607: Principles of Inferential Statistics Comments on/answers to homework on §14 [Non-parametric or "Distribution-free" methods]

IPS4e Q 14.19

Exercise 7.37 (page 520) reports readings from 12 home radon detectors exposed to 105 picocuries per liter of radon:

```
  91.9 97.8 111.4 122.3 105.4  95.0
 103.8 99.6  96.6 119.3 104.8 101.7
```

We wonder if the median reading differs significantly from the true value 105 (i.e. if a machine is just as likely to under- as to over-read)

(a) *Graph the data, and comment on skewness and outliers. A rank test is appropriate.*

> ***There do seem to be a few more on the higher side than the lower side, but it is very difficult with n=12 to judge what would the histogram look like if n were 100 or 1000. A visual test of Normality involves plotting the observed values against their expected value (or expected Z-value) under a Normal distribution with the same mean and SD as observed in the sample.***

(b) *We would like to test hypotheses about the median reading from home radon detectors:*

H_0: median = 105
H_a: median ≠ 105

To do this, apply the Wilcoxon signed rank statistic to the differences between the observations and 105. (This is the one sample version of the test.) What do you conclude?

The departures from 105, in order of (absolute magnitude) are

Rank	1	2	3	4	5	6	7	8	9	10	11	12
	0.2	0.4	1.2	3.3	5.4	6.4	7.2	8.4	10.0	13.1	14.3	17.2
	-	+	-	-	-	+	-	-	-	-	+	+

The sums of the signed ranks are

S+ = 2 + 6 + 11 + 12 = 31 , & S- = 1+3+4+5+7+8+9+10 = 47
(Check: 31 + 47 = 78, and 1+2+... + 12 is 6x13 = 78).

We can focus on one of the 2, say S+. Under the null H that there is no systematic tendency in either direction from the target, the 95% and 99% ranges for S+ are [see excerpt from table] (13,65) and (7,71) respectively ...

Number of **paired** observations showing differences	P = 0.05 (2-sided)	P = 0.01 (2-sided)
12	**13, 65**	**7, 71**

The observed S+ is well within these limits or random variation, and so the data do not provide evidence of any shift (up or down) in the values provided measurements.

If you don't want to carry the specialized table around with you, or wish to be able to calculate the p-value itself, rather than determine where the observed values is with respect to certain landmarks (e.g. 0.05, 0.01, etc.) you could use the Normal approximation (cf. JH or M&M notes)

E[S+ | H_0) = 1/2 of 78 = 39; SD[S+ | H_0) = Sqrt[12(13(25)/12] = 12.7, so that z = (31 - 39)/12.7 , or (31.5-39)/12.7 if use the continuity correction.. Then looking up in the Normal tables the probability of a z-value this or more extreme.

(c) *[added by JH] What is the corresponding p-value if you use a simple sign test ?*

NB: A number of you thought that the formal name of this test is the "Simple sign test". This adjective was added by JH to emphasize that it is the easiest to carry out (and the weakest of the options available). In formal publications, refer to it as "the sign test"

Simply counting signs, and ignoring magnitudes, or even the ranks of the magnitudes, we have 4+ and 8-. If the true distribution of over- and under-readings is symmetric around 105, then the probability of a + reading is 0.5, and so the probability of observing just 4 or fewer is (from the Binomial[12,0.5] distribution) 90/1000 or 0.09, so the probability of this extreme a count, or one more extreme, on either side, is 2 times 0.09 or P = 0.18 ... again, compatible with random (as opposed to systematic) deviation from the target.

IPS4e Q 14.22

Exercise 12.11 presents the following data from a study of the loss of vitamin C in bread after baking:

Condition	Vitamin C (mg/100 g)		Ranks & Sum
Immediately after baking	47.62	49.79	9,10 19
One day after baking	40.45	43.46	7,8 15
Three days after baking	21.25	22.34	5,6 11
Five days after baking	13.18	11.65	3,4 7
Seven days after baking	8.51	8.13	1,2 3

The loss of vitamin C over time is clear, but with only 2 loaves of bread for each storage time we wonder if the differences among the groups are significant.

