DEFINITIONS AND FORMULAE WITH STATISTICAL TABLES FOR ELEMENTARY STATISTICS AND QUANTITATIVE METHODS COURSES

Department of Statistics
University of Oxford

October 2015

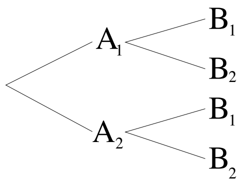
Contents

1 Laws of Probability

Multiplication law $P(A \text{ and } B) = P(A)P(B|A) = P(B)P(A|B)$

Addition law $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Bayes' Rule



$$P(B_1) = P(B_1|A_1)P(A_1) + P(B_1|A_2)P(A_2)$$

$$P(A_1|B_1) = \frac{P(B_1|A_1)P(A_1)}{P(B_1)}$$

2 Theoretical mean and variance for discrete distributions

$$\mu = \Sigma x p(x) \qquad \sigma^2 = \Sigma(x-\mu)^2 p(x)$$

3 Mean and variance for sums of Normal random variables

If $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2),\ i = 1, \ldots, n$ and let $Y = \sum_{i=1}^{n} a_i X_i$ then

$$\mu_Y = \sum_{i=1}^{n} a_i \mu_i \qquad \sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

4 Estimates from samples

Ungrouped data:

sample mean $\bar{x} = \frac{\Sigma x_i}{n}$ sample variance $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{\Sigma x_i^2 - n\bar{x}^2}{n-1}$

Grouped data: $\bar{x} = \frac{\Sigma f x}{\Sigma f}$ $s^2 = \frac{\Sigma f(x-\bar{x})^2}{\Sigma f - 1}$

Counted events: for x in a sample of size n, sample proportion $\hat{p} = \frac{x}{n}$

5 Two common discrete distributions

(i) *Binomial*

$p(x) = {}^n\mathrm{C}_x\, p^x q^{n-x}$, $x = 0, 1, \ldots, n$

$\mu = np,\ \sigma^2 = npq,\ \sigma = \sqrt{npq}$

${}^nC_x = \frac{n!}{x!(n-x)!}$

(ii) *Poisson*

$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, \ldots, \infty$

$\mu = \lambda,\quad \sigma^2 = \lambda,\quad \sigma = \sqrt{\lambda}$

6 Standard errors

Single sample of size n

$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$ or, if σ unknown, $\frac{s}{\sqrt{n}}$

$SE(\hat{p}) = \sqrt{\frac{pq}{n}}$ with $q = 1 - p$, or, if p unknown, $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Sampling without replacement

When n individuals are sampled from a population of N **without replacement**, the standard error is reduced. The standard error for no replacement SE_{NR} is related to the standard error with replacement SE_{WR} by the formula

$$SE_{\text{NR}} = SE_{\text{WR}}\sqrt{\left(1 - \frac{n-1}{N-1}\right)} = \frac{\sigma}{\sqrt{n}}\sqrt{\left(1 - \frac{n-1}{N-1}\right)},$$

where σ is the known standard deviation of the whole population.

Two independent samples of sizes, n_1 *and* n_2

$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ or, if σ_1 and σ_2 unknown and different, $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

For common but unknown σ , $SE(\bar{x}_1 - \bar{x}_2) = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ with $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$, or,

if p_1 and p_2 unknown and unequal, $\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

For common but unknown p , $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})}$ where $\hat{p}$ is a pooled estimate of p defined as $\hat{p} = \frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$ and $\hat{q} = 1 - \hat{p}$.

7 95% confidence limits for population parameters

Mean: when σ known use $\bar{x} \pm 1.96\,\frac{\sigma}{\sqrt{n}}$

when σ unknown use $\bar{x} \pm t\frac{s}{\sqrt{n}}$

where t is the tabulated two-sided 5% level value with degrees of freedom, d.f. $= n - 1$

Proportion: $\hat{p} \pm 1.96\sqrt{\hat{p}\hat{q}/n}$

8 z-tests

Single sample test for population mean μ (known σ): $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

Single sample test for population proportion p: $z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$

Two sample test for difference between two means (known σ_1 and σ_2): $z = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$

Two sample test for difference between two proportions : $z = \frac{\hat{p_1}-\hat{p_2}}{\sqrt{\hat{p}\hat{q}(\frac{1}{n_1}+\frac{1}{n_2})}}$

where $\hat{p}$ is a pooled estimate of p defined as $\hat{p} = \frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$ and $\hat{q} = 1 - \hat{p}$

9 t-tests

Population variance σ^2 unknown and estimated by s^2

Single sample test for population mean μ $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ with d.f.$= n - 1$

Paired samples : test for zero mean difference, using n pairs (x, y), $d = x - y$

$t = \frac{\bar{d}}{s_d/\sqrt{n}}$ with d.f.$= n - 1$, where $\bar{d}$ and s_d are the mean and standard deviation of d.

