Activity 22: Testing difference in means between two independent groups

Scenario: In 2010, the Centers for Disease Control and Prevention reported 35.7% of American adults are obese[1]. Some believe obese individuals face discrimination; being viewed as having physical, moral, and emotional impairments.

Physicians, who are trained to treat all patients warmly and who have access to research suggesting uncontrollable and hereditary aspects of obesity, may also believe obese individuals are undisciplined and suffer from self-control issues.

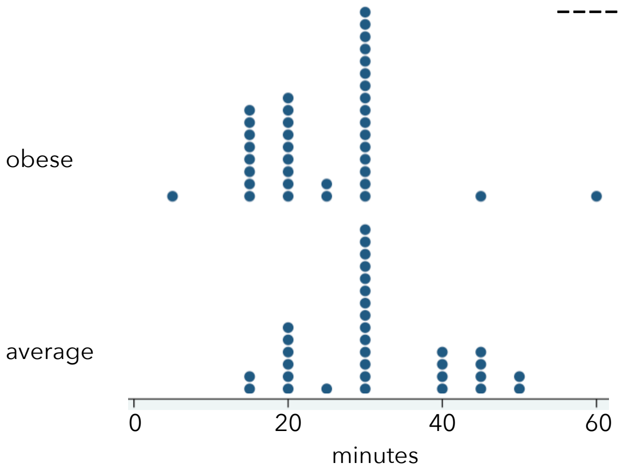
A 2001 study examined physicians' behavioral intentions as well as their expressed attitudes towards average-weight and obese patients. 71 primary care physicians in Houston participated in this study. The doctors were sent a packet containing a medical chart similar to the one they view upon seeing a patient. This chart portrayed a patient who was displaying symptoms of a migraine headache but was otherwise healthy. The weight of the patient was manipulated so that:

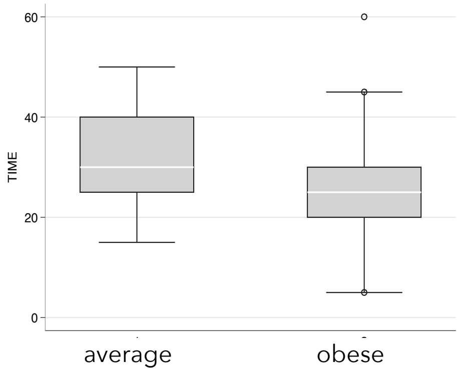
- 33 doctors received a chart from a patient of average weight (body mass index = 23)
- 38 doctors received a chart from an obese patient (body mass index = 36)

The doctors were instructed to examine the charts and then asked, among other questions, how much time they believed they would spend with the patient. Does the amount of time doctors spend with a patient depend on whether the patient is obese?

The following data were obtained:

```
   weight |         N       mean        p50         sd
----------+--------------------------------------------
  average |        33   31.36364         30   9.864134
    obese |        38   24.73684         25   9.652571
----------+--------------------------------------------
    Total |        71    27.8169         30   10.23762
-------------------------------------------------------
```





[1] *National Obesity Trends*, CDC NCHS, 2010, retrieved 2012-03-26: http://www.cdc.gov/obesity/data/adult.html

Hebl, M., & Xu, J. (2001). Weighing the care: Physicians' reactions to the size of a patient. International Journal of Obesity, 25, 1246-1252.

1) Identify the independent and dependent variables in this study.

Independent variable: ______________________________________

Dependent variable: ______________________________________

2) State the null and alternate hypotheses. Express the consequences of both Type I and Type II errors in this study.

Null hypothesis: ______________________________________

Alternate hypothesis: ______________________________________

Type I error consequence: ____

Type II error consequence: ____

3) Why can't we simply look at the sample averages and conclude that physicians report they would spend less time with obese patients?

4) Consider the assumptions necessary for us to conduct an independent samples t-test. Are these assumptions reasonable in this situation?

