Statistiek I

Nonparametric Tests

John Nerbonne

CLCG, Rijks*universiteit* Groningen

http://www.let.rug.nl/nerbonne/teach/Statistiek-I/

Overview

1. Mann-Whitney U-Test
2. Wilcoxon's Signed Rank Test
3. Reporting Statistical Analyses
4. Models, Reflections

Nonparametric Tests

NONPARAMETRIC, DISTRIBUTION-FREE Tests —aren't summarized as parameters to distributions, i.e. $N(0,1), t(18), F(3,36)$ or $B(10,0.3)$

- applied when distribution unknown
  & when dist. violates condition of parametric test
- often best option for nonnumeric data
- less sensitive than parametric tests!
- χ^2 is also non-parametric
- several popular tests
  - Mann-Whitney (U-Test)—like *t*-test
    Kruskal-Wallis ($>$ 2 groups) —where dist. not normal (but still symm.)
  - Wilcoxon Signed-Rank Test—like paired t-test where dist. not normal (but still symm.)
  - Sign Test—where asymmetry possible

Mann-Whitney U-Test

alternative to *t*-test (independent samples)

- applicable to ordinal data
- compares two samples
- tests H_0 : samples from same population
  vs. H_a : samples from diff. populations
- alternative to independent sample *t*-test
- gives same results as Wilcoxon's Rank Sum test (M&M)
- example: SSHA (Survey of Study Habits & Attitudes) compares men, women on motivation, study habits and attitudes

Women's Scores:	$154, 109, 137, 115, 140, 154, \ldots$
Men's Scores:	$108, 140, 114, 91, 180, 115, \ldots$

(see exercises)

Mann-Whitney U-Test: Example

Women's Scores: 154, 109, 137, 115, 140, 154, . . .
Men's Scores: 108, 140, 114, 91, 180, 115, . . .

Take the combined set $W \cup M$, order it from lowest to highest rank

1	2	3	. . .
91	108	109	. . .
M	*M*	*W*	

Sum the ranks for both groups, $\Sigma M, \Sigma F$

$$\begin{aligned} U_M &= n_M n_W + \tfrac{n_M(n_M+1)}{2} - \Sigma M \\ U_W &= n_M n_W + \tfrac{n_W(n_W+1)}{2} - \Sigma W \end{aligned}$$

Mann-Whitney U-Test: Definition

Sum the ranks for both groups, $\Sigma M, \Sigma W$

Use smaller of U_1, U_2 (here U_M, U_W), call it U

Note: if distribution is skewed, this will tend to be small (sum of ranks will be large)

Test often applied to Likert ['lI.k@rt] data, i.e. of the form

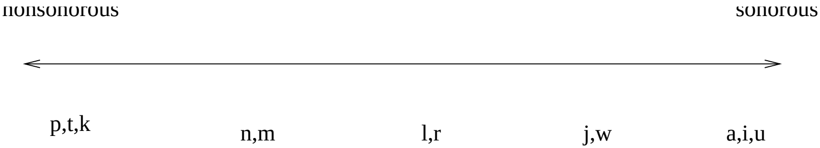
`On a scale of` 1 `(easiest)` −7 `(hardest), the difficulty of this sheet is` ......

Generalization to several groups: **Kruskal-Wallis**

Mann-Whitney — Example

Bastiaanse, Gilbers, v/d Linde 'Sonority Substitutions in Broca's & Conduction Aphasia' *J.Neurolinguistics* 8(4), '94

Sonority scale: phonological **not** phonetic notion



nonsonorous

sonorous

p,t,k

n,m

l,r

j,w

a,i,u

Sonority substitution: one that replaces a segment, changing the sonority, e.g. /pIn/ → /pIt/

Bastiaanse et al.'s Use of Mann-Whitney

- background hypothesis: conduction aphasia has more to do with higher levels of linguistics organization
- expectation: errors involving change in sonority indicate phonological problems
- therefore we expect more sonority errors in conduction aphasia than Broca's aphasia
- H_0: about the same proportion in both aphasia's
- looks like *t*-test, but distribution not normal, therefore Mann-Whitney test
- result: confirmation of alternative hypothesis (more sonority substitutions in conduction aphasia)