Independent samples test for difference between population means μ_x and μ_y using n_x x's and n_y y's. Provided that s_x^2 and s_y^2 are similar values, use *the pooled variance estimate,*

$s^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$, and $t = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}$ with d.f $= n_x + n_y - 2$

10 The χ^2 -test

(Note that the two tests in this section are *nonparametric* tests. There are χ^2 tests of variances, not included here, that are *parametric.*)

χ^2 *Goodness-of-fit tests* using k groups have
d.f.$=(k - 1) - p$ where p is the number of independent parameters estimated
and used to obtain the (fitted) expected values.

χ^2 *Contingency table tests* on two-way tables with r rows and c columns have
d.f.$= (r - 1)(c - 1)$

For both tests, $\chi^2 = \Sigma\frac{(O-E)^2}{E}$ where O is an observed frequency and E is the corresponding expected frequency

11 Correlation and regression

For n pairs (x_i, y_i), with sample variances s_x^2 and s_y^2 as in section 3, define *sample covariance,* $s_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{(n-1)}$. May be computed as $s_{xy} = \frac{1}{n-1}\sum x_i y_i - \frac{n}{n-1}\bar{x}\bar{y}$.
Computed from the sample variances s_{x+y}^2 of $x+y$ and s_{x-y}^2 of $x-y$ as $s_{xy} = \frac{1}{4}(s_{x+y}^2 - s_{x-y}^2)$.

Sample product-moment correlation coefficient, $r = \dfrac{s_{xy}}{s_x s_y}$

Test the significance of the correlation coefficient, $\rho = 0$, or equivalently, of the regression slope $\beta = 0$: $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ with d.f.$= n-2$

Linear Regression of y on x. Equation $y = \alpha + \beta x$ with α and β estimated by a and b.

Estimates: $b = \dfrac{s_{xy}}{s_x^2}$ $a = \bar{y} - b\bar{x}$.

Root-mean-square error of regression prediction given by $s_y\sqrt{1-r^2}$.

Test significance of regression as above or $t = \dfrac{b}{SE(b)}$ with d.f.$= n-2$ where $SE(b) = \dfrac{b\sqrt{1-r^2}}{r\sqrt{n-2}}$

95% confidence limits for the slope β are: $b \pm t.SE(b)$, where t is the tabulated two-sided 5% level value with d.f.$= n-2$

12 Analysis of variance

Single factor or One-way analysis for a completely randomized design.

The test statistic F is calculated as a ratio of two mean squares. If the numbers in the k groups are $n_1, n_2, \ldots, n_k$ then the total sample size is $\Sigma n_i = n$. Calculate the "total sum of squares", $TSS = (n-1)s_T^2$, where s_T^2 is variance of all n observations. Calculate the sample means and sample variances of the k groups by $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ and $s_1^2, s_2^2, \ldots, s_k^2$ and then the "within groups sum of squares" (also known as the"error sum of squares"), $ESS = \Sigma(n_i - 1)s_i^2$. The "between groups sum of squares" may be computed in two different ways: $BSS = \sum n_i(\bar{x}_i - \bar{x}_T)^2$, where $\bar{x}_T$ is the mean of all n observations; or $BSS = TSS - ESS$.

These together with their degrees of freedom are entered into the ANOVA table:

Source of variation	*Degrees of freedom*	*Sum of squares*	*Mean square*	*F*
Between samples	$k-1$	BSS	$BMS = \frac{BSS}{k-1}$	$\frac{BMS}{EMS}$
Within samples	$n-k$	ESS	$EMS = \frac{ESS}{n-k}$	
Total	$n-1$	TSS		

13 Median test for two independent samples

For two independent samples, sizes n_1 and n_2, the median of the whole sample of $n = n_1 + n_2$ observations is found. The number in each sample above this median is counted and expressed as a proportion of that sample size. The two proportions are compared using the Z-test as in §8.

