BMEG 802 – Advanced Biomedical Experimental Design and Analysis

Hypothesis Testing

Joshua G. A. Cashaback, PhD

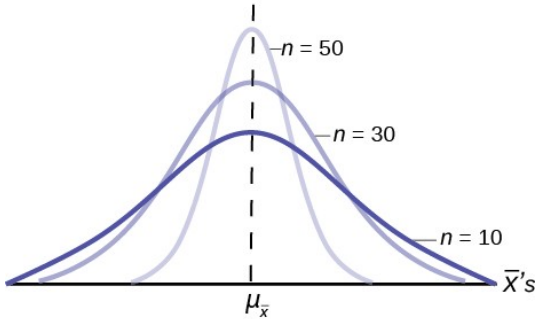
Recap

- Sampling Distribution
- T-distribution
- Confidence Interval

Recap - Sampling Distribution

- Sampling Distribution: The probability distribution of a given statistic (e.g., mean) taken from a random sample
- Constructing a Sampling Distribution
  - Randomly draw n sample points from a finite population with size N
  - Compute statistic of interest
  - List different observed values of the statistic with their corresponding frequencies

Recap - Distribution of the Sample Mean

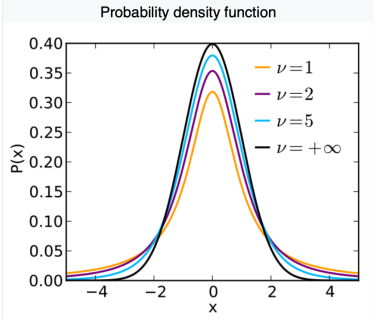


- as n increases we become more confident (less spread) of our estimate of the population mean
- for example, as we sample the heights of people in a country we get a better estimate of that nation's average height.

Recap - Distribution of the Sample Mean

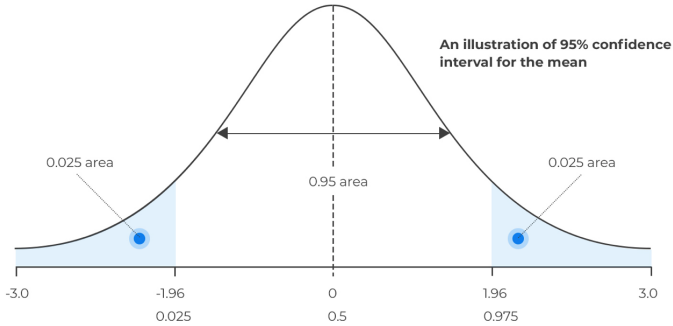
- Rarely know the true population mean or population variance (see primer)
  - can use z-tests / z-tables in this case (I've never used this once...)
- Sample Mean (unknown variance)
  - *Mean*: $\mu_{\bar{X}} = \mu$
  - *Standard Error of the Mean*: $s_{\bar{x}} = \frac{s}{\sqrt{n}}$
- Difference Between Two Means (unknown and equal variance)
  - *Mean Difference*: $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$
  - *Standard Error of the Mean Difference*: $s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Recap - Student's T-Distribution



- ν influence spread, the degrees of freedom (ν = n - 1)
- *t-score of a sample mean*: $t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$
- *t-score of the difference between two sample means*: $t = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
- Normal? -> tend to underestimate variance with n $<$ 30

Confidence Intervals



- Confidence intervals (see Primer)
  - Normal Distribution
  - t-distribution
  - Welsh's
- All rely on calculating a critical value

Learning Objectives

- Null hypothesis testing
- Defining an α
- 1 vs 2 tailed tests
- Parametric
  - 1 sample, 2 sample (Welch's), & paired t-tests
- Non-parametric
  - Mann-Witney U test (between), Wilcoxon signed-rank (paired)
- Correct for Multiple Comparisons

Null Hypothesis Testing

- Null Hypothesis (H_0): there is no significant difference between specified populations, any observed difference being due to sampling or experimental error.
- Alternative Hypothesis (H_A): a position that states something is happening, a new theory is preferred instead of an old one (null hypothesis)
- THE critical ingredient in an inferential statistical test (frequentist approach):
  - determining the probability of obtaining the observed data, assuming the null hypothesis is true.
- Remember, we cannot prove a theory to be true, we can only prove one to be false!

