Two Parameter HT Exercises

Pages	Suggested Reading
423 – 431	Section 10.1, 10.2 (for means), 10.4 (for proportions)

Pages	Problems
441 – 451	(***Section 10.9***) 1 – 10 (also identify the two that are kinda ludicrous, for reasons we discussed in class. Hint: they're the ones that fit answer choice "A"), 11, 13, 15, 16, 18, 29, 39, 40, 46, 48

I'm going to include this exercise here, since a) it fits this section nicely, and b) based on what I've been seeing around COCC/Oregon research, it seems very pertinent.

Sometimes, data is collected in Likert scale format. You most likely have seen them, somewhere. Here is an example Likert scale question, taken from COCC's student evaluations, as of 2011:

I would recommend this course:

Strongly Agree	…..					Strongly Disagree
7	**6**	**5**	**4**	**3**	**2**	**1**

If you strongly agreed with the belief that you'd recommend the course, you'd most likely select 7. If you hated it, and wouldn't wish it on your worst enemy, you'd probably select 1. But when would you select, say, 5? 3? 2? What do these numbers even ***mean***?

Likert data is neat, and simple…but its use is also amazingly non-intuitive and easily misused. You see, it's called **ordinal** data (you might remember this from your first project) in that the distance between those values is, for all intents and purposes, not well – defined. I mean, what's the difference (numerical) between a score of 5 and a score of 3 on this scale? Two…***what's***?

Now, the part that frosts my doughnut is this: often times, results of these Likert data items are simply averaged, and that average is used to make decisions. The reason said frosting occurs is that ***the average is meaningless*** in the case of these data! Consider these results, gotten by yours truly a few years back on two different student evals (same question, two MTH 111 courses two years apart):

I would recommend this course.

		Winter 2006	Spring 2008
Strongly Agree	7	13	17
.	6	1	3
.	5	3	1
.	4	0	0
.	3	1	1
.	2	0	0
Strongly Disagree	1	2	0

So, make sure you can understand what these numbers mean…for example, in the Winter 2006 column, I got thirteen 7's ("strongly agree"), one 6 ("a little less than strongly agree"), three 5's ("even a little less, but I'm not sure how much because no one taught me what the intervals are"), and so on.

What COCC does is average these data. For example, $\mu_{\text{Winter 2006}} = 5.85$, and $\mu_{\text{Spring 2008}} = 6.59$.

E1. Make sure you can see how they got those averages.

E2. Explain why those averages are meaningless (a similar problem occurred in # 21 of your last problem set).

What troubles me is this: those averages can (and often, are) used to make large decisions (promotion, tenure, etc.). For example, since my average in 2008 is "one point higher" (whatever ***that*** means) than the one in 2006, I might be viewed as having "improved".

That's silly.

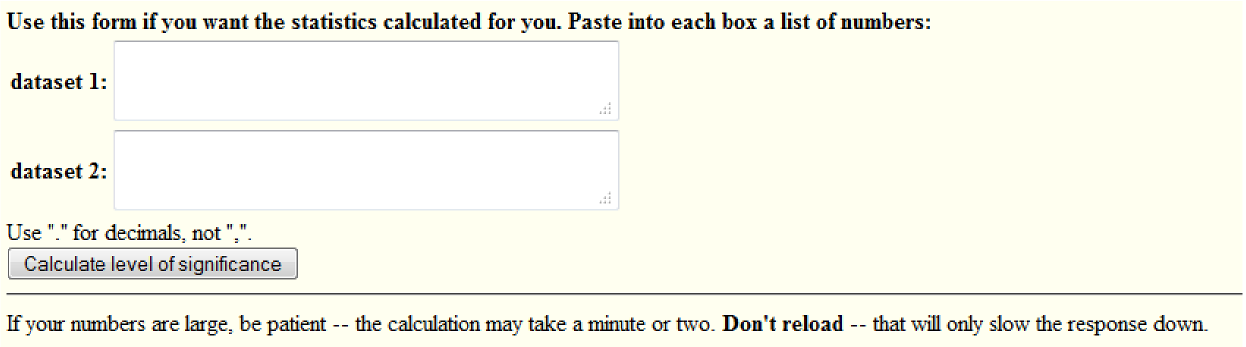
What we need is another approach to Likert data, a way to a) treat larger numbers as, indeed, higher scores (for example, "6" is higher than "5"), but also b) remove the "interval" between the scores. And we have just that, in what's called a Wilcoxon Rank Sum Test (AKA the Mann Whitney U test). The WRST/MWUT is a **non – parametric** test (i.e., the data don't have to be interval, nor normally distributed…unlike a two sample T – test, which requires that the data be continuously normally distributed, or CLT – applied). In fact, all we have to have is at least 10 data points in each of two independent samples. Voila...we're there. The test will see which of the following are more believable:

