low*.

Example: The sequence of genders **FMFMFMFMFM** is not random because there are 10 runs, which is very *high*.

- **CAUTION: Runs Test for Randomness**
- **Focus on Order:** The runs test evaluates randomness based on the sequence or order in which data points occur, not their frequency.

- **Example Clarification:** A sequence with **3 men and 20 women** might appear random in order but does not consider whether this represents a **biased sample** (e.g., disproportionately more women).

- **Limitation:** The test does not address concerns about sample composition or imbalance; it strictly assesses the randomness of the arrangement.

Runs Test for Randomness

Objective

Apply the runs test for randomness to a *sequence* of sample data to test for randomness in the *order* of the data. Use the following null and alternative hypotheses:

H_0: The data are in a random order.

H_1: The data are in an order that is not random.

Notation

n_1 = number of elements in the sequence that have one particular characteristic. (The characteristic chosen for n_1 is arbitrary.)

n_2 = number of elements in the sequence that have the other characteristic

G = number of runs

Requirements

1. The sample data are arranged according to some ordering scheme, such as the order in which the sample values were obtained.

2. Each data value can be categorized into one of ***two*** **separate categories** (such as male/female).

Test Statistic and Critical Values

For Small Samples and $\alpha = 0.05$: If $n_1 \leq 20$ and $n_2 \leq 20$ and the significance level is $\alpha = 0.05$, the test statistic, critical values, and decision criteria are as follows:

Test statistic: number of runs *G*

Critical values of G: Use Table A-10.

- **Decision criteria:** Reject randomness if the number of runs *G* is such that
- $G \leq$ smaller critical value found in Table A-10.
- or $G \geq$ larger critical value found in Table A-10.

For Large Samples or $\alpha \neq 0.05$: If $n_1 > 20$ or $n_2 > 20$ or a 0.05, the test statistic, critical values, and decision criteria are as follows:

Where

$$z = \frac{G - \mu_G}{\sigma_G}$$

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

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

- **Critical values of *z*:** Use Table A-2.
- **Decision criteria:** Reject randomness if the test statistic *z* is such that
- $z \leq$ negative critical *z* score (such as -1.96).
- or $z \geq$ positive critical *z* score (such as 1.96).