Null and Alternative Example

Research Hypothesis: I hypothesize that the drug will change survival time compared to no treatment

- Null Hypothesis: $H_0 : \mu_{drug} = \mu_{control}$
- Alternative Hypothesis: $H_A = \mu_{drug} \neq \mu_{control}$

Two-sided and One-sided Alternative Hypotheses

- Two-tailed: $H_0 : \mu_{drug} = \mu_{control}; H_A = \mu_{drug} \neq \mu_{control}$
- One-tailed: $H_0 = \mu_{drug} \leq \mu_{control}; H_A = \mu_{drug} > \mu_{control}$
- One-tailed: $H_0 = \mu_{drug} \geq \mu_{control}; H_A = \mu_{drug} < \mu_{control}$
- Only use a one-tailed / directional hypothesis if you have a strong theoretical prediction (for example, from a model).
  a. can gain statistical power
  b. but sometimes findings are meaningful and interesting if they go in an unexpected way...
- We can accommodate both two-tailed and one-tailed tests statistically

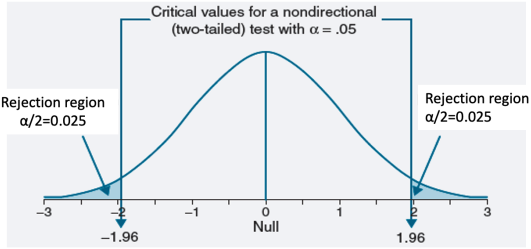
Define your level of significance (α)

Level of significance (α): Probability of rejecting the NULL hypothesis due simply to chance.

- $\alpha = 0.05$
- some use others (e.g. 0.01, 0.0000003 particle discovery in high energy physics)
  - somewhat arbitrary lines in the sand. Sort out what your research area typically uses.
- How do we determine what the probability of rejecting the Null hypothesis given our data?
  - i.e., what is our p-value and is higher or lower than our level of significance?

2-Tailed Test

e.g., $H_A = \mu_{drug} \neq \mu_{control}$

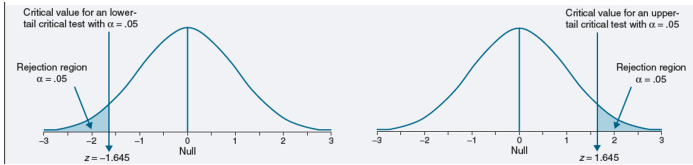


2-tails: We are interested in any significant deviations from H_0

- the sum of the tails sums to α (0.025 in each tail for a two-tailed when $\alpha = 0.05$)
- Calculate t-statistic (as defined above)
- see where the t-statistic lies relative to 'critical score' that depends on defined alpha (same procedure used to calculate confidence intervals)
- for the right tail, if the t-statistics $>$ critical value $=$ reject null & accept alternative
- for the left tail, if the t-statistics $<$ critical value $=$ reject null & accept alternative
- if the t-test is between the rejection region $=$ fail to reject the null

1-Tailed Test

e.g., $H_A = \mu_{drug} < \mu_{control}$ (left plot) or $H_A = \mu_{drug} > \mu_{control}$ (right plot)



- 0.05 in each tail
- if t-statistic within rejection region = reject null & accept alternative
- do you see why you get power with a one-sided / directional hypothesis?