(**a**) *Use the Kruskal-Wallis test to assess significance, then write a brief summary of what the data show.*

Before doing any test, and especially one that does not take account of time, it is quite obvious what the data show: Vitamin C disappears with time .. the half-life seems to be just less than 3 days.. In this example,

Moore and McCabe are guilty of type IV error...asking the wrong question. A better example, where there is no obvious structure is their exercise 14.24 (a study to see which of four colours best attracts cereal leaf beetles). Virtually all of the exercises in section 14.3 have 3 or more groups that have a natural order.. Even the one 14.32 {a study of iron-deficiency anemia in Ethiopia. The issue is whether Ethiopian food loses more iron when cooked in some types of pots [aluminum, clay and iron} has a-priori structure -- metal versus clay, or iron versus non-iron pots.[1]

But, for the sake of illustration, here is the Kruskal Wallis statistic, which tests the null hypothesis that measurements fluctuate around the <u>same level no matter which day</u>, against the (silly) alternative that they vary in some UNSPECIFIED way .. maybe higher on days 1 and 7, and lower on days 3 and 5, or maybe level until day 7, **or maybe up one day, down the next...**

$$T = \{12/(10(11))\} \times \{19^2 /2 + 15^2/2 + 11^2/2 + 7^2/2 + 3^2/2 \} - 3(10+1) = \underline{8.73}$$

The reference distribution against which this has to be compared is the Chi-Squared distribution with 4 df .. 1 less than the number of groups. It is the Chi-Squared rather than the F distribution because there is no separate estimation of a unit variance (σ^2) the way there is with a traditional ANOVA on the raw data. The Chi-square distribution has only 1 tail, and does not distinguish directions.. so it is already omni-directional.. Also note that the 4 df is a penalty for looking for differences in <u>all</u> directions, even in silly directions or patterns. The critical values of the distribution are 9.5 for alpha=0.05, and 13.2 for alpha = 0.01. So the statistic has a P-value somewhere just greater than 0.05, it is 0.068 in fact, and so not statistically significant at the conventional 0.05 level. This is at variance with your ocular test.. where it is clear that Vitamin C decreases. This is because your are using the time structure.. The correct test here is a test of trend or correlation using either the raw data or the ranks.. either way, it should be a test that looks in one direction with time .. down [or also up (i.e. 2-sided) if you are a skeptic or expect miracles!].

(b) *Because there are only 2 observations per group, we suspect that the common chi-square approximation to the distribution of the Kruskal-Wallis statistic may not be accurate. The exact p-value (from the SAS software) is 0.0011. Compare this with your p-value from (a). Is the difference large enough to affect your conclusion?*

Technically, yes the results are discrepant, but as per above, all of these are focused on the wrong test in the first place.

C**omment** by JH

The above analyses are a good example of the mistake made by the person who looks under the lamppost for his lost keys, even though he believes he lost them at a different place on the street-- just because there is more light under the lamppost! This is not a question of 5 "groups" or "conditions": the so-called groups have a very clear ***time structure***, but the analysis used by M&M does not use this structure (If you interchange the rows (times) you still get the same p-value). Fortunately, the 'signal is strong enough here, and the noise from loaf to same-day loaf small enough that the differences are clear. A better -- more sensitive and focused -- analysis is to measure the trend (slope of regression line) in vitamin C over time -- just as your eye does! It is ***not a question of whether*** bread loses vitamin C, but ***how quickly*** it does. If the question were "after 1 day, is the loss more with certain of 5 ***types*** of bread than others, then a (***global***) Kruskal-Wallis or other statistic might be more appropriate-- but again, it is probably not a question of whether, but of how much.

[The same comment applies whether we are parametric or non-parametric. And JH will be returning to this issue when we study chi-square tests for proportions rather than means.]