5) Let's assume your null hypothesis is true. Under this assumption, sketch the sampling distribution we would get if we repeatedly took samples of size 33 and 38 and calculated the difference between the means for each sample. Label the mean and standard error of this distribution. Identify the critical value and shade-in the rejection region.

6) The difference in means we actually observed in this study was 6.6268. Under your null hypothesis, how likely were we to obtain results at least as extreme as this? Calculate & interpret this p-value. What conclusions can you make?

7) Here's the output when I conducted an independent samples t-test in Stata. Does this match your calculations?

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 average |      33    31.36364    1.717125    9.864134    27.86597    34.86131
   obese |      38    24.73684    1.565854    9.652571    21.56412    27.90956
---------+--------------------------------------------------------------------
combined |      71     27.8169    1.214982    10.23762     25.3937    30.24011
---------+--------------------------------------------------------------------
    diff |            6.626794    2.320283                1.997955    11.25563
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =   2.8560
Ho: diff = 0                                     degrees of freedom =       69

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9972         Pr(|T| > |t|) = 0.0057          Pr(T > t) = 0.0028
```

8) Create a 90% confidence interval for the difference in time spent with average weight and obese patients. Before you calculate the interval, predict if it will contain zero. Why are you able to predict this?

9) Make sure you can conduct the t-test and calculate the confidence interval on your calculator or computer.

Scenario: Will a smiling person accused of a crime be treated more leniently than one who is not smiling? If so, does the type of smile make a difference?

A 1995 study asked 136 students to serve as members of a college disciplinary panel and judge a student accused of cheating. Each subject received a file that contained

- a letter from the chair of the Committee on Discipline
- a summary of the evidence against the suspected cheater
- background information on the suspect, including prior academic performance
- a color photo portraying one of the four following facial expressions

a "felt" smile

a false smile

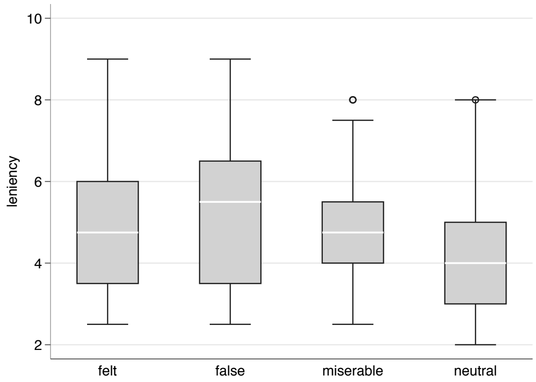
a miserable smile

a neutral expression

The subjects were then asked to indicate their judgments. They did this by answering 5 questions about the likelihood of the suspect's guilt and how severe the punishment should be. These questions were combined to form a single "leniency score" (where higher scores = less severe punishment)

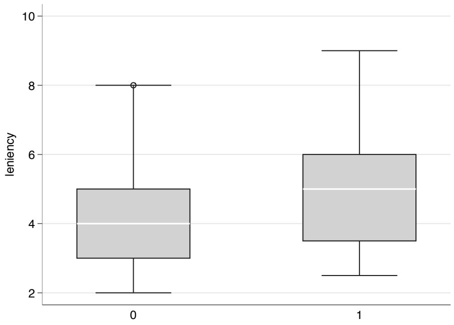
The following data were obtained:

```
           |          N      mean        sd
-----------+------------------------------
      felt |         34  4.911765  1.680866
     false |         34  5.367647  1.827023
 miserable |         34  4.911765  1.453682
   neutral |         34  4.117647  1.522850
-----------+------------------------------
     Total |        136  4.827206  1.671525
------------------------------------------
```



If we combine the first three groups into a "smile" group, our data are:

