Activity 18: Comparing two treatment groups

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

Doctors, 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.

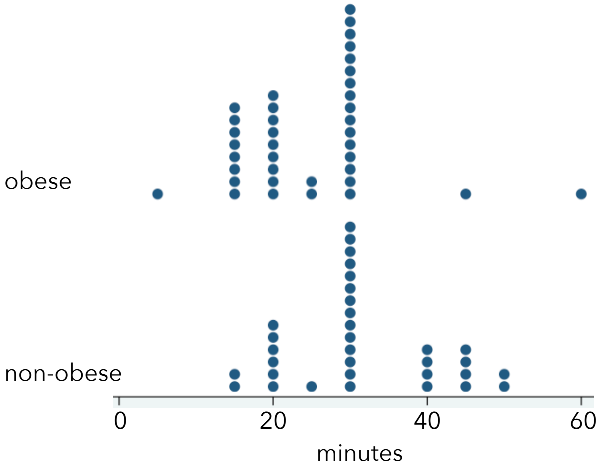
A 2001 study examined physicians' behavioral intentions towards average-weight and obese patients. 71 primary care physicians in Houston were sent a packet containing a medical chart similar to the one they view upon seeing a new patient. This chart portrayed a patient who was displaying symptoms of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

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

After reviewing the chart, the doctors were asked to indicate how much time they would spend with the patient. **Do doctors indicate they would spend less time with obese patients?**

Here's a summary of the data from this study:

```
  weight |   N      mean   median         sd
---------+--------------------------------
non-obese|  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.

In this scenario, we may want to investigate a couple different questions using a variety of analysis methods:

General question: **By how much do the treatment groups differ?**

In this scenario: How much less time do doctors spend with obese patients?

Method: Effect size

Options: Cohen's d (effect size based on difference between two means)
D-statistic (effect size based on discrimination of individual objects from two groups)

Method: Confidence interval

Options: Bootstrap methods (perhaps for $\mu_1 - \mu_2$ or median$_1$ - median$_2$)
Theory-based (perhaps using t-distribution with/without equal variances assumption)

Question: **Do the treatment groups differ?**

In this scenario: Do doctors spend less time with obese patients?

Method: Null hypothesis significance test

Options: Randomization-based test (perhaps for $\mu_1 - \mu_2$ or median$_1$ - median$_2$)
t-test to test H_0: $\mu_1 - \mu_2 = 0$ with equal variances assumption
Welch-Sattertwaite method to test H_0: $\mu_1 - \mu_2 = 0$ without equal variances assumption

1) Suppose we want to estimate the magnitude of the difference between the two groups. One way to do this would be to estimate the standardized difference between the group means: $\delta = \dfrac{\overline{X}_1 - \overline{X}_2}{s}$

Calculate and interpret this effect size (Cohen's d):