HOMEGROWN

2 [from A&B] **Obstetric records of (the mothers of) a group of children who died "suddenly and unexpectedly' (SUD) were compared with those of a group of live 'control' children**. Observations on the duration of the 2nd stage of labour were as follows:

```
S.U.D.   60, 25,  6, 8,  5, <5,   10,  25, 15, 10
Controls 13, 20, 15, 7, 75, 120*, 10, 100,  9, 25, 30
           *: terminated by surgical intervention.
```

a [for students in one of the Epidemiology programmes] Do you agree that this should be called a case-control study? Why, or why not?

YES. Because that's what a case-control study is! The so-called case control study in Q1 (lead levels) is not a case-control study: it is a comparison of children of employees in the lead industry with children of parents employed elsewhere.

b Compute and compare the median duration of labour in each group and evaluate the statistical significance of the difference.

The median of the SUD group is 10 (both the 5th and 6th observations are 10), that of the control group 20 (the 6th out of 11).

[1] Incidentally, the authors of that were Abdulaziz A Adishb, a, Steven A Esreyb, c, d, a, Theresa W Gyorkosf, , d, Johanne Jean-Baptistee and Arezoo Rojhanig, a a School of Dietetics and Human Nutrition, McGill University, St Anne-de-Bellevue, Quebec, Canada, b Jimma Institute of Health Sciences, Ethiopia c UNICEF, New York, NY, USA d Department of Epidemiology and Biostatistics, McGill University Canada e Pharmacokinetics at Phoenix International, Montreal Canada f Division of Clinical Epidemiology, Montreal General Hospital, Montreal, Quebec, Canada g Department of Family and Consumer Sciences, Western Michigan University, MI, USA

Rank Sum Test (Two independent samples):

S.U.D. group has n = 10, control has n = 11; therefore if using tables, we use the (smaller) SUD group in the rank test.

Sum of Ranks in SUD group = **86.5**.

According to the null hypothesis [implying that the ranks are randomly distributed over the two groups, since the two distributions are hypothesized to have the same shape and location], one would expect the sum of ranks in the sample of size 10 to be 10/21-ths of the total of 231 ranks (1 + 2 ... +21).   We observe 86.5 when we expect 110, and ask what is probability of getting 86.5 or less or of getting something this extreme or more extreme on the other side of 110? The mirror image of 86.5 relative to 110 is 110 + (110-86.5) or 134.5 . So we are looking for the tail areas outside of (86.5, 134.5).

The Gaussian approximation works very well here [see on the ch 14 Resources, the remarkably close to Gaussian distribution even in the situation where n1 is as low as 3 and n2 as low as 5]

To use it, we need the SD of W under the null. This is

SD   = sqrt[n1 × n2 × (n1+n2+1) / 12]
          = sqrt[10 × 11 × 22 /12]
          =**14.2**.

So 86.5 corresponds to

z = (86.5 – 110)/14.2 = **–1.65** ,

corresponding to a **lower tail** area of just under 5%, or a **two-sided** P-value of **just under 10%**.

See Resources for Ch 14 (template, using as example attendance for exercise classes) for how to run Rank sum test via SAS PROC NPAR1WAY.

OR Table in my notes shows that the numbers 81 and 139 cut off 95% of the distribution i.e. a **rank sum of <u>81 or less or 139 or more</u> would be significant at the 0.05 level. So the data we observe do not reach this 0.05 level of significance.**

Armitage's table only gives the lower limit (81) and lets you do the "mirror imaging" yourself - if you need to.   Colton gives you both.

Notice that M&M do not make a fuss as to which sample you choose to compute the sum of ranks for. This is because the expected value under H_0 reflects the sample size of the one you choose to be "n1". Naturally, if doing it by hand, to save on summing, you would choose the smaller sample size.

Q4. Medication to prevent acute mountain sickness

a "Those taking acetazolamide reached a higher altitude (11 versus 4 reached the summit)" (abstract).