Mann-Whitney

- useful fallback for *t*-test for independent samples
- no applicability to single-sample situations, paired data

Wilcoxon’s Signed Rank Test

- like t-test, applicable to single samples and paired samples!
- normally applied to numeric data outside normal dist.
- numeric data is translated into ranked, signed data
- distribution should be roughly symmetric, not skewed
  —since hypothesis is about mean μ
- potentially applicable to pure rankings
  —need to rank differences

Wilcoxon Applied to Single Sample

Translation into ranked, signed data

Example: test reports claim $\mu = 92$ (for dyslexics) on test of dyslexia. You suspect that 92 is too high and arrange to have it administered to 10 randomly chosen dyslexics.

$$\begin{aligned} H_0 &: \mu = 92 \\ H_a &: \mu < 92 \end{aligned}$$

Results:

78	95	84	70	96
73	87	85	76	94

Wilcoxon Calculations

Convert the data

col 2 convert to ± diff. to μ

col 4 rank **unsigned** data

col 5 add signs to ranks

Score	**Diff.**		**Rank**	**Signed**
				Rank
x	$\delta = x - \mu$	$\lvert\delta\rvert$	**of**$\lvert\delta\rvert$	r
78	-14	14	7	-7
95	3	3	2	2
84	-8	8	6	-6
70	-22	22	10	-10
96	4	4	3	3
73	-19	19	9	-9
87	-5	5	4	-4
85	-7	7	5	-5
76	-16	16	8	-8
94	2	2	1	1

W, the test statistic, is the sum of **positive** ranks (here, $W = 6$)

Wilcoxon *p*-Values

$$H_0\colon \mu = 92 \qquad \textbf{and} \qquad H_a\colon \mu < 92$$

If H_0 is true, then positive and negative magnitudes should be roughly the same, i.e.

$$\frac{1}{2}\sum_{i=1}^{n} i, \text{where } n \text{ is the size of the sample}$$

Refer to tables (or SPSS or S+) for critical values of W

$$P(W_{10} \leq 8) = 0.025$$

W_{10} since the prob. of W depends on sample size n
This is **one-tailed** prob. —since hypothesis is one-tailed.

Wilcoxon in Two-Sided Hypotheses

$p = 0.025$ low enough to reject $H_0: \ \mu = 92$ in favor of one-tailed $H_a: \ \mu < 92$

If we'd examined two-sided $H'_a: \ \mu \neq 92$, then we should have obtained:

$$P(W_{10} \leq 8) = 0.05$$

naturally, less strong **against** H_0.

Probabilities of W

	2-tailed Signif.	
	0.05	0.01
	1-tailed Signif.	
N	0.025	0.005
6	0	—
7	2	—
8	9	0
9	5	1
10	8	3
11	10	5
12	13	7
13	17	9
14	21	12
15	25	15
⋮	⋮	⋮
20	52	37
⋮	⋮	⋮
25	89	68

i.e. $P(W_{10} < 8) = 0.05$ in 2-sided hypothesis

Wilcoxon vs. t-test

Sometimes, tables list only small positive values, but right skewing results in large positive value

—To test hypothesis of right skew, use magnitude of sum of negative ranks

To compare mean in single sample of unknown σ (to some hypothesis), use the *t*-test

- **unless** population symmetric but not normal, e.g., some bimodal distributions
  then use Wilcoxon
- what if the population is non-normal **and** asymmetric?
  —*t* test OK, but use large sample, > 100)
  —recall, too, that *t* test differs little from *z* test (with sd $= \sigma$) if $n > 100$

Paired Samples in Wilcoxon

Wilcoxon also used as substitute for paired-sample *t*-test

Example: S+ exercise (French Listening Test before and after course)

Person	**Before**	**After**
1	32	34
2	31	31
⋮	⋮	⋮
20	23	26

(Assumption: dist. non-normal)

$H_0 : \mu_b = \mu_a$ **(no diff.)**; $H_a : \mu_b < \mu_a$ **(improvement)**

1. Calculate $\delta_i = t_{ia} - t_{ib}$, convert this to signed ranks (as above), etc.
2. Use $\mu_{\delta_i} = 0$ as H_0, $\mu_{\delta_i} > 0$ as H_a, treat as single sample.
3. See laboratory exercise.