Steps for Hypothesis Testing

1. Define your hypotheses and choose your level of significance
2. t-tests -> calculate the corresponding "test statistic" and compare the result against the "critical value"
   - is the t-score $>$ or $<$ than critical value?
     - reject or fail to reject the null.
   - remember direction for 1 tailed tests, and α for 1 (e.g., t_{α}) vs. 2 tailed tests (e.g., $t_{\alpha/2}$)
3. Determine the p-value for your test
   - area under the curve – remember direction, and whether 1 vs. 2 tailed
   - reject the null $p < \alpha$ or fail to reject null if $p >$ *alpha* (should correspond with critical value assessment)
4. State your conclusion

Parametric Tests

- 1 sample t-test
  - Example 1: 1-tailed (left tail)
  - Example 2: 1-tailed (right tail)
  - Example 3: 2-tailed
- 2 sample Welch's t-test
  - Example 4: 2-tailed (can also do 1-tailed)
- Paired t-test
  - Example 5: 2-tailed (can also do 1-tailed)

Given Sample Mean & SD, OR simulate by drawing from a prob dist, OR given an array of data.

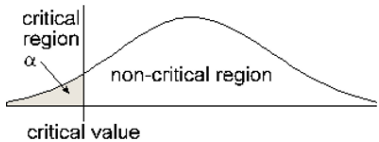
1 Sample t-test (1 tailed)

- 1 sample t-test (1 tailed)
- $t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$, where μ is some value we are comparing our sample to.
- find t_{crit}
- find p-value

1 Sample t-test (Example 1)

Sample of 12 people lost 0.61 kg with standard deviation of s= 1.62 kg. Can we conclude their weight is **less than** than their original weight?

- Known: $\bar{x} = -0.61, s = 1.62, n = 12$
- Assumptions: Population is normally distributed
- Hypotheses: $H_0 : \mu \leq 140$ and $H_A : \mu > 140$
- $t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$
- $-1.3 = \frac{-0.61 - 0}{\frac{1.62}{\sqrt{12}}}$



1 Sample t-test (Example 1 cont'd)

```
t0 = (-0.61 - 0)/(1.62/(12)^(1/2))
alpha = 0.05
tcrit0 = qt(alpha, 12 - 1) # (alpha, df) - NOTICE alpha!
pval0 = pt(t0,12 - 1) # looking at LEFT tail
t0
```

```
## [1] -1.304384
```

```
tcrit0
```

```
## [1] -1.795885
```

```
pval0
```

```
## [1] 0.1093673
```

1 Sample t-test (Example 1 cont'd)

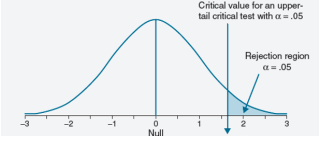
Decision, Conclusion, and p value

- -1.3 NOT less than -1.796, so fail to reject H_0
  - More generally, p-value (0.109) $>$ 0.05.
- There was no significant weight loss with the diet

1 Sample t-test (Example 2)

Among 157 African-American men seen in the emergency department at the hospital, the mean systolic blood pressure was 146 mm Hg with a standard deviation of 27. Can we conclude based on this data that the mean systolic blood pressure for a population of African-American men is **greater than** 140 (i.e., μ) at the 95% confidence level?

- Known: $\bar{x} = 146, s = 27, n = 157$
- Assumptions: Population is normally distributed
- Hypotheses: $H_0 : \mu \leq 140$ and $H_A : \mu > 140$
- $t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$
- $2.78 = \frac{146 - 140}{\frac{27}{\sqrt{157}}}$



1 Sample t-test (Example 2 cont'd)

```
t0 = (146-140)/(27/(157)^(1/2))
alpha = 0.05
tcrit0 = qt(1 - alpha, 157 - 1) # (1-alpha, n - 1) - NOTICE 1-alpha!
pval0 = 1 - pt(t0,157 - 1) # looking at RIGHT tail (1 - p)!
t0
```

```
## [1] 2.784436
```

```
tcrit0
```

```
## [1] 1.65468
```

```
pval0
```

```
## [1] 0.003013078
```