```
    smile |          N      mean        sd
----------+------------------------------
        0 |         34  4.117647   1.52285
        1 |        102  5.063725  1.658568
----------+------------------------------
    Total |        136  4.827206  1.671525
-----------------------------------------
```



LaFrance, M., & Hecht, M. A. (1995). Why smiles generate leniency. Personality and Social Psychology Bulletin, 21, 207-214.

10) Let's first compare the combined smile group to the neutral expression. State the null and alternate hypotheses. Express the consequences of both Type I and Type II errors in this study.

Null: ______________________________________ Alternate: ______________________________________

Type I error consequence: ____

Type II error consequence: ____

11) Consider the assumptions necessary for us to conduct an independent samples t-test. Are these assumptions reasonable in this situation? How can we check these assumptions?

12) Calculate a 99% confidence interval for the difference in means between the smile and neutral groups. What conclusions can you make?

13) Sketch the sampling distribution of mean differences under your null hypothesis. Label the mean and standard error, identify the critical value, and shade-in the rejection region. Calculate your observed test statistic and make a conclusion.

14) The observed difference in means is 0.946078. Calculate and interpret a p-value in this scenario.

15) Here's the output when I conducted an independent samples t-test in Stata. Does this match your calculations?

```
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      34    4.117647    .2611667     1.52285    3.586299    4.648995
       1 |     102    5.063725    .1642227    1.658568    4.737952    5.389499
---------+--------------------------------------------------------------------
combined |     136    4.827206    .1433321    1.671525    4.543739    5.110673
---------+--------------------------------------------------------------------
    diff |           -.9460784     .322035               -1.583007   -.3091494
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -2.9378
Ho: diff = 0                                     degrees of freedom =      134

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0019         Pr(|T| > |t|) = 0.0039          Pr(T > t) = 0.9981
```

16) Suppose we wanted to test the difference between means of the "felt" and "false" smile groups. Conduct this test on your calculator (or look at the Stata output below) and state any conclusion(s) you can make.

```
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    felt |      34    4.911765    .2882662    1.680866    4.325283    5.498247
   false |      34    5.367647    .3133318    1.827023    4.730169    6.005125
---------+--------------------------------------------------------------------
combined |      68    5.139706    .2131142    1.757384    4.714328    5.565084
---------+--------------------------------------------------------------------
    diff |           -.4558824    .4257631               -1.305946    .3941812
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -1.0707
Ho: diff = 0                                     degrees of freedom =       66

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1441         Pr(|T| > |t|) = 0.2882          Pr(T > t) = 0.8559
```

17) Based on this p-value, can we conclude there is no difference in leniency between the felt and false smile groups?

18) With 4 groups in this study, how could we use t-tests to compare the means of all 4 groups? How many t-tests would we need to conduct?

19) Suppose we set $\alpha = 0.05$ for each of our t-tests. If we conducted all those t-tests, what would be the overall probability that we would make at least one α-error across all our tests? How could we reduce the chances of making an α-error?

20) Let's generalize the results of our answers to the previous two questions. Suppose we have a study with G groups. If we conduct t-tests to compare all possible pairs of means, what would be our overall α-error rate? What are the implications of this?

Scenario: Researchers have established that sleep deprivation has a harmful effect on visual learning, but is it possible to "make up" for sleep deprivation by getting a full night's sleep in subsequent nights? A 2000 study investigated this question by giving a visual discrimination task to 21 subjects. Afterwards, the 21 subjects (volunteers between the ages of 18 and 25) were randomly assigned to one of two groups:

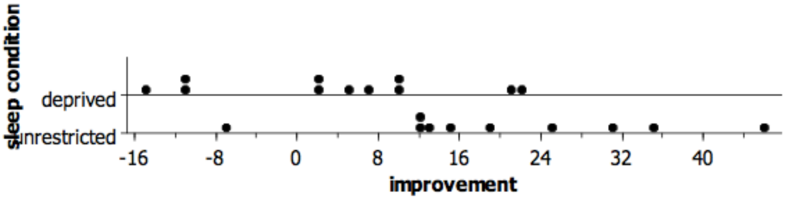
- one group was deprived of sleep the night following the test
- the other group was permitted unrestricted sleep

Both groups were then allowed as much sleep as they wanted on the following two nights. On the 3rd day, all 21 subjects were given another visual discrimination task.

Subjects' performance on the test was recorded as the minimum time (in milli-seconds) between stimuli appearing on a computer screen for which they could accurately report what they had seen on the screen. The sorted data and dotplots presented here are the improvements in those reporting times between the pre-test and post-test (a negative value indicates a decrease in performance):

n=11 sleep deprived:	-14.7	-10.7	-10.7	2.2	2.4	4.5	7.2	9.6	10.0	21.3	21.8
n=10 unrestricted:	-7.0	11.6	12.1	12.6	14.5	18.6	25.2	30.5	34.5	45.6	

