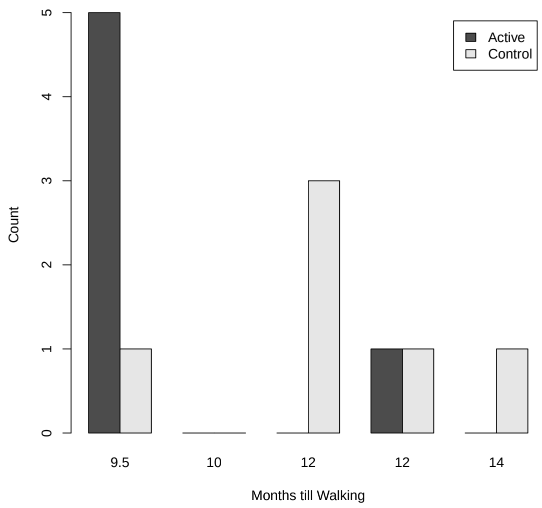
1/21/2010 36-402/608 ADA-II H. Seltman
Breakout #4 Comments

Question 1: A manual rank sum calculation requires finding the sum of all of the ranks. Without using a calculator or computer can you find the sum of the integers from 1 to 1000?

This relates to a story told about Carl Friedrich Gauss. In elementary school the students were asked to add all of the integers from 1 to 100. He did it very quickly by realizing that you can write the sum as (1+100) + (2+99) + (3+98) + $\cdots$ + (50+51) = 50*101 = 5050. In general the sum of n ranks is $(n + 1)n/2$.

Question 2:

Sleuth data problem 29 (Exercise and Walking Times) presents an experiment reported in the journal *Science*. The subjects are 12 one-week-old infants recruited as a "convenience sample" from white, middle class families. The infants were randomly assigned the "active" group which had stimulation of the walking reflex via four 3-minute long exercise sessions daily from week 2 to week 8. The control group received no stimulation. The time of first walking was recorded in months.



```
> with(d25, t.test(months~group))
        Welch Two Sample t-test
t = -1.8481, df = 9.976, p-value = 0.09442
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.4929271  0.3262604
sample estimates:
 mean in group Active mean in group Control
```

```
            10.12500             11.70833

> with(d25, wilcox.test(months~group, conf.int=TRUE))
        Wilcoxon rank sum test with continuity correction
W = 9, p-value = 0.1705
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -3.500013  0.750049
sample estimates:
difference in location
            -2.000005

Warning messages:
1: In wilcox.test.default(x = c(9.75, 9.5, 9.5, 9, 10, 13), y = c(9,  :
  cannot compute exact p-value with ties
```

Question 2: Which two tests were done? How do you know that they are both potentially appropriate? Which test is likely to be more reliable in the problem and why? Comment on internal and external validity if you know what they are.

We see an independent samples (unpaired, two-sample) t-test and a Wilcoxon rank-sum test (Mann Whitney U test). Either can be used for two indpendent samples (with uncorrelated errors). The t-test assumes Normality, but is quite robust to moderate degrees of non-Normality. Both tests assume equal variance. The t-test requires a quantitative outcome, while the rank-sum test also works for ordinal data. Although it's hard to be sure with such a small amount of data, there may be a high degree of non-Normality, so the rank sum test may be preferred.

Internal validity is assured by random assignment of treatment, as in this problem. This results in the only difference (on average) between the two groups being due to treatment, as opposed to additional differences due to confounding variables, as in an observational study. Good internal validity allows one to make causal conclusions.

External validity refers to how well the tested sample represents the population of interest. A convenience sample my have poor external validity unless the effect we are studying does not vary within the population. Poor external validity means that we cannot generalize our results to a population of interest.

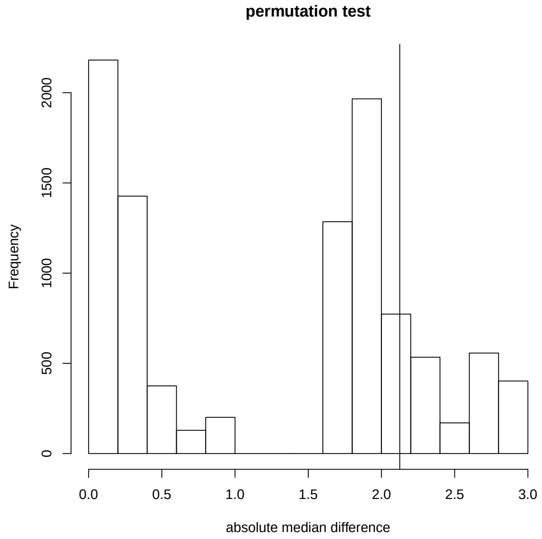
Question 3: Permutation test for Walking data

```
### Permuatation test of median difference
# For fun, find total number of possible randomizations of 12 items into 2 groups
choose(12,2) # [1] 66
```

```
# Function to test median differences by permutation
fm = function(y, x, nsim=1000, plotit=FALSE) {
     if (!is.numeric(x)) stop("x must be numeric")
     if (length(table(x))!=2) stop("x doesn't have 2 values")
     vals = as.numeric(names(table(x)))
     xsim = rep(1:2, table(x))
     fsim = function(dummy) {
               xperm=sample(xsim)
               stat=abs(median(y[xperm==1])-median(y[xperm==2]))
             }
     rslt = sapply(rep(1,nsim), fsim)
     stat = abs(median(y[x==vals[1]])-median(y[x==vals[2]]))
     p.value=mean(rslt>=stat)
     if (plotit) {
       hist(rslt, main="permutation test", xlab="absolute median difference")
       abline(v=stat)
     }
     return(p.value)
}

fm(d25$months,as.numeric(d25$group), 10000, plotit=TRUE) # 0.2479 (takes 12 seconds)
fm(d25$months,as.numeric(d25$group), 100000) # 0.24366
fm(d25$months,as.numeric(d25$group), 1000000) # 0.242174 (takes tens of minutes)
```