1 Sample t-test (Example 2 cont'd)

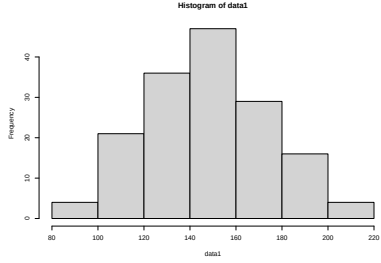
Decision, Conclusion, and p value

- $2.78 > 1.65$ so reject H_0
  - More generally, p-value $(0.003) < 0.05$.
- The mean systolic blood pressure for the sampled population is greater than 140

1 Sample t-test (Example 2 cont'd)

Same Example, but lets sample from a Normal Distribution to get the data. Here we are pretending to simulate the experiment.

```
n1 = 157
alpha1 = 0.05
data1 <- rnorm(n1, mean = 146, sd=27)
hist(data1)
```



1 Sample t-test (Example 2 cont'd)

Same Example, but lets sample from a Normal Distribution to get the data. Here we are pretending to simulate the experiment.

```
t1 = (mean(data1) - 140) / (sd(data1) / (157)^(1/2.))
tcrit1 = qt(1 - alpha1, n1 - 1)
pval1 = 1 - pt(t1,n1 - 1)
t1
```

```
## [1] 3.778607
```

```
tcrit1
```

```
## [1] 1.65468
```

```
pval1
```

```
## [1] 0.0001119896
```

1 Sample t-test (Example 2 cont'd)

Alternatively, use the built in t.test function

```
res1 <- t.test(data1, mu = 140, alternative = "greater")
#alternative options: "less", "greater", "two.sided"
```

Pay attention when you simulate different experiments and how the p-value changes!

1 Sample t-test (Example 2 cont'd)

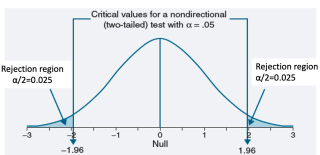
Alternatively, use the built in t.test function

```
res1
```

```
##
##  One Sample t-test
##
## data:  data1
## t = 3.7786, df = 156, p-value = 0.000112
## alternative hypothesis: true mean is greater than 140
## 95 percent confidence interval:
##  144.5096      Inf
## sample estimates:
## mean of x
##  148.0228
```

1 Sample t-test (Example 3)

A cookie company claims that there are 15 chocolate chips per cookie, but you aren't convinced. You take 10 cookies and count the number of chocolate chips in each cookie. Here is what the data looks like: [13,14,15,17,18,19,21,20,19,20]. Are the number of chocolate chips significantly different from 15?



1 Sample t-test (Example 3 cont'd)

```
data2 = c(13,14,15,17,18,19,21,20,19,20)
alpha = 0.05
t2 = (mean(data2) - 15) / (sd(data2) / (length(data2))^(1/2.))
tcrit2 = qt(alpha/2, length(data2)-1) # Left tail (alpha/2 = 0.025)
pval2 = 2*pt(-abs(t2),length(data2) - 1) # -abs() do calc on left-tail
# multiply by 2 (accounting for both sided)
abs(t2)
```

```
## [1] 2.982405
```

```
abs(tcrit2)
```

```
## [1] 2.262157
```

```
pval2
```

```
## [1] 0.01538941
```

1 Sample t-test (Example 3 cont'd)

```
res2 <- t.test(data2, mu = 15, alternative = "two.sided")
res2
```

```
##
##  One Sample t-test
##
## data:  data2
## t = 2.9824, df = 9, p-value = 0.01539
## alternative hypothesis: true mean is not equal to 15
## 95 percent confidence interval:
##  15.6279 19.5721
## sample estimates:
## mean of x
##      17.6
```

1 Sample t-test (Example 2 cont'd)

Decision, Conclusion, and p value

- $2.98 > 2.23$ so reject H_0
  - More generally, p-value $(0.015) < 0.05$.
- There are significantly more chocolate chips in the cookies than 15. Yah!