```
deprived |          N      mean         sd
---------+--------------------------------
      No |         10     19.82   14.72532
Deprived |         11      3.90   12.17185
---------+--------------------------------
   Total |         21  11.48095   15.42827
------------------------------------------
```



Allan Rossman: http://statweb.calpoly.edu/arossman/stat325/notes.html

21) Based on the dotplot and summary statistics, it appears as though the subjects who got unrestricted sleep the first night outperformed the subjects who were deprived of sleep. In fact...

- The difference in means was 19.82 – 3.90 = 15.92
- The difference in medians was 16.55 – 4.5 = 12.05
- 9 out of the 10 lowest gain scores belong to the subjects who were sleep deprived

With all this in mind, is it possible that there really is no harmful effect of sleep deprivation and that random chance produced the differences we observed between these two groups? What would it mean if we were to find a *significant* difference?

22) We could type this data into a calculator or computer and conduct an independent samples t-test. Do you think the necessary assumptions for an independent samples t-test are satisfied in this scenario? Explain.

23) Most students who are completing their first statistics class, when presented with this scenario, would conduct an independent samples t-test. Below, I've pasted output from a t-test. From this, what conclusions could we make?

```
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      10       19.82    4.656556    14.72532    9.286139    30.35386
       1 |      11         3.9    3.669952    12.17185   -4.277162    12.07716
---------+--------------------------------------------------------------------
combined |      21    11.48095    3.366725    15.42827    4.458087    18.50382
---------+--------------------------------------------------------------------
    diff |               15.92    5.873229                 3.62719    28.21281
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.7106
Ho: diff = 0                                     degrees of freedom =       19