Sign Test

When all else fails ... **sign test** (use PROPORTIONS, M&M, § 5.1, other sheets)

- divides data into classes $+, -$ and 0 (only)
  e.g. positive, negative, and no change
- use: when dist. nonnormal, asymmetric
- compares proportion of positive to negative
- tests whether division is roughly chance-like
  H_0 : no weighting toward $+$ (or $-$), no change
- More next week (on proportions)

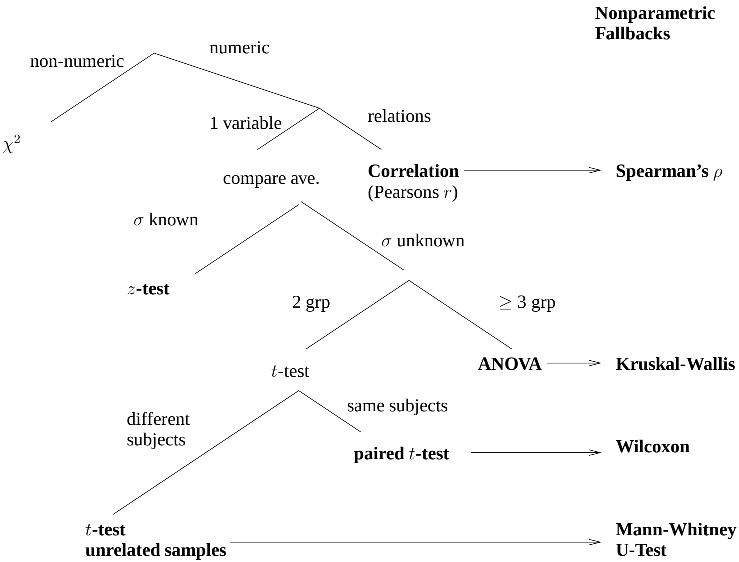
example

1. 22 aphasics judged subjectively (as belonging to one of two categories)
   question: are the judgements roughly similar?
   method: count same as $+$, different as $-$

Nonparametric Tests

NONPARAMETRIC, DISTRIBUTION-FREE Tests

- applied when distribution unknown
  & when dist. violates condition of parametric test
- often best option for non-numeric data
- less sensitive than parametric tests!
- several easy, useful tests
  - Mann-Whitney (U-Test)—for indep. sample t-test
    Kruskal-Wallis—allows $>$ 2 groups
    —assumes symmetry, but not normal dist.
  - Wilcoxon Signed-Rank Test—for paired t-test
    —assumes symmetry, not normal dist.
  - Sign Test—when all else fails



Nonparametric Fallbacks

non-numeric

numeric

χ^2

1 variable

relations

compare ave.

Correlation (Pearsons r) → **Spearman's ρ**

σ known

σ unknown

z-**test**

2 grp

$\geq$ 3 grp

t-test

ANOVA → **Kruskal-Wallis**

different subjects

same subjects

paired t-test → **Wilcoxon**

t-**test unrelated samples** → **Mann-Whitney U-Test**

Reporting Statistics

From the third lab:

It is suspected that the Spanish language proficiency of social workers in larger cities is different from that of social workers from smaller cities and towns (simply due to their different exposure to the language). Your agency wishes to test this since training programs may differ depending on proficiency levels. You obtain data from twenty social workers, ten from each group, and you wish to test whether the groups are different.

Hypotheses? One-sided or two-sided?

Hypotheses

We compare cities (c) and non-cities ($\not{c}$).
$H_0 : \mu_c = \mu_{\not{c}}$
$H_a : \mu_c \neq \mu_{\not{c}}$

This is obviously two-sided.

