NONPARAMETRIC TESTS II (DA_2022)

Andrea Giansanti

Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

Lecture n. 19, Rome mon 11th of May 2022

Dipartimento di Fisica

Sapienza
Università di Roma

Outline L 19

- Basic distinctions and concepts in inference: one sample/two samples/ Multiple samples; power/significance (type I/type II errors) sample size/ experimental design

- Overwiew of statistical test (see Rosner's general flow-chart p.895;

See also W&S interleaf on p. 465 which test I should use?

- The Wilcoxon rank sum test (Mann-Whitney U test)

The 1 BILLION DOLLAR QUESTION: WHICH TEST I SHOULD USE? (see W&S interleaf n. 7)

List of sub-questions:

- Does your test involve **just one variable**,

or are you testing **the association between two or more variables**?

- Are the variables **categorica**l or **numerical**?

- Are your data paired?

(e.g.**diachronic vs. synchronic** treatment studies)

- What are the assumptions of the tests, and do your data meet those assumptions? (**Normality/non-normality**)

TABLE 1 Commonly used statistical tests for data on a single variable. These methods test whether a population parameter equals the value proposed in the null hypothesis or whether a specific probability model fits a frequency distribution. (Red numbers in parentheses refer to the chapter that discusses the test. Some refer to future chapters.)

Data type	Goal	Test
Categorical	Use frequency data to test whether a population proportion equals a null hypothesized value	Binomial test (7) χ^2 Goodness-of-fit test with two categories (used if sample size is too large for the binomial test) (8)
	Use frequency data to test the fit of a specific population model	χ^2 Goodness-of-fit test (8)
Numerical	Test whether the mean equals a null hypothesized value when data are approximately normal (possibly only after a transformation) (13)	One-sample *t*-test (11)
	Test whether the median equals a null hypothesized value when data are not normal (even after transformation)	Sign test (13)
	Use frequency data to test the fit of a discrete probability distribution	χ^2 Goodness-of-fit test (8)
	Use data to test the fit of the normal distribution	Shapiro-Wilk test (13)

Most hypothesis tests are carried out to determine whether two variables are associated or correlated. This question can be addressed when the two variables are both categorical, both numerical, or when there is one of each. Table 2 lists the most common tests used for each combination when the appropriate assumptions are met.

TABLE 2 Commonly used tests of association between two variables. (Red numbers in parentheses refer to the chapter that discusses the test.)

		Type of explanatory variable	
		Categorical	**Numerical**
Type of response variable	Categorical	Contingency analysis (9)	Logistic regression (17)
	Numerical	*t*-tests, ANOVA, Mann-Whitney *U*-test, etc. [See *Table 3 for more defails.*]	Linear and nonlinear regression (17) Linear correlation (16) Spearman's rank correlation (when data are not bivariate normal) (16)

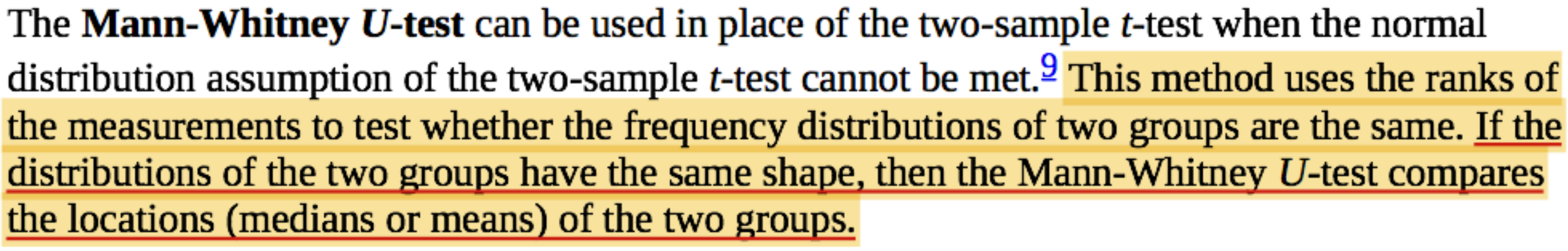
Many methods allow hypothesis tests of differences in a numerical response variable among different groups (see the bottom left corner of Table 2). Testing differences between groups is equivalent to a test of association between a categorical explanatory variable (group) and a response variable. Table 3 summarizes these tests and gives the particular circumstances in which each is used, along with alternatives that make fewer assumptions.

TABLE 3 A comparison of methods to test differences between group means according to whether the tests assume normal distributions. (Red numbers in parentheses refer to the chapter that discusses the test.)