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9931         Pr(|T| > |t|) = 0.0139          Pr(T > t) = 0.0069
```

24) I want to be clear – **you should not conduct an independent samples t-test (or any other parametric test) without first investigating assumptions and your data!** If you are worried that some assumptions may not be satisfied, you have several options (such as running a different test with different assumptions, using a nonparametric test, transforming your data, or not running a test at all).

In this activity, I want to present you with one alternative to the t-test. Remember the first day of class this semester when we dealt with a study to determine if dolphins could communicate? On that first day, we wrote out hypotheses, conducted a test, estimated a p-value, and wrote out conclusions. We did all of that without even knowing what a t-test was. Instead, we used the concept of randomization.

We've already stated that even though our sample data indicate the unrestricted group outperformed the sleep deprived group, it's possible the differences were just due to random chance. Our key question, therefore, is:

How likely were we to observe data that favor the unrestricted group by at least as much as our data suggest?

To address this question, we will use a simulation process like we did at the beginning of the semester. We will:

- Randomize: We will assume sleep deprivation has no negative effect (the null model) and replicate the random assignment of the 21 subjects (and their improvement scores) between the two groups

- Repeat: We will repeat this random assignment a large number of times and calculate a measure of how different the groups are in order to get a sense for what is expected and what is surprising

- Reject?: If the results we observed in our study are in the tail of the null model's distribution, we will reject that null model

In this analysis, we will calculate the difference in means between the two groups after each new random assignment. If we do this a large number of times, we will have a good idea of whether the difference in means we observed is surprising under our null model of no real difference between the two groups. Note that we could just as easily use the medians instead of the means, which is a nice feature of this analysis strategy.

25) We could conduct this simulation by hand by:

- Taking 21 index cards and writing one observed data value on each card
- Shuffling the cards and randomly select 11 to represent the sleep deprivation group
- Calculating the mean of the 11 cards you randomly selected and the mean of the 10 unselected cards
- Calculating the difference in means
- Repeating this process thousands of times
- Determining the proportion of results that are at least as extreme as what we observed from the data

We could also, thankfully, use technology to greatly speed up this process. We'll use a computer to:

- Randomly assign groups to the 21 improvement scores (11 sleep deprived and 10 unrestricted sleep)
- Calculate the difference in group means
- Store that difference
- Repeat 1000 times
- Display the results and calculate the proportion of results that are at least as extreme as what was observed

Open the *randomization applet* at http://www.rossmanchance.com/applets/randomization20/Randomization.html. You'll notice the data from this study are already entered into the applet. You can see, once again, that the observed difference in means was found to be 15.92.

a) Click on **re-randomize** to randomize the 21 improvement scores between the two groups. Notice the new difference between group means (calculated from this randomization). Click **re-randomize** one more time to ensure you get a different difference between group means (although it is possible for you to get the same difference twice).

b) Now un-check the **animate** feature and run 998 more randomizations. Look at the distribution of the 1000 simulated differences in group means. Is the center where you would expect? Does the shape have a recognizable pattern?

c) Count how many of your 1000 simulated differences in group means are as extreme (or more extreme) than what the researchers actually observed (greater than 15.92). To do this, you can enter 15.92 in the **count samples beyond** field and click **GO**. What approximate p-value did you obtain from your simulation?

26) Does this simulation analysis provide strong evidence that sleep deprivation has harmful effects three days later?

27) Even if you found a statistically significant difference between the mean improvement scores of the two groups, is it legitimate to draw a cause-and-effect conclusion between sleep deprivation and lower improvement scores? Explain. Was this an observational or experimental study?

28) On my computer, I could use Stata to run the same randomization analysis. The code I would use for 10,000 randomizations would be:

```
permute score diff=(r(mu_1)-r(mu_2)), reps(10000): ttest score, by (deprived)
```

The output I obtained is pasted below, along with a display of the results from my 10,000 simulations. Interpret the results. What conclusions can you make?

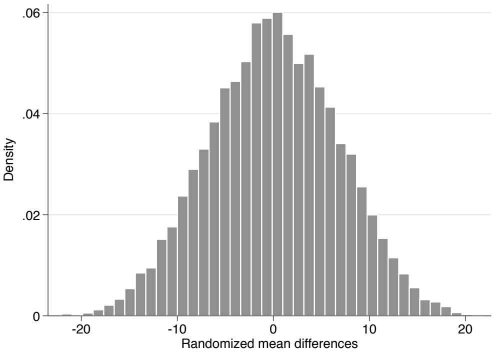
observed mean diff.

n=10,000 replications

p-value