[treating the outcome as <u>binary</u>] Carry out a statistical test to evaluate the 11 vs. 4 "<u>successes</u> <u>in reaching the summit</u>" ***(i)*** using the pairing ***(ii)*** ignoring the pairing.
[for (i), the information provided is not sufficient, so do the test with each of the possible configurations]

<u>unpaired</u>: 11/12 vs. . 4/12 can be tested with the z-test for 2 proportions
z   = (11/12 - 4/12)/sqrt[(15/24)(9/24){1/12 + 1/12}]
           = 2.95 (or a little less if use continuity correction)
giving a 2-sided P-value of approx. 0.0032.

<u>OR</u> by X^2: X^2 = 24(11*8 – 1*4)/{12*12*15*9} = 8.71 = 2.95^2.

<u>OR</u> by Fisher's exact test

```
Table    Prob.
 3   9
12   0   0.000

 4   8
11   1   0.005

 5   7
10   2   0.040

 6   6
 9   3   0.155

 7   5
 8   4   0.300

 8   4
 7   5   0.300

 9   3
 6   6   0.155

10   2
 5   7   0.040

11   1
 4.  8   0.005  <------observed

12   0
 3.  9   0.000
```

Here, the 1-tail area is 0.005 + 0.000 = 0.005, so the 2-tail P-value is 0.010 [in this case, because the 2 sample sizes are the same the distribution is symmetric].

Notice the P-value is not as extreme with Fisher's exact test.

Paired: If we take the scenario that is less favourable to the treatment, and pair the one failure on treatment with the success ion placebo (like the actual case we had related to us in class today) we would have (going back to the layout in Chapter 8)

(McNemar) Test of equality of proportions (paired):

		Result in Placebo member		
		+	–	Total Sets
Result in Tx member	+	a	b	
	–	c	d	
				12 Sets

In the abstract, the results are reported unmatched. It is often easier to report unmatched. From Figure 2, there are only 2 possible tables, depending on who the "x" who failed is matched with. If it is with an "o" who reached the summit, then a=3, b=8,c=1, d=0; otherwise it is a=4, b=7,c=0, d=1. Either way, its is impressive.

Prob(a 7/0 or more extreme split under H_0:50:50) is

BinomialProb(k=7, n=7, p=0.5) = 0.0078 (table C)

So P-value (2-sided) = 0.0158.

Prob(a 8/1 or more extreme split under H_0:50:50) is

BinomialProb(k=8, n=9, p=0.5)
+
BinomialProb(k=9, n=9, p=0.5) {don't forget the "more extreme"}

= 0.0176 + 0.0020 = 0.0196

So P-value (2-sided) = 0.0392.

Z or X^2 versions of McNemar in Ch 9 notes are large sample approximations, but would still not do that badly here, since Binomial with p=0.5 is symmetric.

Either way(scenario), difference is impressive.

BUT would have to wonder if the subjects became unblinded.

b *[treating the outcome as ordinal]*

"Fig. 2 compares the altitudes reached by subjects taking the drug and those taking the placebo... the drug group showed a striking advantage (Wilcoxon signed rank sum test p < 0.01)" (last paragraph, 2nd page)

Presumably, they carried out a "Wilcoxon signed rank test" for paired data. They call it a "signed rank sum test" ... they used the terminology in A&B's textbook rather than in Bradford Hill's or M&Ms'. It would be better if publications called it the "Wilcoxon test for paired data".

Again, there is a small ambiguity, from the data supplied in the Figure, as to what the 12 pairs of 'altitudes reached' must have been. Try to match what the configuration must have been with the reported p-value.

Again, cannot tell the exact pairings. However, a possible scenario... (see footnote)

say it was x who failed matched with o who succeeded...