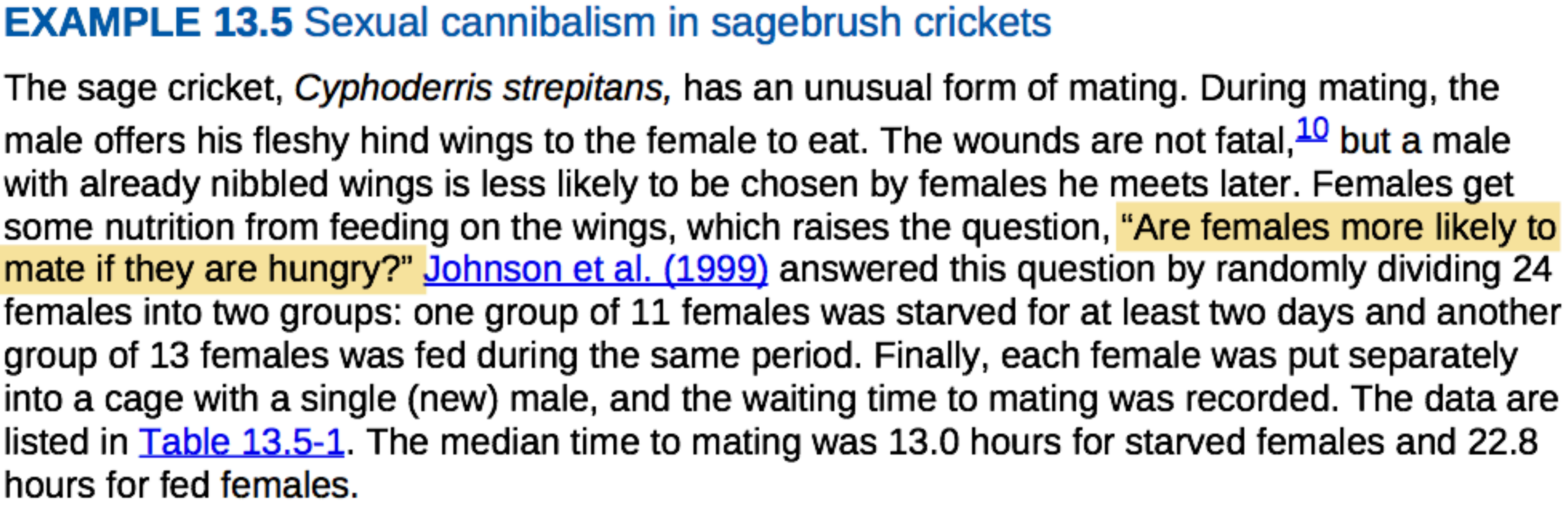
Number of treatments	**Tests assuming normal distribution**	**Tests not assuming normal distributions**
Two treatments (independent samples)	Two-sample *t*-test (12) Welch's *t*-test (used when variance is unequal in the two groups) (12)	Mann-Whitney *U*-test (13)
Two treatments (paired data)	Paired *t*-test (12)	Sign test (13)
More than two treatments	ANOVA (15)	Kruskal-Wallis test (15)

Comparing two groups: the Mann-Whitney *U*-test

RULE OF THUMB: Use the Wilcoxon signed-rank test when you'd like to use the paired *t*–test, but the differences are severely non-normally distributed.



The **Mann-Whitney *U*-test** can be used in place of the two-sample *t*-test when the normal distribution assumption of the two-sample *t*-test cannot be met.[9] This method uses the ranks of the measurements to test whether the frequency distributions of two groups are the same. If the distributions of the two groups have the same shape, then the Mann-Whitney *U*-test compares the locations (medians or means) of the two groups.



EXAMPLE 13.5 Sexual cannibalism in sagebrush crickets

The sage cricket, *Cyphoderris strepitans*, has an unusual form of mating. During mating, the male offers his fleshy hind wings to the female to eat. The wounds are not fatal,[10] but a male with already nibbled wings is less likely to be chosen by females he meets later. Females get some nutrition from feeding on the wings, which raises the question, "Are females more likely to mate if they are hungry?" Johnson et al. (1999) answered this question by randomly dividing 24 females into two groups: one group of 11 females was starved for at least two days and another group of 13 females was fed during the same period. Finally, each female was put separately into a cage with a single (new) male, and the waiting time to mating was recorded. The data are listed in Table 13.5-1. The median time to mating was 13.0 hours for starved females and 22.8 hours for fed females.

From: Michael C. Whitlock and Dolph Schluter - The Analysis of Biological Data-W. H. Freeman and Company (2015) chap 13

TABLE 13.5-1 Times to mating (in hours) for female sagebrush crickets that were recently starved or fed. The measurements of fed females are in red to facilitate comparison after ranking (see Table 13.5-2).