```
------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
        diff |      15.92     136   10000  0.0136  0.0012  .0114224   .0160672
------------------------------------------------------------------------------
Note:  confidence interval is with respect to p=c/n.
Note:  c = #{|T| >= |T(obs)|}
```

C = 136. This tells us 136 of the 10,000 replications resulted in mean differences at least as big as 15.92. So, 1.36% of our replications gave us results as or more extreme than what we observed.



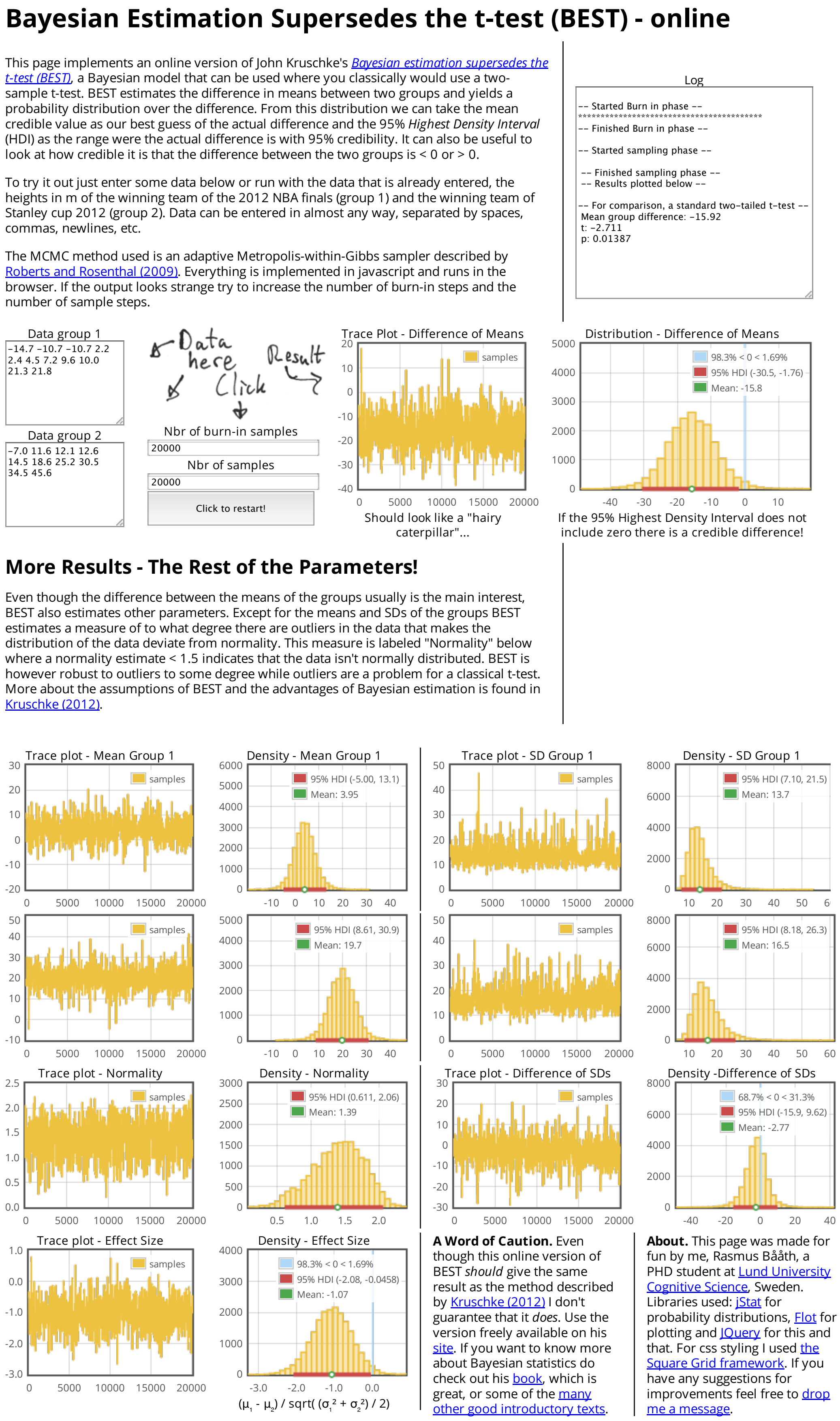
29) Instead of running 10,000 simulations (or more), we could also list out every possible randomization for this study. If we have 21 subjects, in how many ways could we randomly assign 11 subjects to one group and 10 to another?

Answer = 352,716. How did I calculate this? Are each of these randomizations equally likely to occur?

We could then see how many of these randomizations give us results as or more extreme than 15.92. It turns out that number is 2,533. From this, calculate the exact p-value. Do our p-values from the t-test, the randomization method, and the exact calculation all agree? Which method is *best*?

30) Go back to the website and run a simulation to determine if the group medians differ significantly. Report the p-value and write out your conclusions.

Below, I've pasted results from a Bayesian approach to the t-test. If we have time, I'll explain what's going on and the advantages to this Bayesian approach. The website I used for this was: http://www.sumsar.net/best_online/



Bayesian Estimation Supersedes the t-test (BEST) - online

This page implements an online version of John Kruschke's *Bayesian estimation supersedes the t-test (BEST)*, a Bayesian model that can be used where you classically would use a two-sample t-test. BEST estimates the difference in means between two groups and yields a probability distribution over the difference. From this distribution we can take the mean credible value as our best guess of the actual difference and the 95% *Highest Density Interval* (HDI) as the range were the actual difference is with 95% credibility. It can also be useful to look at how credible it is that the difference between the two groups is < 0 or > 0.

To try it out just enter some data below or run with the data that is already entered, the heights in m of the winning team of the 2012 NBA finals (group 1) and the winning team of Stanley cup 2012 (group 2). Data can be entered in almost any way, separated by spaces, commas, newlines, etc.

The MCMC method used is an adaptive Metropolis-within-Gibbs sampler described by Roberts and Rosenthal (2009). Everything is implemented in javascript and runs in the browser. If the output looks strange try to increase the number of burn-in steps and the number of sample steps.

Log