2 Sample Welch's t-test

- 2 sample tests are interested in whether there are differences between 2 groups
- There are several traditional 2 sample t-tests, where the appropriate version depending on equal or unequal sample size or variance (giant flow charts!)
- Welch's t-test gives gives equivalent answer to traditional t-test when there is an equal sample size or variances, BUT can also handle unequal sample size and variance.
  - http://daniellakens.blogspot.com/2015/01/always-use-welchs-t-test-instead-of.html
- Same t-statistic calculation and t-distribution, just have to apply correction for degrees of freedom (df)

2 Sample Welch's t-test

- 2 sample t-test (1 or 2 tailed)
- $t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
- $\nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}}$
- find t_{crit}
- find p-value
- t.test() defaults to Welch's t-test

2 Sample Welch's t-test (Example 4)

A math test was given to 300 17 year old students in 1978 and again to another 350 17 year old students in 1992.

- Group 1: $X_1 = 300.4, S_1 = 34.9, n = 300$
- Group 2: $X_2 = 306.7, S_2 = 30.1, n = 350$

Is there a significant difference between math scores? Use $\alpha = 0.01$. Solve this by calculating the t-score and critical value, as well as solving the p-value. Then, simulate this experiment by drawing the values from a normal distribution and use a built in Welch's t-test function (e.g., t.test() in R).

2 Sample Welch's t-test (Example 4 cont'd)

```
t4 = ((306.7 - 300.4)-(0-0))/(34.9^2 / 300 + 30.1^2 / 350)^(1/2)
v4 = (34.9^2 / 300 + 30.1^2 / 350)^2 / (34.9^4 / (300^2 * (300 - 1)) + 30.
alpha4 = 0.01
tcrit4 = qt(alpha4/2, v4)
pval4 = 2*pt(-abs(t4),v4)
abs(t4)
```

```
## [1] 2.443286
```

```
abs(tcrit4)
```

```
## [1] 2.584122
```

```
pval4
```

```
## [1] 0.01484393
```

2 sample Welch's t-test (Example 4 cont'd)

```r
group1 <- rnorm(300, mean = 300.4, sd=34.9)
group2 <- rnorm(350, mean = 306.7, sd=30.1)
t.test(group1, group2, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  group1 and group2
## t = -2.6843, df = 589.82, p-value = 0.007471
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.844209  -1.835492
## sample estimates:
## mean of x mean of y
```

Paired t-test (1 or 2 tailed)

- $t = \frac{\bar{X}_D - \mu_0}{s_D/\sqrt{n}}$, where $\bar{X}_D$ and s_D are the mean and standard deviation of the differences between paired samples.
  - e.g., collecting the same person before and after some intervention
- Sometimes, it is useful to treat populations of pairs as one population of differences (d's)
- Gain power using a 'within design', because you control for a lot of unwanted variability.
- Example A: Does Gas A give better mileage than Gas B?
- Example B: Does Sunscreen A work better than Sunscreen B?
  - Paired comparisons eliminate unwanted variability (car type, mileage, skin tone, time in sun...)

Paired t-test (Example 5)

- A manufacturer claims it has developed an additive that increases gas mileage. But you are not sure whether the additive will increase or decrease performance. They recruit 10 drivers. Each driver drives a car on a well-conditioned track. They record the gas mileage without any additive, then with additive.

Driver	with additive	w/out additive	difference
1	22	18	4
2	25	21	4
3	17	16	1
4	24	22	2
5	16	19	-3
6	29	24	5
7	20	17	3
8	23	21	2
9	19	23	-4
10	20	18	2

Paired t-test (Example 5 cont'd)

```
data5a = c(22,25,17,24,16,29,20,23,19,20)
data5b = c(18,21,16,22,19,24,17,21,23,18)
data5diff = data5a - data5b
t5 = (mean(data5diff) - 0) / (sd(data5diff) / (length(data5diff))^(1/2.))
tcrit5 = qt(0.05/2, length(data5diff)-1)
pval5 = 2*pt(-abs(t5),length(data5diff)-1)
abs(t5)
```

```
## [1] 1.714286
```

```
abs(tcrit5)
```

```
## [1] 2.262157
```

```
pval5
```

```
## [1] 0.1206207
```