Starved	Fed
1.9	1.5
2.1	1.7
3.8	2.4
9.0	3.6
9.6	5.7
13.0	22.6
14.7	22.8
17.9	39.0
21.7	54.4
29.0	72.1
72.3	73.6
	79.5
	88.9

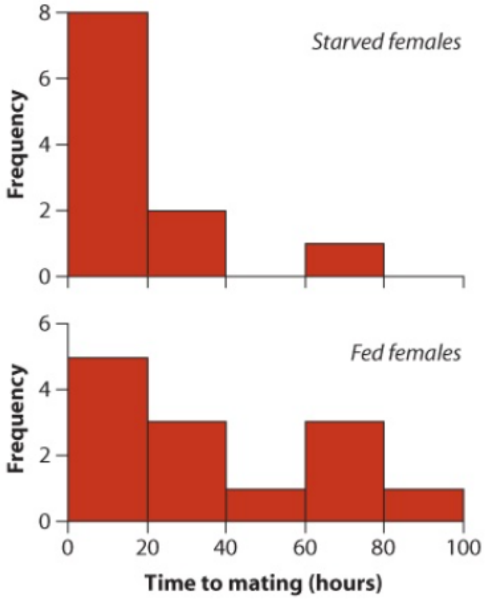


Figure 13.5-1
Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

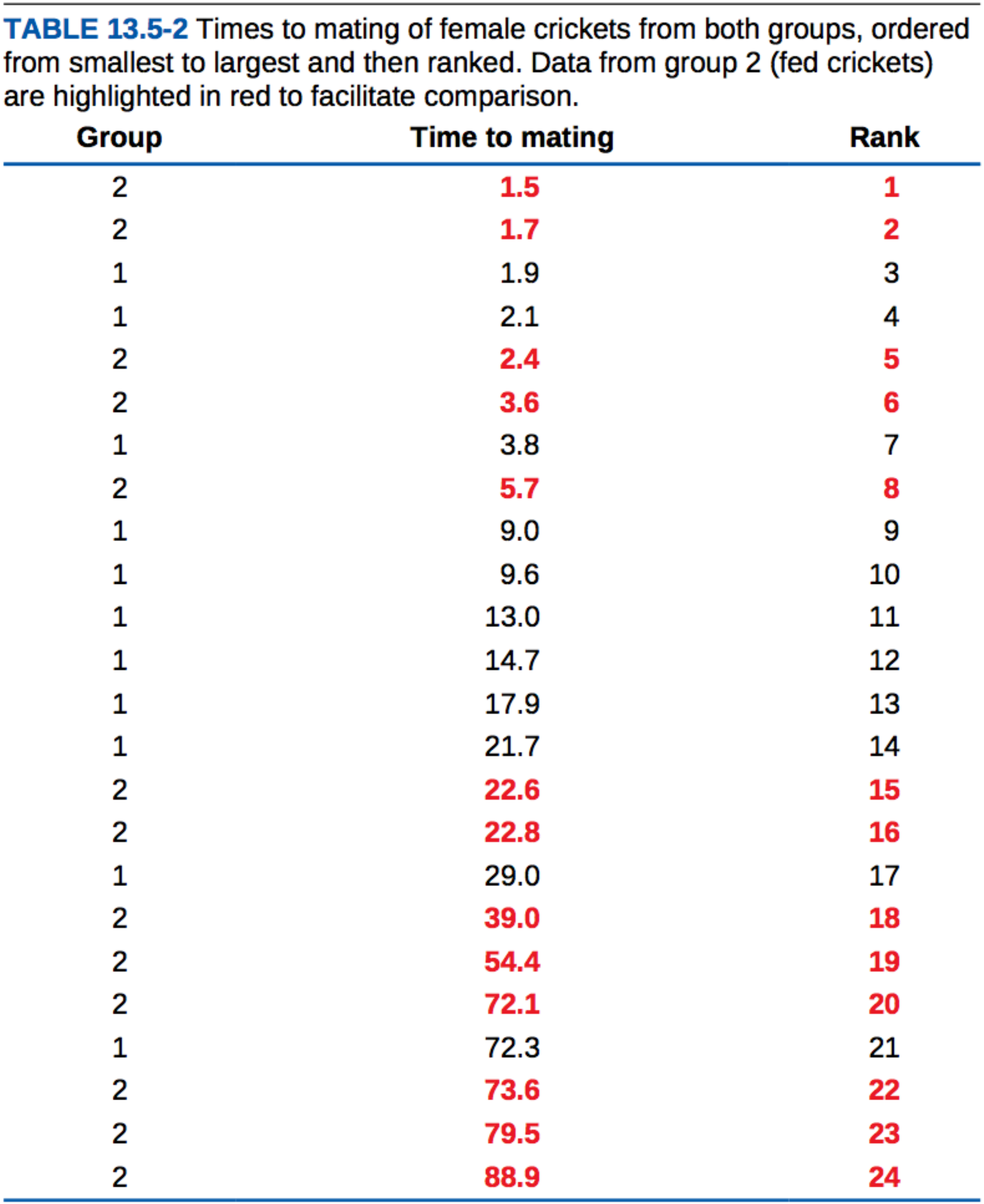


TABLE 13.5-2 Times to mating of female crickets from both groups, ordered from smallest to largest and then ranked. Data from group 2 (fed crickets) are highlighted in red to facilitate comparison.