Tx	Placebo	Diff	Rank*	Signed
Uhuru	Uhuru	0	--	
Uhuru	Uhuru	0	--	
Uhuru	Uhuru	0	--	
Uhuru	Gillman's	+1	3	+3
Uhuru	Gillman's	+1	3	+3
Uhuru	Gillman's	+1	3	+3
Uhuru	Gillman's	+1	3	+3
Uhuru	B/w Kibo and Gillman's	+2	7	+7
Uhuru	B/w Kibo and Gillman's	+2	7	+7
Uhuru	B/w Kibo and Gillman's	+2	7	+7
Uhuru	Kibo	+3	9	+9
Gillman's	Uhuru	–1	3	–3

$(W^+, W^-) = (42, 3)$

*There are 5 1's so they each get the average of 1,2,3,4,5 = 3.
*There are 3 2's so they each get the average of 6,7,8 = 7.

NOTICE that the distances between points reached are somewhat arbitrary.. changing them to actual metres would not change the ***ranks***.

The Table from Bradford Hill (my notes) says that a (W^+, W^-) split of (5,40) (***or vice versa***) or one more extreme would be significant at the 0.05 level 2 sided. So, since we observed a (42,3) split, P < 0.05. A split of (1,44) would have been significant at the 0.01 level. (technically speaking, these tables are for situations where there are no "ties").

Using Normal Approximation to W+ (see justification by enumeration under Resources for Ch 14), and under the null; hypothesis,

E[W+] = 9*10/4 = 22.5
SD[W+] = sqrt[9*19*19/24] = 8.44 giving z = (42 - 22.5)/8.44 = 2.31

Prob [Z > 2.31] = 0.0107 so P-value (2-sided) = 0.0214

* 4 scenarios are x at Gillman's paired with o at ...

(1) Uhuru (2) Gillman's (3) B/w Kibo and Gillman's (4) Kibo

You can also obtain the signed Rank test from PROC UNIVARIATE in SAS,

```
data a;               output.............
input pairdiff;       Variable=PAIRDIFF
lines;                N              9  Sum Wgts       9
+1                    Mean       1.333  Sum           12
+1
+1                    Std Dev    1.118  Variance   1.25
+1                    Skewness   -0.84  Kurtosis   1.9428
+2
+2                    USS           26  CSS       10
+2                    CV         83.85  Std Mean   0.3726
+3                    T:Mean=0   3.577  Pr>|T|     0.0072
-1                    Num ^= 0       9  Num > 0        8
;
proc univariate;      M(Sign)      3.5  Pr>=|M|  0.039[sign test]
var pairdiff;         Sgn Rank 19.5 Pr>=|S| 0.0234
run;
```

Do not be upset if you cannot do so exactly, since it is not clear whether the authors' p-value is 1- or 2-sided, or how they handled ties, or whether they used exact methods (by enumeration, as I do in the diagonals of the table I worked out and put as a separate item "Wilcoxon Signed Rank Test: by Enumeration" on the web page), or by a Gaussian approximation (with/without correction to variance for ties, or continuity correction)

c "In every pair the partner on acetazolamide had the lower symptom score." (first sentence of third page)

i What value of the Wilcoxon signed rank statistic does this imply? (Think of Gauss!)

All 12 differences were negative

so T– = 1+2+...+12=**78**, T+=0. (M&M use "W")

ii What other non-parametric test is suggested by this statement?

Sign Test: test 12/0 split versus 50:50

ie P-value = 2 x BinomialProb(12/12, p=0.5)
= 2 x 0.0002

d Kilimanjaro vs. Mt Kenya:

We could make two contrasts:

i using the data from the "self paired" crossover: use the data from the two expeditions; compare each person's data from the expedition on which (s)he was taking active treatment with the same person's data from the expedition on which (s)he was taking placebo... a one-sample test (within-person comparisons, a paired t-test with 23 df if we were using parametric tests)

ii using the data from the "matched pairs": use only the data from the Mt. Kilimanjaro expedition; compare each treated person's data with his/her partner's data... again a paired t-test but with only 11 df, and between-person comparisons)

Although contrast (i) looks more powerful statistically (and is the one implied in the title of the paper), why is it the scientifically weaker one of the two in this study?

The acclimatization rendered the second half non-comparable. Also, carry over of drug. etc..

One of you showed the problem in a striking way.. comparing the mean of the values when on placebo on one mountain with mean values when on placebo on the other one.