14 Rank sum test or Mann-Whitney test

For two independent samples, sizes n_1 and n_2, ranked without regard to sample, call the sum of the ranks in the smaller sample R. If $n_1 \leq n_2 \leq 10$ refer to Table 5, otherwise use a Z test with $z = (R - \mu)/\sigma$ where $\mu = \frac{1}{2}n_1(n_1 + n_2 + 1)$ and $\sigma = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$, assuming $n_1 \leq n_2$. In case of ties, ranks are averaged.

15 Sign test for matched pairs

The number of positive differences from the n pairs is counted. This number is binomially distributed with $p = \frac{1}{2}$, assuming a population zero median difference. So apply the Z test for a binomial proportion with $p = \frac{1}{2}$.

16 Wilcoxon test for matched pairs

Ignoring zero differences, the differences between the values in each pair are ranked without regard to sign and the sums of the positive ranks, R_+ and of the negative ranks, R_-, are calculated. (Check $R_+ + R_- = \frac{1}{2}n(n+1)$, where n is the number of nonzero differences). The smaller of R_+ and R_- is called T and may be compared with the critical values in Table 6 for a two-tailed test. (For one-tailed tests, use R_- and R_+ with the same table, remembering to halve P.) In case of ties, ranks are averaged.

17 Kolmogorov-Smirnov test

Two samples of sizes n_1 and n_2 are each ordered along a scale. At each point on the scale the empirical cumulative distribution function is calculated for each sample and the difference between the pairs are recorded as D_i. The largest absolute value of the D_i is called D_{max} and this value is compared with the 5% one-tailed value

$$D_{crit} = 1.36\sqrt{\frac{n_1 + n_2}{n_1 n_2}}.$$

Single sample version, compares sample with theoretical distribution,

$$D_{crit} = 1.36\sqrt{\frac{1}{n}}.$$

Should only be used with no ties, but it commonly is used otherwise. With ties, the value of D_{max} tends to be too small, so that the p-value is an overestimate.

18 Kruskal-Wallis test for several independent samples

(Analysis of variance for a single factor). For k samples of sizes $n_1, n_2, ..n_k$, comprising a total of n observations, all values are ranked without regard to sample, from 1 to n. The rank sums for the samples are calculated as $R_1, R_2, .., R_k$. (Check $\Sigma R_i = \frac{1}{2}n(n+1)$. The test statistic is

$$H = \left[\frac{12}{n(n+1)}\Sigma\frac{R_i^2}{n_i}\right] \; -3(n+1),$$

which is compared to χ^2 table with d.f. $= k-1$

19 Spearman's Rank Correlation Coefficient

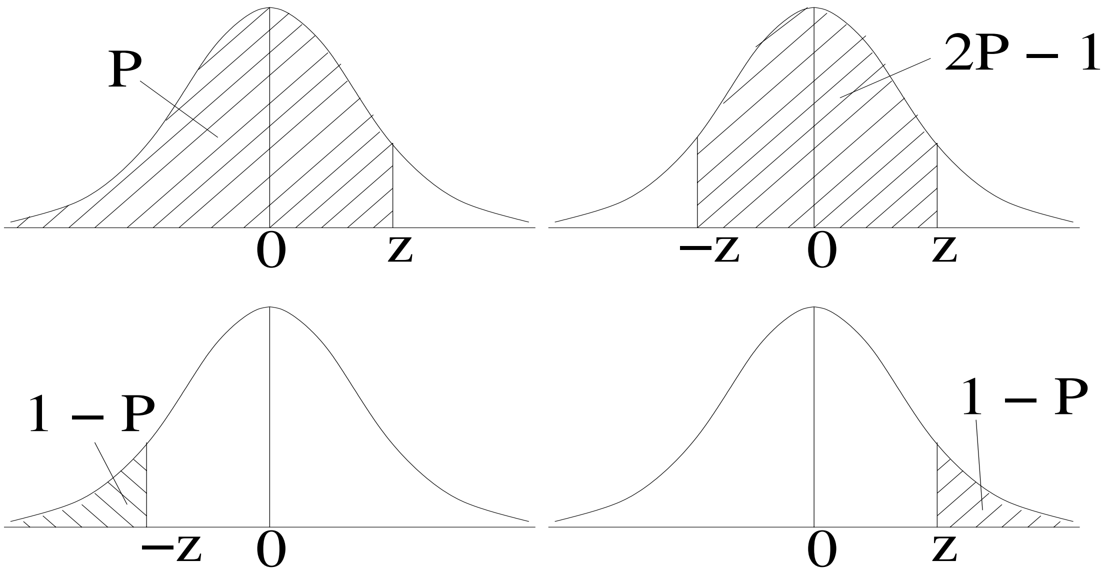
If x and y are ranked variables the Spearman Rank Correlation Coefficient is just the sample product moment correlation coefficient between the pairs of ranks, r_s, which may also be computed by