Group	Time to mating	Rank
2	**1.5**	**1**
2	**1.7**	**2**
1	1.9	3
1	2.1	4
2	**2.4**	**5**
2	**3.6**	**6**
1	3.8	7
2	**5.7**	**8**
1	9.0	9
1	9.6	10
1	13.0	11
1	14.7	12
1	17.9	13
1	21.7	14
2	**22.6**	**15**
2	**22.8**	**16**
1	29.0	17
2	**39.0**	**18**
2	**54.4**	**19**
2	**72.1**	**20**
1	72.3	21
2	**73.6**	**22**
2	**79.5**	**23**
2	**88.9**	**24**

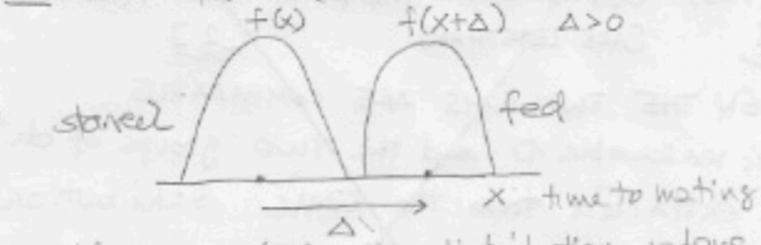
RANKING THE DATA
FROM BOTH GROUPS

DA2020 L#22

NOTE ON WILCOXON RANK SUM NON PARAMETRIC TEST
(AKA MANN-WHITNEY U-TEST)

Consider the data from Johnson 1999 referred to in W&S chap13 'Cannibalism in crickets' (sexual)
The objective of the test is to show that the two samples either came from the same distribution (H_0) or came from a SHIFTED DISTRIBUTION (H_1)

FIG.0



In the specific case study the distribution refers to time to mating with the idea that fed cricket females should have a delay at mating w.r. to starved females.
The data are shown in clides 8 & 9 of this presentation. Thus :

* $H_0 : F_S = F_F$ $H_1 : F_S(x) = F_F(x+\Delta)$
where F(x) is the cumutative distribution function related to f(x) where x is mating time.

This is the typical test on 'location' of 2 distribution if f(x) were normal we would do the TWO sample t-test

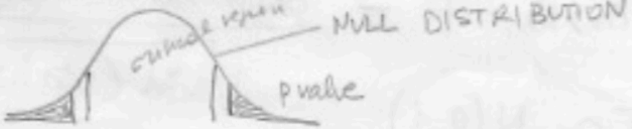
FURTHER NOTE ON THE ESTIMATES OF RANKSUMS

THIS IS THE SETTING IN WHICH AGAIN WE HAVE A CONTROL SITUATION, WE PERTURB THE SITUATION AND WE WOULD LIKE TO ASSESS WHETHER THE PERTURBATION HAS SHIFTED THE DISTRIBUTION IN A STAT. SIGNIFICANT WAY.

- IN A NON PARAMETRIC TEST THE STATISTIC TO BE CONFRONTED WITH THE NULL DISTRIBUTION FOR ESTREMALITY IS BASED ON RANKS

- FIRST STEP RANKING
  COMBINE DATA FROM THE TWO GROUPS OF OBSERVATIONS INTO ONE RANKED LIST FROM THE LOWEST TO THE HIGHEST VALUE

- COMPUTE THE SUM OF RANKS RELATIVE T TO THE SMALLEST GROUP AND THEN CHECK FOR ESTREMALITY

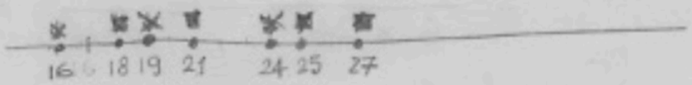


- COMPARE THE OBSERVED T VALUE WITH THE DISTRIBUTION OF ALL THE POSSIBLE SUMS OF RANKS TO CHECK FOR ESTREMALITY IN OTHER WORDS TO CHECK FOR THE PROBABILITY OF GETTING A T VALUE AS EXTREME AS THE ONE OBSERVED UNDER THE NULL IPOTHESIS

- IN THE CASE OF SMALL SAMPLES THIS CAN BE DONE COMBINATORIALLY (BY HAND OR USING TABLES

- EXAMPLE PLACEBO / TREATMENT SCORE

P	rank	T	rank
19	3	21	4
24	5	27	7
16	1	18	2
		25	6



$$T = 1 + 3 + 5 = 9$$

- what is the distubution of sums of ranks separating 7 ordered into two groups at random

PRODUCE A TABLE OF POSSIBLE SUMS

RANKS							SUMS VARIABLE
1	2	3	4	5	6	7	
x	x	x					6
x	x		x				7
x	x			x			8
x	x				x		9
x	x					x	10
x		x	x				8
x		x		x			9
x		x			x		10
x		x				x	11