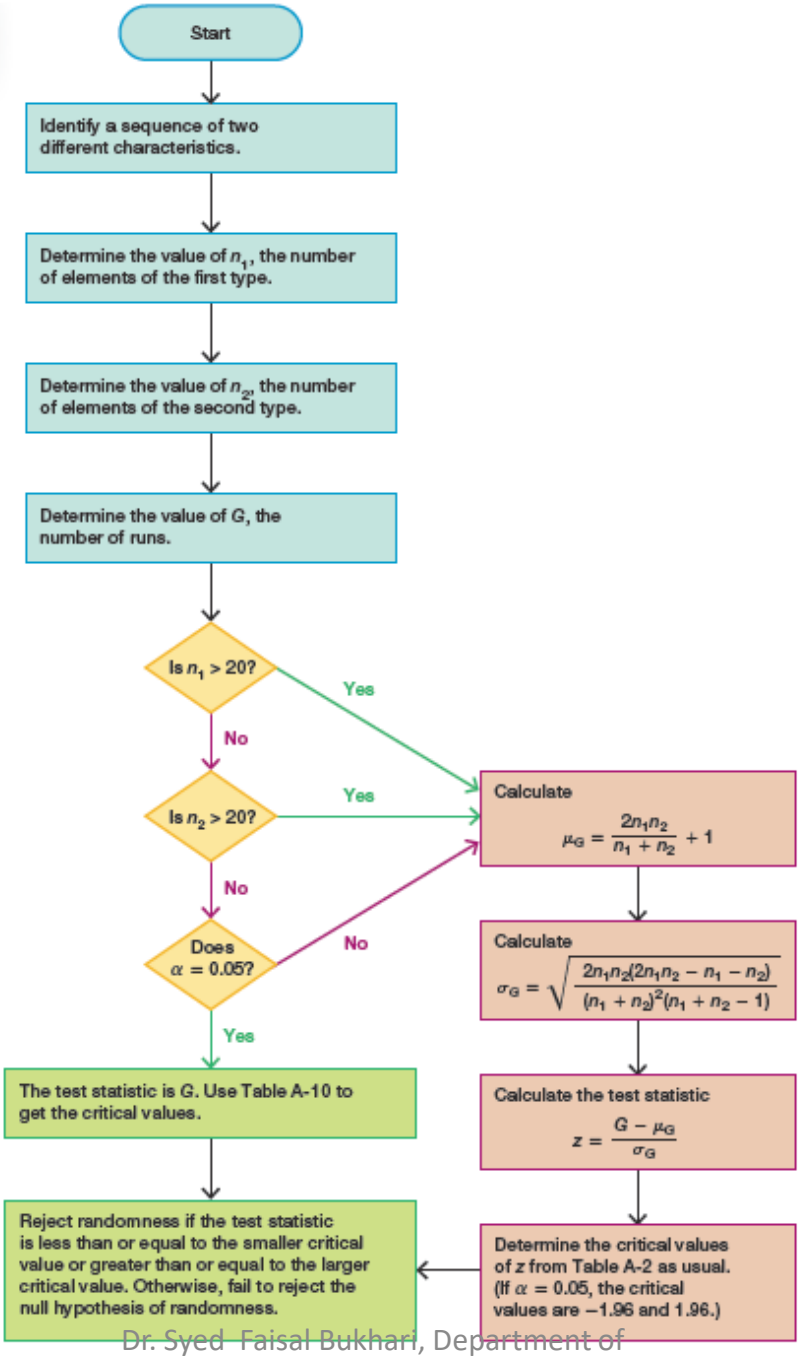


FIGURE 13-5 Procedure for Runs Test for Randomness

EXAMPLE 1 Small Sample: Political Parties of Presidents

Listed below are the political parties of the **past 15 presidents of the United States (as of this writing)**. The letter **R represents a Republican president** and the letter **D represents a Democratic president**. Use a 0.05 significance level to test for randomness in the sequence.

R D D R D D R R D R R D R D R

SOLUTION

REQUIREMENT CHECK

(1) The data are **arranged in order**.

(2) Each data value is Categorized into one of **two separate categories** (Republican/Democrat).

The requirements are satisfied.

We will follow the procedure summarized in Figure 13-5. The sequence of two characteristics (Republican/Democrat) has been identified.

We must now find the values of n_1, n_2, and the number of runs *G*. The sequence is shown below with spacing adjusted to better identify the different runs.

R	DD	R	DD	RR	D	RR	D	R	D	R
1st run	2nd run	3rd run	4th run	5th run	6th run	7th run	8th run	9th run	10th run	11th run

The above display shows that there are **8 Republican presidents** and **7 Democratic presidents**, and **the number of runs is 11**. We represent those results with the following notation

n_1 = number of Republican presidents = **8**

n_2 = number of Democratic presidents = **7**

G = number of runs = **11**

Because $n_1 \leq 20$ **and** $n_2 \leq 20$ **and the significance level is** α **= 0.05**, the test statistic is ***G* = 11 (the number of runs),** and we refer to Table A-10 to find the critical values of **4 and 13**. Because ***G* = 11 is neither less than or equal to the lower critical value of 4**, nor is it greater than or equal to the upper critical value of 13, ***we do not reject randomness***.

There is not sufficient evidence to reject randomness in the sequence of political parties of recent presidents. Based on the given data, it appears that **Republicans and Democrats become presidents in random order.**

Large Sample: Testing Temperatures for Randomness Above and Below the Mean

EXAMPLE 2

Use the following mean global temperatures (℃) for 50 recent and consecutive years to test for **randomness above and below the mean**. Use a 0.05 significance level. The data are listed in order by row.

13.98 14.10 14.05 14.03 13.65 13.75 13.93 13.98 13.91 14.00

14.04 13.90 13.95 14.18 13.94 13.98 13.79 14.16 14.07 14.13

14.27 14.40 14.10 14.34 14.16 14.13 14.19 14.35 14.42 14.28

14.49 14.44 14.16 14.18 14.31 14.47 14.36 14.40 14.71 14.44

14.41 14.56 14.70 14.64 14.60 14.77 14.64 14.66 14.68 14.70

SOLUTION

REQUIREMENT CHECK

(1) The data are arranged in order.

(2) Each data value can be categorized into one of two separate categories: below the mean or above the mean.

The requirements are satisfied.

The null and alternative hypotheses are as follows:

H_0: The sequence is random.

H_1: The sequence is not random.

The mean of those 50 temperatures is **14.250**℃. If we replace each temperature with B if it is below the mean and A if it is above the mean, we get the following sequence. For example, the first temperature of 13.98℃ is *below* the mean of **14.250**℃, so 13.98 is replaced with B.

B B B B B B B B B B

B B B B B B B B B B

A A B A B B B A A A

A A B B A A A A A A

A A A A A A A A A A

Examination of the sequence of B's and A's shows that the letter B occurs 26 times, the letter A occurs 24 times, and the number of runs is 8, so we have the following:

n_1 = number of B's = 26

n_2 = number of A's = 24

G = number of runs = 8

Since, $n_1 > 20$, we need to calculate the test statistic z, so we must first evaluate μ_G and σ_G as follows:

$$\mu_G = \frac{2n_1n_2}{n_1+n_2} + 1 = \frac{2(26)(24)}{24+26} + 1$$
$$= 25.96$$

and

$$\sigma_G = \sqrt{\frac{(2n_1n_2)(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}} = \sqrt{\frac{(2)(26)(24)_{[2)(26)(24)-26-24]}}{(24+26)^2(24+26-1)}}$$
$$= 3.49355558$$

$$z = \frac{G-\mu_G}{\sigma_G} = z = \frac{8-25.96}{3.49355558} = -5.14$$

We can use the test statistic z = -5.14 to find that the P-value is 0.000, which is less than the significance level of Type equation here. = 0.05, so we reject the null hypothesis of randomness.

Also, because the significance level is α = 0.05 and we have a two-tailed test, the critical values are **z = -1.96 and z = 1.96. The test statistic of z = -5.14** does fall within the critical region, so we again reject the null hypothesis of randomness.

INTERPRETATION

We have sufficient evidence to reject randomness of the sequence of 50 mean global temperatures. If we simply examine the list of B's and A's, we see that the B's are mostly in the beginning of the sequence and the A's are mostly at the end, suggesting that over this period of 50 years, the mean global temperature is *increasing*