H_0: The two independent samples have equal medians.
H_1: The two independent samples have non – equal medians.

E3. Why would comparing medians be better than comparing means?

Our conclusion will follow, as usual, from the **P** – value, which can be gotten through a somewhat tedious– process that uses a new distribution ("U") and require a lot of crunching of data[1]. Here's a much easier way: use the applet in this link (http://elegans.som.vcu.edu/~leon/stats/utest.html) to get the **P** – value (the only downside is that you'll have to enter the data in raw form; it doesn't accept distributions)

Use the part of the page marked "**Use this form if you want the statistics calculated for you. Paste into each box a list of numbers":**



Use this form if you want the statistics calculated for you. Paste into each box a list of numbers:

dataset 1:

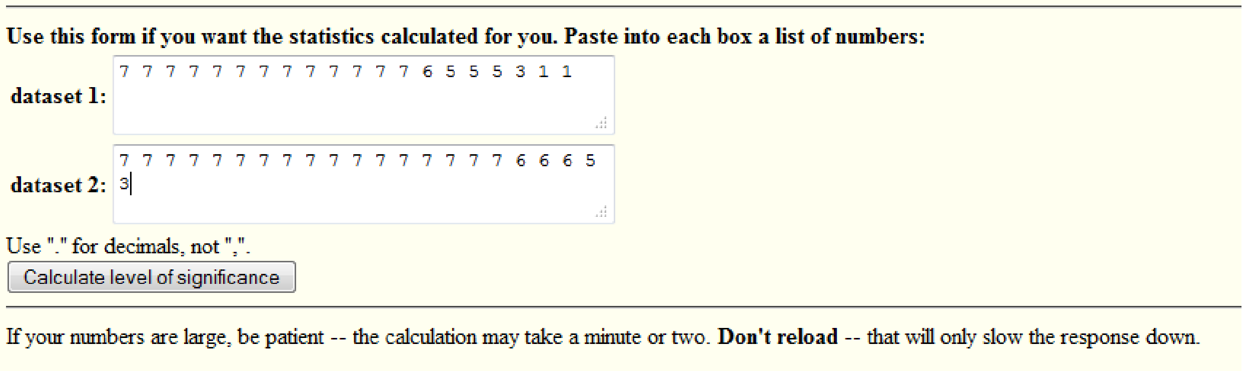
dataset 2:

Use "." for decimals, not ",".

If your numbers are large, be patient -- the calculation may take a minute or two. **Don't reload** -- that will only slow the response down.

Here's what it looked like when I entered the data:

[1] The phrase "rank sum" gives an idea to what's going on with the data. If you're interested, give 'er a Google.



Use this form if you want the statistics calculated for you. Paste into each box a list of numbers:

dataset 1: 7 7 7 7 7 7 7 7 7 7 7 7 7 6 5 5 5 3 1 1

dataset 2: 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 6 6 6 5
3

Use "." for decimals, not ",".

If your numbers are large, be patient -- the calculation may take a minute or two. **Don't reload** -- that will only slow the response down.

E4. Run the test, and tell me…did my scores significantly improve from 2006 to 2008?

E5. So now…say you're on a promotions committee, and need to make a decision about whether or not I should be promoted based on all of these data (the original scores, the "averages", the U – test results). What would you advise your colleagues?