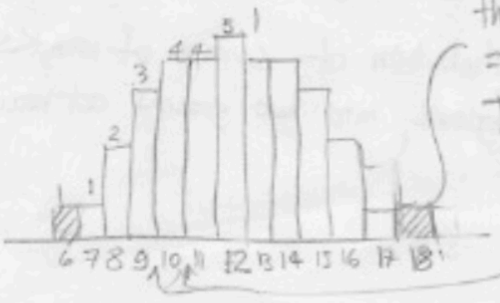
35 righe così

# of way of getting 3 objects from a set of 7

$$\binom{7}{3} = \frac{7!}{3!(7-3)!}$$

$$= \frac{1\times2\times3\times4\times5\times6}{6\times24} = \frac{5040}{6\times24} = 35$$

Make the histogram



this area is 0,057
= P value

The observed value
T = 9
IS NOT SO
EXTREMAL

Under the null hypothesis,
so we reject the alternative hypothesis that the treatment was effective

if the samples are more numerous than 10 then the T distribution can be approximated with a Gaussian

with $\mu_T = n_1 \underbrace{\frac{(n_1+n_2+1)}{2}}_{\text{average rank}}$

and $\sigma_T = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$

and look at tables for $Z_T = \frac{T-\mu_T}{\sigma_T}$

In the case of Crickets

$T = 3+4+7+9+10+11+12+13+14+17+21 = 121$

$n_1 = 11 \quad n_2 = 13 \quad n_1 + n_2 = 24 \qquad n > 10$

$\mu_T = 11(11+13+1)/2 = \frac{11 \times 25}{2} = 137{,}5$

$\sigma_T = \sqrt{\frac{11 \cdot 13 (25)}{12}} = 17{,}26$

$Z = \frac{|121 - 137{,}5| - 0{,}5}{17{,}26} = -0.956$

$p = 2 \times [1 - \Phi(-0{,}956)] = 0{,}17$ big!

let us use the statistical table –

following the U - variable (WRS)

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 88$$

$$U_2 = n_1 n_2 - U_1 = 55$$

$U = 88$ look for $U_{0,05, 11, 13} = 106$

Since 88 is less than the critical value The null hypothesis ~~that~~ should not ~~as rank sum~~ be rejected since a value as big or bigger than what is observed can be obtained by chance with a probability $> 0,05$

$> 0,05$!