Paired t-test (Example 5 cont'd)

```r
t.test(data5a, data5b, paired = TRUE, alternative = "two.sided")
```

```
##
##  Paired t-test
##
## data:  data5a and data5b
## t = 1.7143, df = 9, p-value = 0.1206
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5113467  3.7113467
## sample estimates:
## mean of the differences
##                     1.6
```

1 Sample t-test (Example 2 cont'd)

Decision, Conclusion, and p value

- Assumptions: Differences are normally distributed
- Hypotheses: $H_0 : \mu = 0$ and $H_A : \mu \neq 0$
- $1.71 < 1.65$ so we fail to reject H_0
  - More generally, 0.1206 (p-value) > 0.05 (α).
- Therefore, the additive did not lead to significantly different gas mileage.

Nonparametric Tests

- Sometimes data is non-normal (skewed, bimodal, etc.), or ordinal, so what do we do?
  - Use Shapiro-Wilk normality test
    - Can transform data (e.g., log, sqrt, etc.), but these also make assumptions
- Mann-Witney U test (2 sample) & Wilcoxon (paired)
  - rank method
  - based on calculating all possibilities to form distribution
- Bootstrapping (we'll get to this later)

2 Sample, Mann-Witney U test

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

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$U = min(U_1, U_2)$$

- If $U < U_{crit}$ we reject the null (opposite from t-test; always reject if $t > t_{crit}$)

  - A researcher designed an experiment to assess the effects of prolonged inhalation of cadmium oxide. 8 animals inhaled cadmium oxide and 10 animals were controls. The dependent variable was hemoglobin level. Does cadmium oxide influence hemoglobin levels?

Mann-Witney U Example

A	B
4	23
7	6
8	3
9	24
13	17
13	14
17	24
11	29
	13
	33

Simple Rank	A	Merge	B
1		3	1
2	2	4	
3		6	3
4	4	7	
5	5	8	
6	6	9	
7	7	11	
8	9	13	
9	9	13	
10		13	9
11		14	11
12	12.5	17	
13		17	12.5
14		23	14
15		24	15.5
16		24	15.5
17		29	17
18		33	18
total	54		116.5

Mann-Witney U Example

$$U_1 = 61.5 = 8 \cdot 10 + \frac{8(8+1)}{2} - 54.5$$

$$U_2 = 18.5 = 8 \cdot 10 + \frac{10(10+1)}{2} - 116.5$$

$$U = 18.5 = min(U_1, U_2)$$

```
alpha = 0.05
qwilcox(alpha/2,8,10) - 1 # two-tailed, critical value! (1 tailed = 0.05)
```

```
## [1] 17
```

fail to reject the null since $U(18.5) > U_{crit}$ (17).

Mann-Witney U Example

```r
alpha = 0.05
G1 = c(4,7,8,9,13,13,17,11)
G2 = c(23,6,3,24,17,14,24,29,13,33)
wilcox.test(G1, G2, alternative = "two.sided")
```

```
## Warning in wilcox.test.default(G1, G2, alternative = "two.sided"): canno
## compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  G1 and G2
## W = 18.5, p-value = 0.06125
## alternative hypothesis: true location shift is not equal to 0
```

Paired, Wilcoxon signed-rank

- nonparametric test for paired samples

$$W = \sum_{i=1}^{N_r} [sgn(x_{2,i} - x_{1,i}) \cdot R_i]$$

- sgn is +1 or -1
- exclude pairs where difference equals zero, N_r is the reduced sample size
- If $W < W_{crit}$ we reject the null (opposite from t-test; always reject if $t > t_{crit}$)

Wilcoxon signed-rank Example

i	$x_{2,i}$	$x_{1,i}$	$x_{2,i} - x_{1,i}$			
			sgn	abs	R_i	sgn $\cdot R_i$
5	140	140		0		
3	130	125	1	5	1.5	1.5
9	140	135	1	5	1.5	1.5
2	115	122	–1	7	3	–3
6	115	124	–1	9	4	–4
10	135	145	–1	10	5	–5
8	125	137	–1	12	6	–6
1	125	110	1	15	7	7
7	140	123	1	17	8	8
4	140	120	1	20	9	9