E6. Certain statisticians will only use two – tailed tests; they make the (correct) claim that the difference in parameter, if it exists, will be more significant with a two – tailed test than a one – tailed one. Let's see if that makes sense.

Make up data (x_1, n_1, x_2, and n_2) so that:

a) When you run a 2 – Proportion Z interval ($\boxed{\text{STAT}}$ $\boxed{\blacktriangleleft}$ choice "B" on mine) on your data, there is no significant difference between p_1 and p_2.
b) A 2 – Proportion Z Test ($\boxed{\text{STAT}}$ $\boxed{\blacktriangleleft}$ $\boxed{6}$) testing for a difference ("≠") between p_1 and p_2 in agrees with part a (i.e., you get a $P > 5\%$).
c) However, a 2 – Proportion Z Test testing for one of p_1 and p_2 being greater than the other finds significance ($P < 5\%$)
d) OK…which of the following is always true?

   1) If you get statistical significance with a one – tailed test, you will also get statistical significance with a two – tailed test.
   2) If you get statistical significance with a two – tailed test, you will also get statistical significance with a one – tailed test.

e) Why, then, do many statisticians only run two – tailed tests?

E7. In class, for the assumptions for this kind of test, I told you that you should have

"Independent SRS, with $np \geq 5$ and $nq \geq 5$ in each sample...use a worst – case scenario, if necessary."

Now what could I have possibly meant by a "worst case scenario"?

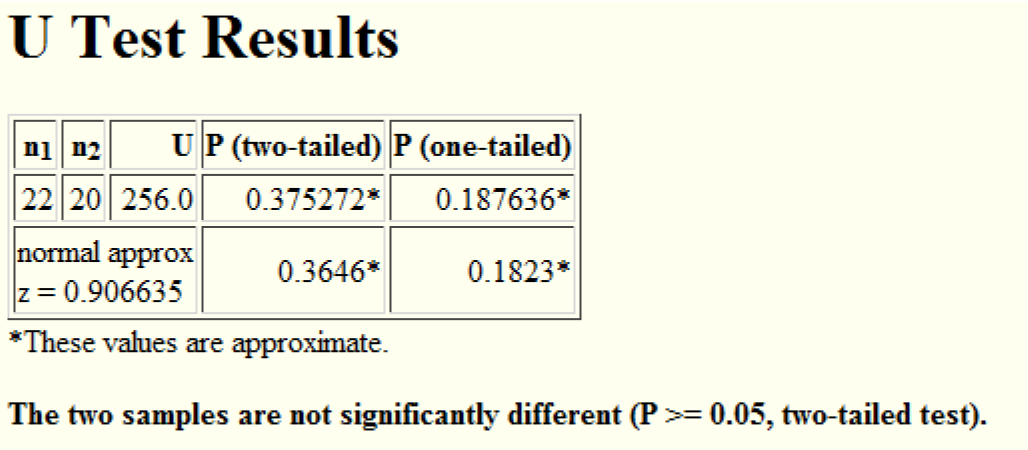
<u>Answers.</u>

<u>E1</u>. They’re weighted means, like a GPA. If you’re using your TI, place the scores 7, 6, 5, …1 into L1, and the frequencies into L2.

<u>E2</u>. They’re meaningless because their computation involves adding numbers together. Ask yourself: what does it mean to add? For example, what is 2 + 1? It’s 3. Why? Try to answers that question; it’s harder than it seems. Then, ask yourself – what do those numbers actually stand for in the Likert scale survey?

<u>E3</u>. Look at the original data again. You might graph them, too.

<u>E4</u>. Nope.



U Test Results

n_1	n_2	U	P (two-tailed)	P (one-tailed)
22	20	256.0	0.375272*	0.187636*
normal approx z = 0.906635			0.3646*	0.1823*

*These values are approximate.

The two samples are not significantly different (P >= 0.05, two-tailed test).

So, even though the 2008 “average” is “higher” than the 2006 “average”, it’s a meaningless measure, meaninglessly different.