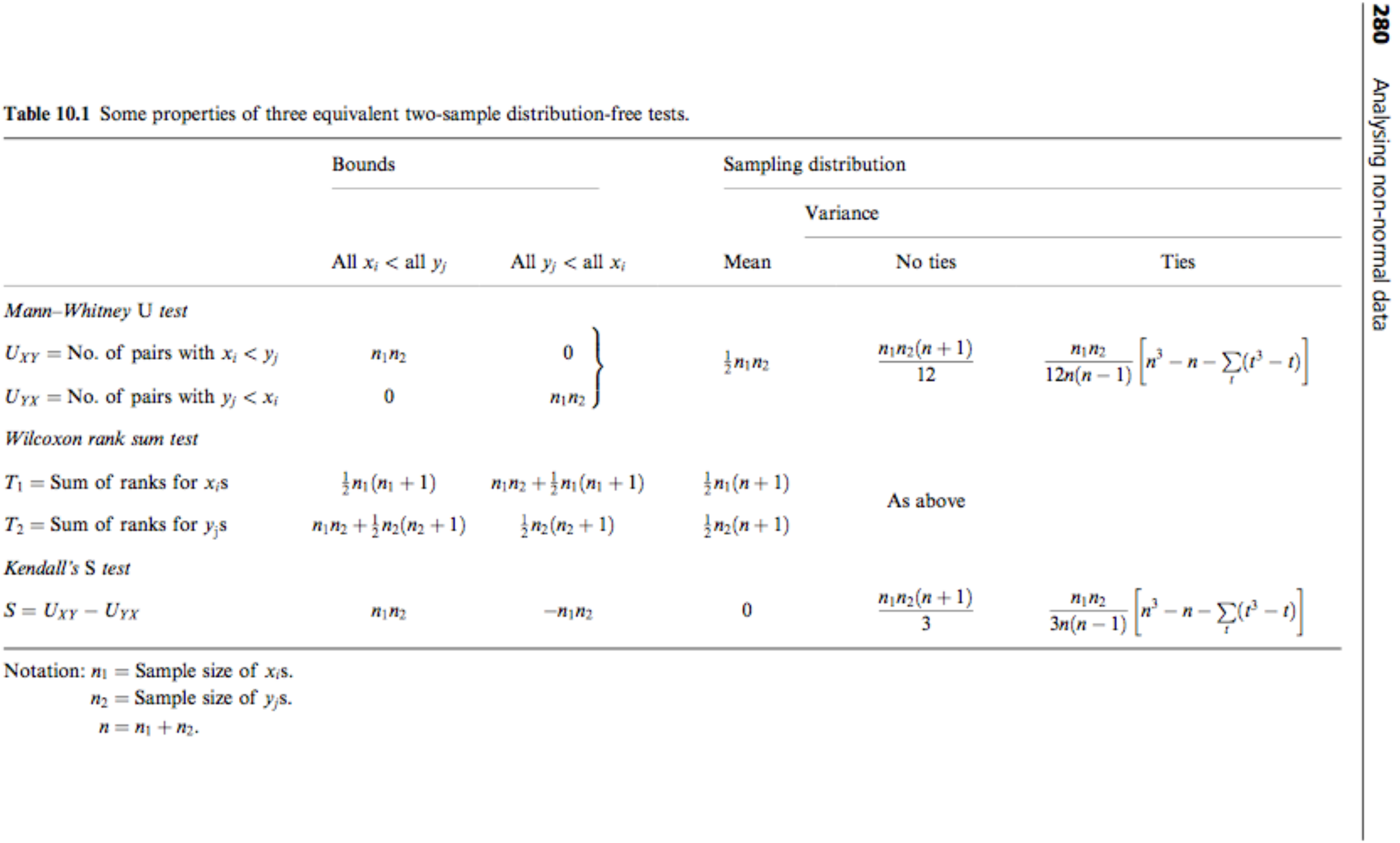


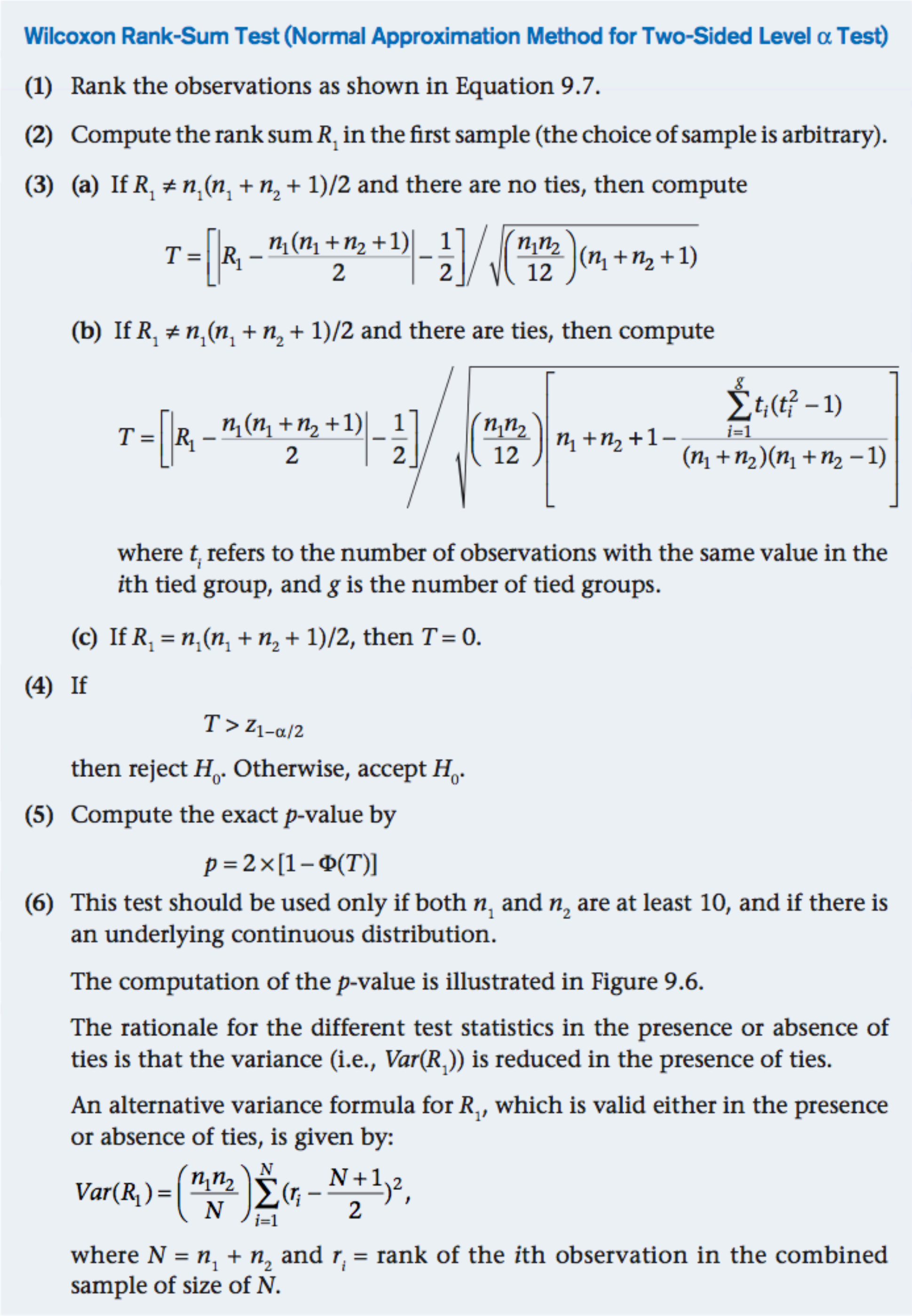
Table 10.1 Some properties of three equivalent two-sample distribution-free tests.