- ranked by absolute value
- when ties, you average (e.g., 1 and 2 ranks tied, so use average of 1.5 for each)

Wilcoxon signed-rank Example

```
alpha = 0.05
Nr = 9
W = abs(1.5 + 1.5 - 3 - 4 - 5 - 6 + 7 + 8 + 9)
# calculate Wcrit (instead of lookup table)
# 0.05/2 = 2-tailed; 0.05 = 1 tailed
Wcrit = -(qsignrank(alpha/2, Nr, lower.tail=FALSE)+1)+Nr*(Nr+1)/2
W
```

```
## [1] 9
```

```
Wcrit
```

```
## [1] 5
```

Since W $>$ Wcrit, we cannot reject the null hypothesis

Wilcoxon signed-rank Example

```
G1 = c(125,115,130,140,140,115,140,125,140,135)
G2 = c(110,122,125,120,140,124,123,137,135,145)
wilcox.test(G1, G2, paired = TRUE, alternative = "two.sided")
```

```
## Warning in wilcox.test.default(G1, G2, paired = TRUE, alternative =
## "two.sided"): cannot compute exact p-value with ties
```

```
## Warning in wilcox.test.default(G1, G2, paired = TRUE, alternative =
## "two.sided"): cannot compute exact p-value with zeroes
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  G1 and G2
## V = 27, p-value = 0.6353
## alternative hypothesis: true location shift is not equal to 0
```

Correcting for Multiple Comparisons

- We want our tests to find true positives and true negatives
- Multiple comparison
  - spurious findings due to chance
  - Type I error (false positive)
  - many different versions
- Perform omnibus test before multiple corrections (e.g, ANOVA)
  - Later in course

Bonferroni

- Simplest: $\alpha_{adjusted} = \frac{\alpha}{n}$
- e.g., $\alpha_{adjusted} = 0.0167 = \frac{0.05}{3}$
- Controls for false positives (Type I errors)
- Overly conservative
  - leads to false negatives (Type II errors)

```
pvals = c(0.01, 0.02, 0.04)
p.adjust(pvals, method = "bonferroni", n = length(pvals))
```

```
## [1] 0.03 0.06 0.12
```

Holm-Bonferroni

- strikes a balance between Type I and Type II errors
- step-wise correction:
  1. sort p-values from smallest to largest
  2. Test whether $P_k < \frac{\alpha}{m+1-k}$. If so, reject H_k and move to the next (k^{th}) p-value from the total number of comparisons (m), otherwise STOP (and remaining p-values are deemed not significant)
- e.g., with $m = 4$ the **adjusted alpha's** would be: $\frac{0.05}{4}, \frac{0.05}{3}, \frac{0.05}{2}, \frac{0.05}{1}$
- Typically you report the **adjusted p-value**. Just multiply your p-value by the adjusted alpha's denominator.

```
pvals = c(0.01, 0.02, 0.04)
p.adjust(pvals, method = "holm", n = length(pvals))
```

```
## [1] 0.03 0.04 0.04
```

Many Multiple Comparison Corrections

- Tukey - all possible comparisons: TukeyHSD()
- Scheffe
- Dunnett
- Fisher's LSD (least significant difference)
- Newman-Keuls
- Find what your field does and, more importantly, justify your decisions!

Interpretation of p-value

- if $p < \alpha$, then you can reject the null hypothesis.
- even if you reject the Null, this does not mean the alternative hypothesis is necessarily true (remember, you can only falsify a theory)
- can find statistically significance, but it is functionally irrelevant (e.g., 1 IQ point difference after sampling 1 million people)
- no degree of significance (e.g., higly significant). It is either significant or it isn't. Could say, that it is highly reliable.