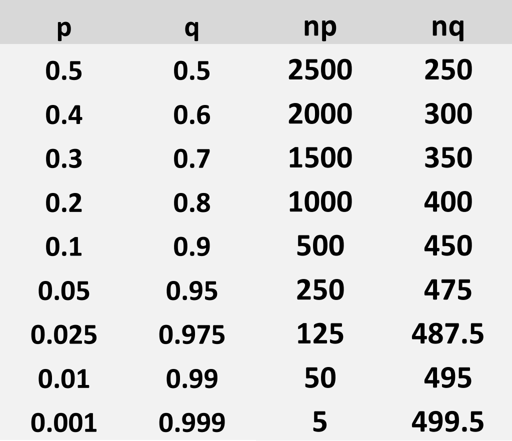
<u>E5</u>. I can’t ***wait*** to hear what you say!

<u>E6</u>. Hint: Make sure your data create a P – value that’s between 5% and 10%, non – inclusive.

<u>E7</u>. Since you don’t know either **p** nor **q**, pick one to be very, very small. Here’s why (assume ***n*** = 1000, and realize that **p** and **q** are interchangeable):

p	q	np	nq
0.5	0.5	50	50
0.4	0.6	40	60
0.3	0.7	30	70
0.2	0.8	20	80
0.1	0.9	10	90
0.05	0.95	5	95
0.025	0.975	2.5	97.5
0.01	0.99	1	99
0.001	0.999	0.1	99.9

As your **p** (or **q**) gets further and further away from 0.5, you need a larger and larger sample size to offset that disparity (as you might remember from our talks way back about single – proportion CIs). See how, once **p** falls below 5%, we don't satisfy the **np** ≥ 5 anymore? Well, here are the same proportions, with **n** = 5000:



p	q	np	nq
0.5	0.5	2500	250
0.4	0.6	2000	300
0.3	0.7	1500	350
0.2	0.8	1000	400
0.1	0.9	500	450
0.05	0.95	250	475
0.025	0.975	125	487.5
0.01	0.99	50	495
0.001	0.999	5	499.5

¡Bueno! We now satisfy the requirements for the test (assuming, of course, that our data are well – collected, random, and unbiased). So, by "worst case scenario", I mean "since you don't know what **p** and **q** are, assume one will be tiny and sample enough to satisfy **np** ≥ 5 or **nq** ≥ 5".

By having a larger sample size, a few other wondrous things happen, as well. Can you think of one?

Two Parameter HT Quizzes

Quiz 1.

The Princeton public school system did a statistical research project to see if the incidence of hungry children was the same across two schools in a low – income area of the town. A random sample (n=80) of elementary school students the first school revealed that 20% did not have breakfast before coming to school. A second school's random sample (n=180) showed that 12% did not have the corresponding breakfast. Now, we all know that 20% is different than 12%...however, is there a ***statistically significant*** difference between the hunger rates of hunger in students at these two low – income schools? Define "hunger rates", for the purposes of this problem, as the percentage of those who came to school without breakfast.

Quiz 2.

You have obtained the number of years of education from one random sample of 38 police officers from City A and the number of years of education from a second random sample of 30 police officers from City B. The average years of education for the sample from City A is 15 years with a standard deviation of 2 years. The average years of education for the sample from City B is 14 years with a standard deviation of 1.5 years. Is there a statistically significant difference (at the 5% level) between the education levels of police officers in City A and City B?

Quiz 3.

An astute MTH 244 student told me that there seemed to be more females than males in my classes, so I decided to collect some data (from my own historical records and COCC fact books) to see if there are proportionally more females in my classes than at COCC in general. What I found was this: in a random sample of 50 COCC students[2], there were 30 females. In a random sample of 50 of my students, there were 36 females. Does this data support the claim that there is a higher percentage of females in my classes (than at COCC in general)?

[2] You might not think 50's enough. It is...remember, I'm not trying to estimate the percentage of females, because COCC already knows that parameter. I'm trying to see if the percentage of females in my class is different. To achieve a large enough sample size to do that, I just need to ensure than ***np*** and ***nq*** are both at least 5 (or 10). A sample size of 50 ensures this.

Quiz 4.

(*no template for this one*)

Fully and correctly answer Question **E6** above (**2 points** for each of the sections **a** through **e**).