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

where d is the difference $x-y$, and n is the number of pairs (x, y).
Test r_s using $t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$ with d.f.$= n-2$

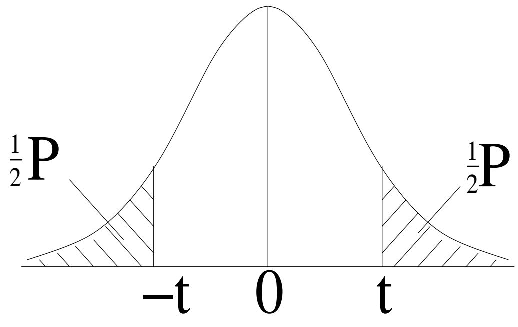
TABLE 1 : *The Normal Integral*

Tabled value is P



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0.1	0.5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0.2	0.5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0.3	0.6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0.4	0.6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0.5	0.6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0.6	0.7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0.7	0.7580	7611	7642	7673	7704	7734	7764	7794	7823	7852
0.8	0.7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0.9	0.8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1.0	0.8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1.1	0.8643	8665	8686	8708	8729	8749	8770	8790	8810	8830
1.2	0.8849	8869	8888	8907	8925	8944	8962	8980	8997	9015
1.3	0.9032	9049	9066	9082	9099	9115	9131	9147	9162	9177
1.4	0.9192	9207	9222	9236	9251	9265	9279	9292	9306	9319
1.5	0.9332	9345	9357	9370	9382	9394	9406	9418	9429	9441
1.6	0.9452	9463	9474	9484	9495	9505	9515	9525	9535	9545
1.7	0.9554	9564	9573	9582	9591	9599	9608	9616	9625	9633
1.8	0.9641	9649	9656	9664	9671	9678	9686	9693	9699	9706
1.9	0.9713	9719	9726	9732	9738	9744	9750	9756	9761	9767
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2.0	0.9772	9821	9861	9893	9918	9938	9953	9965	9974	9981
3.0	0.9987	9990	9993	9995	9997	9998	9998	9999	9999	9999