	Bounds		Sampling distribution		
				Variance	
	All $x_i <$ all y_j	All $y_j <$ all x_i	Mean	No ties	Ties
Mann–Whitney U test					
U_{XY} = No. of pairs with $x_i < y_j$	$n_1 n_2$	0	$\frac{1}{2} n_1 n_2$	$\frac{n_1 n_2 (n+1)}{12}$	$\frac{n_1 n_2}{12n(n-1)}\left[n^3 - n - \sum_t (t^3 - t)\right]$
U_{YX} = No. of pairs with $y_j < x_i$	0	$n_1 n_2$			
Wilcoxon rank sum test					
T_1 = Sum of ranks for x_is	$\frac{1}{2} n_1 (n_1 + 1)$	$n_1 n_2 + \frac{1}{2} n_1 (n_1 + 1)$	$\frac{1}{2} n_1 (n + 1)$	As above	
T_2 = Sum of ranks for y_js	$n_1 n_2 + \frac{1}{2} n_2 (n_2 + 1)$	$\frac{1}{2} n_2 (n_2 + 1)$	$\frac{1}{2} n_2 (n + 1)$		
Kendall's S test					
$S = U_{XY} - U_{YX}$	$n_1 n_2$	$-n_1 n_2$	0	$\frac{n_1 n_2 (n+1)}{3}$	$\frac{n_1 n_2}{3n(n-1)}\left[n^3 - n - \sum_t (t^3 - t)\right]$

Notation: n_1 = Sample size of x_is.
n_2 = Sample size of y_js.
$n = n_1 + n_2$.

From:Armitage, StatisticalMethodsInMedicalResearch

Continuity correction



Wilcoxon Rank-Sum Test (Normal Approximation Method for Two-Sided Level α Test)

(1) Rank the observations as shown in Equation 9.7.

(2) Compute the rank sum R_1 in the first sample (the choice of sample is arbitrary).

(3) (a) If $R_1 \neq n_1(n_1 + n_2 + 1)/2$ and there are no ties, then compute

$$T = \left[\left|R_1 - \frac{n_1(n_1+n_2+1)}{2}\right| - \frac{1}{2}\right] \Big/ \sqrt{\left(\frac{n_1 n_2}{12}\right)(n_1+n_2+1)}$$

(b) If $R_1 \neq n_1(n_1 + n_2 + 1)/2$ and there are ties, then compute

$$T = \left[\left|R_1 - \frac{n_1(n_1+n_2+1)}{2}\right| - \frac{1}{2}\right] \Bigg/ \sqrt{\left(\frac{n_1 n_2}{12}\right)\left[n_1+n_2+1-\frac{\sum_{i=1}^{g} t_i(t_i^2-1)}{(n_1+n_2)(n_1+n_2-1)}\right]}$$

where t_i refers to the number of observations with the same value in the ith tied group, and g is the number of tied groups.

(c) If $R_1 = n_1(n_1 + n_2 + 1)/2$, then $T = 0$.

(4) If

$$T > z_{1-\alpha/2}$$

then reject H_0. Otherwise, accept H_0.

(5) Compute the exact p-value by

$$p = 2 \times [1 - \Phi(T)]$$

(6) This test should be used only if both n_1 and n_2 are at least 10, and if there is an underlying continuous distribution.

The computation of the p-value is illustrated in Figure 9.6.

The rationale for the different test statistics in the presence or absence of ties is that the variance (i.e., $Var(R_1)$) is reduced in the presence of ties.