```
-- Started Burn in phase --
****************************************
-- Finished Burn in phase --

-- Started sampling phase --

 -- Finished sampling phase --
 -- Results plotted below --

-- For comparison, a standard two-tailed t-test --
Mean group difference: -15.92
t: -2.711
p: 0.01387
```

Data group 1

```
-14.7 -10.7 -10.7 2.2
2.4 4.5 7.2 9.6 10.0
21.3 21.8
```

Data group 2

```
-7.0 11.6 12.1 12.6
14.5 18.6 25.2 30.5
34.5 45.6
```

Data here

Click

Result

Nbr of burn-in samples

20000

Nbr of samples

20000

Click to restart!

Trace Plot - Difference of Means

Should look like a "hairy caterpillar"...

Distribution - Difference of Means

If the 95% Highest Density Interval does not include zero there is a credible difference!

More Results - The Rest of the Parameters!

Even though the difference between the means of the groups usually is the main interest, BEST also estimates other parameters. Except for the means and SDs of the groups BEST estimates a measure of to what degree there are outliers in the data that makes the distribution of the data deviate from normality. This measure is labeled "Normality" below where a normality estimate < 1.5 indicates that the data isn't normally distributed. BEST is however robust to outliers to some degree while outliers are a problem for a classical t-test. More about the assumptions of BEST and the advantages of Bayesian estimation is found in Kruschke (2012).

Trace plot - Mean Group 1

Density - Mean Group 1

Trace plot - SD Group 1

Density - SD Group 1

Trace plot - Normality

Density - Normality

Trace plot - Difference of SDs

Density -Difference of SDs

Trace plot - Effect Size

Density - Effect Size

$(\mu_1 - \mu_2) / \mathrm{sqrt}((\sigma_1^2 + \sigma_2^2) / 2)$

A Word of Caution. Even though this online version of BEST *should* give the same result as the method described by Kruschke (2012) I don't guarantee that it *does*. Use the version freely available on his site. If you want to know more about Bayesian statistics do check out his book, which is great, or some of the many other good introductory texts.

About. This page was made for fun by me, Rasmus Bååth, a PHD student at Lund University Cognitive Science, Sweden. Libraries used: jStat for probability distributions, Flot for plotting and JQuery for this and that. For css styling I used the Square Grid framework. If you have any suggestions for improvements feel free to drop me a message.