Comment: The way the problem was stated was two-sided. It would also be natural to asked a one-sided question, namely, is the group from the cities (with more exposure) better? **One needs to read problem statements carefully.**

How to test?

How to test?

We test a hypothesis about differences in means in two groups using a *t*-test for independent samples.

Conditions (assumptions)?

Assumptions?

Since the samples are small, they need to be roughly normal for the *t*-test to yield valid results.

How can you test normality?

Normality?

Test normality using normal quantile plot.

(Display in student report. See Lab 3.)

Results?

$m_c = 28.4, m_{\not{c}} = 26.2$
sd ≈ 5
$t(9) = -0.98$
$p \leq 0.34$

How to report?

Report

We suspected that the Spanish language proficiency of social workers in larger cities is different from that of social workers from smaller cities and towns (simply due to their different exposure to the language). We wished to test this since training programs may differ depending on proficiency levels. We obtained data from twenty randomly selected social workers, ten from each group, verified that the samples were roughly normally distributed, and tested whether the groups differed in means, obtaining $t(9) = -0.98, p \leq 0.34$. The more urban group was only 0.4 sd better. We retained the null hypothesis that the groups do *not* differ.

(but we note that the difference in standard deviation would be significant in a samples of thirty each).

Statistics in Research

Research Article/ Honor's Thesis
Background

- explain background theory clearly, consistently
  minimal wrt deriving testable prediction
- explain novelty
  - genuine novelty
- derive testable predictions
  identify auxiliary assumptions
- if another theory is contrasted
  - be fair
  - show contrast in testable predictions
- summarize relevant earlier studies

Population/Sample

Design

- be clear on how theory is related to test
- describe population, relation to sample, size of sample
- note use of volunteers, drop outs
- use a control group (if possible)
    - assign subjects to control randomly

Reporting Statistical Analysis

Analysis

- make data available (discuss w. advisor, website)
- examine data w. descriptive statistics, tables, graphics
- justify choice of test
- show that requirements met, e.g., normal dist.
- note significance level

Discuss Results

Conclusions

- interpret results esp. vis-a-vis theory
- discuss "failed" hypotheses, too
- be sensitive to size of result vs. significance
- discuss alternative explanations
- sketch further questions

Discuss Results

Some distinctions in Field not yet in the lectures.

- dependent vs. independent variables
  outcome vs. predictor variables
- statistical "models"
- systematic vs. unsystematic variation
- studies vs. experiments

Dependent vs. independent variables

... or OUTCOME vs. PREDICTOR variables

- often a matter of perspective!
- use "web-site design" to predict number of visitors
- use shopping basket analysis to predict other needs
  diapers in basket → baby oil, baby food
  music purchased in iTunes → future purchases
- use postal code to predict income, receptiveness for advertising for expensive commodities

Statistical “Models”

- Statistical models simple
  Unlike e.g. the Ptolemaic (geocentric) model of the solar system! Or Copernican model.
- in t-test for independent samples
  - model $\propto$ two different means (assumed in H_a)
  - alternative model $\propto$ single mean (see H_0)
- in Mann-Whitney U-test, model $\propto$ different medians
- More elaborate models in second course!

Systematic vs. unsystematic variation

- Think of the variation in two samples of participants in a study on communication preferences (digital vs. traditional)
- The groups differ in age: Under 30 vs. 31 and older
  - Systematic variation is that is explained in a model, e.g. a difference of 20% in preference for digital news media (Cohen's $d = 0.5$sd
  - Unsystematic variation is all the rest, e.g. the roughly equal preference for hearing news in radio broadcasts.

Studies vs. experiments

- Studies draw their data from naturally occurring processes
  - Who are the customers of bol.com?
  - Where to they live? How much disposable income do they have?
  - Do they react to advertisements in newspapers? Which ones?
- Experiments carefully control potential potentially influential factors in order to study a small number precisely.
  - What colors are preferred in logos?
  - Control background color, size of logo, proportion of figure vs. ground, use in letter heads vs. t-shirts, ...

Next Week

Proportions, Frequencies