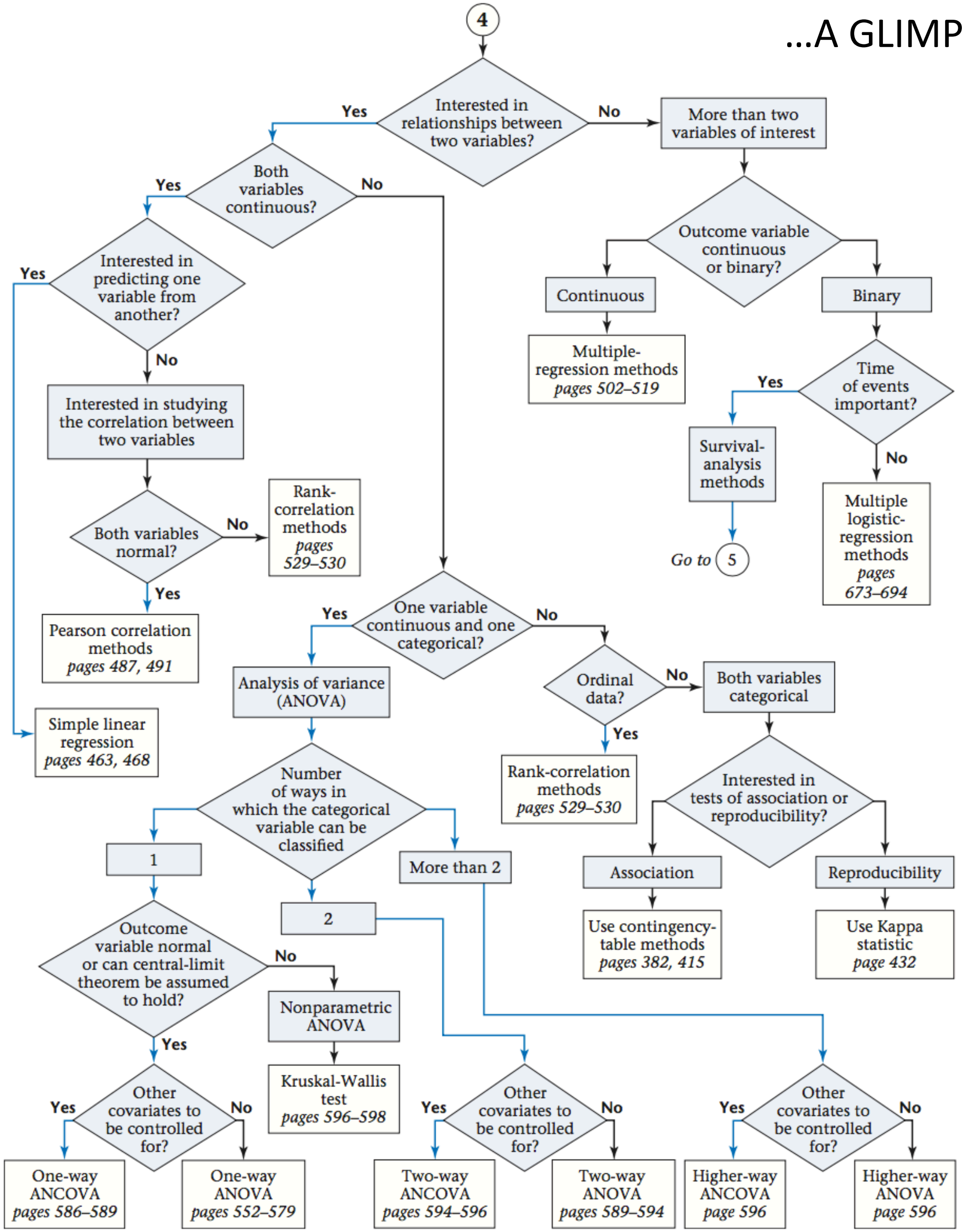
An alternative variance formula for R_1, which is valid either in the presence or absence of ties, is given by:

$$Var(R_1) = \left(\frac{n_1 n_2}{N}\right)\sum_{i=1}^{N}(r_i - \frac{N+1}{2})^2,$$

where $N = n_1 + n_2$ and r_i = rank of the ith observation in the combined sample of size of N.

…A GLIMPSE TO THE JUNGLE

FIGURE 11.33 Flowchart for appropriate methods of statistical inference


4
Interested in relationships between two variables?
Yes
No
More than two variables of interest
Both variables continuous?
Yes
No
Outcome variable continuous or binary?
Continuous
Binary
Multiple-regression methods pages 502–519
Time of events important?
Yes
No
Survival-analysis methods
Go to 5
Multiple logistic-regression methods pages 673–694
Interested in predicting one variable from another?
Yes
No
Interested in studying the correlation between two variables
Both variables normal?
No
Yes
Rank-correlation methods pages 529–530
Pearson correlation methods pages 487, 491
Simple linear regression pages 463, 468
One variable continuous and one categorical?
Yes
No
Analysis of variance (ANOVA)
Ordinal data?
No
Yes
Both variables categorical
Rank-correlation methods pages 529–530
Interested in tests of association or reproducibility?
Association
Reproducibility
Use contingency-table methods pages 382, 415
Use Kappa statistic page 432
Number of ways in which the categorical variable can be classified
1
2
More than 2
Outcome variable normal or can central-limit theorem be assumed to hold?
No
Yes
Nonparametric ANOVA
Kruskal-Wallis test pages 596–598
Other covariates to be controlled for?
Yes
No
One-way ANCOVA pages 586–589
One-way ANOVA pages 552–579
Other covariates to be controlled for?
Yes
No
Two-way ANCOVA pages 594–596
Two-way ANOVA pages 589–594
Other covariates to be controlled for?
Yes
No
Higher-way ANCOVA page 596
Higher-way ANOVA page 596