The top of the `fm()` function just allows the "x" input, which defines which group "y" belonged to, to be any two numbers. Variable "vals" will be the two possible "x" values. Variable "xsim" is just the numbers 1 and 2 (representing group assignment) replicated the right number of times to match the data but in no meaningful order, e.g., 1,1,1,1,1,1,2,2,2,2,2,2 for our data. Function `fsim()` computes the absolute value of the difference of medians for the two groups for one particular random group assignment. The `sapply()` computes a large number (nsim) of these differences. Variable "stat" is the absolute value of the difference of medians for the two groups for the real data.



Question 3: How is the permutation p-value calculated? How do you interpret the plot? What does the p-value mean? Without computational details, how is the exact permutation p-value different from the approximate one calculated here? Do you think this test has more or less power than the Wilcoxon test?

The perputation p-value is calculated by computing the distribution of the statistic of interest (here median difference across treatment groups) over many or all possible randomizations of treatment application. The p.value is the probability of seeing the observed statistic or one more extreme (less like what is expected under H_0) based on the computed (empiric) randomization distribution.

The plot shows the randomization distribution and observed statistic. Due to the use of absolute value, more extreme corresponds to higher values, so the p-value is the area under the curve above the observed value. A small p-value means that the observed statistic is unlikely under random treatment assignment with an ineffectual treatement, so the treatment probably affects the outcome.

An exact p-value would come from computing the distribution of the desired statistic once for each possible randomization.

The test probably has lower power than the Wilcoxon test, because it only looks at the medians, while the Wilcoxon test looks at all of the ranks.

Question 4: Tail feather experiment

Wiebe and Bortolotti (2002) examined color in the tail feathers of northern flickers. Some of the birds had one ”odd” feather that was different in color or length from the rest of the tail feathers, presumably because it was regrown after being lost. They measured the yellowness of one odd feather on each of 16 birds and compared it with the yellowness of one typical feather from the same bird.

The question of interest is whether the odd feather is more or less yellow than the typical feathers.

Here are the data:

```
   bird typical    odd
1     A  -0.255 -0.324
2     B  -0.213 -0.185
3     C  -0.190 -0.299
4     D  -0.185 -0.144
5     E  -0.045 -0.027
6     F  -0.025 -0.039
7     G  -0.015 -0.264
8     H   0.003 -0.077
9     I   0.015 -0.017
10    J   0.020 -0.169
11    K   0.023 -0.096
12    L   0.040 -0.330
13    M   0.040 -0.346
14    N   0.050 -0.191
15    O   0.055 -0.128
16    P   0.058 -0.182
```

Question 4: Show the calculations needed for the sign test. If you have a calculator handy, compute the z-score, and comment on your conclusion (without using a Normal probability table).

There are no ties. The odd feather is more yellow for only 3 of 16 pairs.

K=3, n=16, E(K)=n/2=8, SE(K)=$\sqrt{n/4} = 2$. Z=(3-8)/2=-2.5 With Z¡(-1.96), we would reject H_0, that odd feathers are just as likely to be more yellow than the typical feather as they are to be less yellow.

Question 5: More on the birds

```
> with(tail, t.test(typical, odd, paired=TRUE))
        Paired t-test
t = 4.0647, df = 15, p-value = 0.001017
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.06521848 0.20903152
sample estimates:
mean of the differences
               0.137125

> with(tail, wilcox.test(typical, odd, paired=TRUE, conf.int=TRUE))
        Wilcoxon signed rank test
V = 126, p-value = 0.001312
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 0.0505 0.2150
sample estimates:
(pseudo)median
       0.13025
```

Question 5: What features of the data might suggest use of the signed-rank test over the paired t-test? Under what circumstances do you expect such similar results for the parametric (model based) and non-parametric tests? What do you conclude about the odd feathers? It is not OK to do several tests and choose the one you like best. Explain why not.

If we see gross non-Normality, the signed-rank test is preferred. If the two groups both have normal outcomes, the test results will be similar. Odd feathers do appear to be less yellow (p=0.001), 95% CI for the typical minus odd yellowness = [0.065, 0.209].

When you do several tests, each one has an additional chance of making a type-1 error, so your overall chance of a type-1 error is increased over the nominal 5% that we try to achieve.