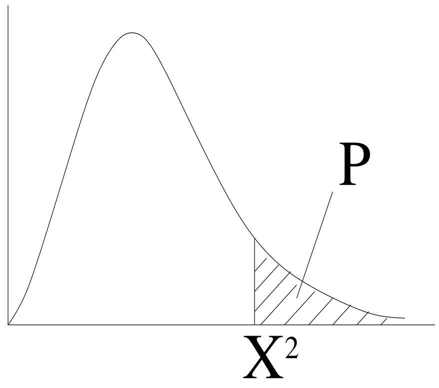
21 **TABLE 2 :** *Table of t*



Probability P of lying outside $\pm t$

d.f.	P=0.10	P=0.05	P=0.02	P=0.01
1	6.31	12.71	31.82	63.7
2	2.92	4.30	6.96	9.93
3	2.35	3.18	4.54	5.84
4	2.13	2.78	3.75	4.60
5	2.02	2.57	3.36	4.03
6	1.94	2.45	3.14	3.71
7	1.90	2.37	3.00	3.50
8	1.86	2.31	2.90	3.36
9	1.83	2.26	2.82	3.25
10	1.81	2.23	2.76	3.17
11	1.80	2.20	2.72	3.11
12	1.78	2.18	2.68	3.06
13	1.77	2.16	2.65	3.01
14	1.76	2.15	2.62	2.98
15	1.75	2.13	2.60	2.95
16	1.75	2.12	2.58	2.92
17	1.74	2.11	2.57	2.90
18	1.73	2.10	2.55	2.88
19	1.73	2.09	2.54	2.86
20	1.73	2.09	2.53	2.85
21	1.72	2.08	2.52	2.83
22	1.72	2.07	2.51	2.82
23	1.71	2.07	2.50	2.81
24	1.71	2.06	2.49	2.80
25	1.71	2.06	2.49	2.79
26	1.71	2.06	2.48	2.78
27	1.70	2.05	2.47	2.77
28	1.70	2.05	2.47	2.76
29	1.70	2.05	2.46	2.76
30	1.70	2.04	2.46	2.75
40	1.68	2.02	2.42	2.70
60	1.67	2.00	2.39	2.66
∞	1.65	1.96	2.33	2.58

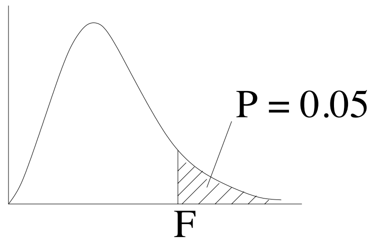
TABLE 3 : *Table of* χ^2



Probability P of a value of χ^2 greater than:

d.f	P=0.05	P=0.01
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.73
12	21.03	26.22
13	22.36	27.69
14	23.68	29.14
15	25.00	30.58
16	26.30	32.0
17	27.59	33.41
18	28.87	34.81
19	30.14	36.19
20	31.41	37.57
21	32.67	38.93
22	33.92	40.29
23	35.17	41.64
24	36.42	42.98
25	37.65	44.31
26	38.89	45.64
27	40.11	46.96
28	41.34	48.28
29	42.56	49.59
30	43.77	50.90
40	55.76	63.69
60	79.08	88.38

22 TABLE 4 : *Table of F for $P = 0.05$*



Variance ratio $F = s_1^2/s_2^2$ with ν_1 and ν_2 degrees of freedom respectively.

ν_1	1	2	3	4	5	6	8	12	24	∞	
ν_2											ν_2
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67	6
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93	8
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54	10
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.69	2.51	2.30	12
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13	14
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01	16
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92	18
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84	20
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62	30
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51	40
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.92	1.70	1.39	60
∞	3.84	3.00	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00	∞

23 TABLE 5 : *Critical values of R for the Mann-Whitney rank-sum test*

The pairs of values below are approximate critical values of R for two-tailed tests at levels $P = 0.10$ (upper pair) and $P = 0.05$ (lower pair). (Use relevant $P = 0.10$ entry for one-tailed test at level 0.05).

		larger sample size, n_2						
		4	5	6	7	8	9	10
smaller sample	4	12,24	13,27	14,30	15,33	16,36	17,39	18,42
size n_1		11,25	12,28	12,32	13,35	14,38	15,41	16,44
	5		19,36	20,40	22,43	23,47	25,50	26,54
			18,37	19,41	20,45	21,49	22,53	24,56
	6			28,50	30,54	32,58	33,63	35,67
				26,52	28,56	29,61	31,65	33,69
	7				39,66	41,71	43,76	46,80
					37,68	39,73	41,78	43,83
	8					52,84	54,90	57,95
						49,87	51,93	54,98
	9						66,105	69,111
							63,108	66,114
	10							83,127
								79,131

24 **TABLE 6 :** *Critical values for T in the Wilcoxon Matched-Pairs Signed-Rank test .*

The values below are the approximate critical values of T for two-tailed tests at level P. For a significant result, the calculated T must be **less than or equal to** the tabulated value. (Values of P are halved for one-tailed tests using R_- and R_+ .)

n	**P = 0.10**	**P = 0.05**
5	**2**	**-**
6	**2**	**0**
7	**3**	**2**
8	**5**	**3**
9	**8**	**5**
10	**10**	**8**
11	**14**	**10**
12	**17**	**13**
13	**21**	**17**
14	**26**	**21**
15	**30**	**25**
16	**36**	**29**
17	**41**	**34**
18	**47**	**40**
19	**53**	**46**
20	**60**	**52**
21	**67**	**58**
22	**75**	**65**
23	**83**	**73**
24	**91**	**81**
25	**100**	**89**