```
   weight |    N       mean          sd
---------+------------------------
non-obese|   33   31.36364    9.864134
   obese |   38   24.73684    9.652571
---------+------------------------
   Total |   71   27.81690   10.237620
---------+------------------------
    Diff |         6.62680
```

Effect size = ______________________________

Interpretation: ______________________________

____

While interpreting the magnitude of Cohen's d depends on its substantive context, Cohen did provide some widely used rules-of-thumb: **Small effect: 0.20 ≤ d ≤ 0.30** **Medium effect: d ≈ 0.50** **Large effect: d > 0.80**

Source: Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (second ed.). Lawrence Erlbaum Associates.

2) Instead of estimating the differences between group means, we may want to estimate the difference between individual observations in each group. One way to do this would be to estimate the probability that an individual observation in one group scores higher than an observation in the other group: $P(x_1 > x_2) = P(x_1 - x_2 > 0)$

To estimate this probability, we need to derive some characteristics of the distribution of $x_1 - x_2$.

- The expected value of $x_1 - x_2$ would be:

$$E[x_1 - x_2] =$$

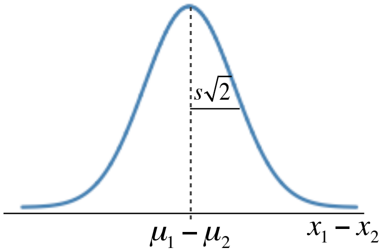
- If we assume the groups are independent and have equal population variances, then:

$$\text{var}[x_1 - x_2] =$$

and

$$SD[x_1 - x_2] =$$

- Assuming the observations from both groups come from populations with normal distributions, calculate D:



$$D = P(x_1 - x_2 > 0) = P\left(z > \frac{0 - (\overline{X}_1 - \overline{X}_2)}{s\sqrt{2}}\right) = P\left(z > \frac{\overline{X}_2 - \overline{X}_1}{s\sqrt{2}}\right) = P\left(z < \frac{\overline{X}_1 - \overline{X}_2}{s\sqrt{2}}\right) =$$

$$P\left(z < \frac{6.6268}{9.75\sqrt{2}}\right) = P(z < 0.481) =$$

Interpretation: ______________________________

3) This time, let's calculate confidence intervals – intervals that will frequently include the value of $\mu_{\text{obese}} - \mu_{\text{non-obese}}$.

To calculate a confidence interval, we can use bootstrap methods or theory-based (parametric) formulas.

Let's first construct a 90% confidence interval for $\mu_{\text{obese}} - \mu_{\text{non-obese}}$ using bootstrap methods. To do this:

a) Copy the data from: http://www.bradthiessen.com/html5/data/doctors.csv
b) Paste it into: http://lock5stat.com/statkey/bootstrap_1_quant_1_cat/bootstrap_1_quant_1_cat.html

Construct a 90% confidence interval based on at least 10,000 bootstrap samples. Record this interval, interpret it, and explain the bootstrap method.

90% confidence interval: ______________________________

Interpretation: ______________________________

Explanation of bootstrap method: ______________________________

____

4) Now, let's construct a 90% confidence interval for $\mu_{\text{obese}} - \mu_{\text{non-obese}}$ using theory-based methods.

Remember that the general form of our confidence intervals has been: $\text{Estimate} \pm (\text{a number of})(\text{standard errors})$

In the previous activity, we derived: $\text{SE}_{\text{pooled}} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

Construct the 90% confidence interval, indicating the degrees-of-freedom for your t-statistic.

90% confidence interval: ______________________________

5) Think about the assumptions you need to make (or what conditions were necessary) to construct each interval.

Assumptions for theory-based interval: ______________________________

Assumptions for bootstrap interval: ______________________________

6) Do your confidence intervals provide evidence to suggest doctors do spend less time with obese patients? Explain.

7) One advantage of bootstrap methods is they can easily be extended to construct intervals for other parameters.

Using the same data and applet, let's use bootstrap methods to construct a 90% confidence interval for the difference between the group **medians**.

a) Copy the data from: http://www.bradthiessen.com/html5/data/doctors.csv
b) Paste it into: http://lock5stat.com/statkey/bootstrap_1_quant_1_cat/bootstrap_1_quant_1_cat.html

90% confidence interval: ____

8) Now let's turn our attention to addressing the question: Do doctors spend less time with obese patients?

To address this question, we're going to compare the means of the obese and non-obese groups. Fill-in-the-blanks:

Null hypothesis: ______________________________

Alternate hypothesis: ______________________________

Type I error consequence: ____

Type II error consequence: ____

9) Looking at our sample data, the 71 doctors indicated they would spend 6.6268 fewer minutes with obese patients. Explain why we can't simply look at our data, reject our null hypothesis, and conclude that doctors spend less time with obese patients?

10) Consider the assumptions necessary for us to conduct an independent samples t-test.

Assumption	Why do we need this assumption?	Is it a reasonable assumption in this scenario?

11) Suppose we do not think the normality and equal variances assumptions are reasonable in this scenario. Instead of running an independent samples t-test, we might employ randomization-based methods.

If you remember the first day of class (the dolphin study), we wrote out hypotheses, conducted a test, estimated a p-value, and stated a conclusion all without knowing anything about probability distributions or t-tests. To do that, we used the concept of randomization.

Even though our sample data indicate doctors spend less time with obese patients, it's possible we could have obtained those results in our sample even if (the population of) doctors really spend equal time with all patients.

Our key question, therefore, is: How likely were we to observe our sample data (or something more extreme) if, in fact, the weight of patients has no effect on the time doctors spend with them?

To address this question, we'll randomize our data and estimate a p-value:

a) **Randomize**: Assuming obesity has no effect (the null model), we'll replicate the random assignment of 71 doctors into the obese and non-obese groups. We will randomly assign 33 doctors to the *non-obese* group and 38 doctors to the *obese* group. Then, we'll calculate our statistic of interest: $\bar{X}_{\text{non-obese}} - \bar{X}_{\text{obese}}$
Note that we could choose another statistic to compare the two groups.

b) **Repeat**: We'll repeat this randomization process many times. Note that there are a huge number of possible randomizations (ways of splitting 71 doctors into groups of size 33 and 38):

$$\binom{71}{33} = \frac{71!}{(71-33)!33!} = 187,265,264,199,657,100,730$$

That would take too long, so we'll get a representative sample of at least 10,000 randomizations. For each replication, we'll calculate $\bar{X}_{\text{non-obese}} - \bar{X}_{\text{obese}}$ to get a sense of what values are typical if obesity does not matter.

c) **Reject?** The 10,000 randomizations represent typical results if our null hypothesis were true. From these randomizations, we can determine how unusual our actual results from the study were. If the actual results from the study look unusual, we can reject the null hypothesis

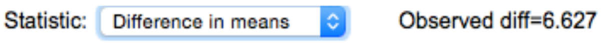
The following table attempts to explain this randomization process for our scenario. As you can see, the first doctor was randomly assigned to the non-obese group and reported spending 15 minutes with that patient. If we could go back in time and, once again, randomly assign this doctor a patient's chart, this doctor might be assigned an obese patient. How would this change the amount of time the doctor would spend with the patient?

Subject	**Actual Data**		**Randomization #1**		**Randomization #2**		**Randomization #3**	
1	Non-obese	15	Obese	_____	Non-obese	15	Obese	15
2	Average	45	Non-obese	_____	Obese	45	Obese	45
3	Non-obese	30	Non-obese	_____	Obese	30	Non-obese	30
…	…	…	…	…	…	…	…	…
71	Obese	60	Non-obese	_____	Obese	60	Non-obese	60
Means	Average = 31.36 Obese = 24.74		Average = 27.58 Obese = 28.03		Average = 28.33 Obese = 27.37		Average = 28.94 Obese = 26.84	
Difference	Avg - Obese = +6.62		Avg - Obese = -0.45		Avg - Obese = +0.96		Avg - Obese = +2.10	

12) Let's conduct this randomization-based test:

a) Copy the data from: http://www.bradthiessen.com/html5/data/doctors.csv
b) Open the applet: http://www.rossmanchance.com/applets/AnovaShuffle.htm?hideExtras=2
c) Paste the data into the box on the top-left and click the **USE DATA** button

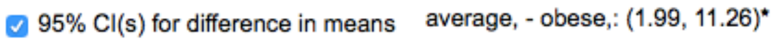
d) We want to calculate the difference between means:



Statistic: Difference in means

Observed diff=6.627

e) At the bottom, you could check a box to get a 95% confidence interval for this difference.

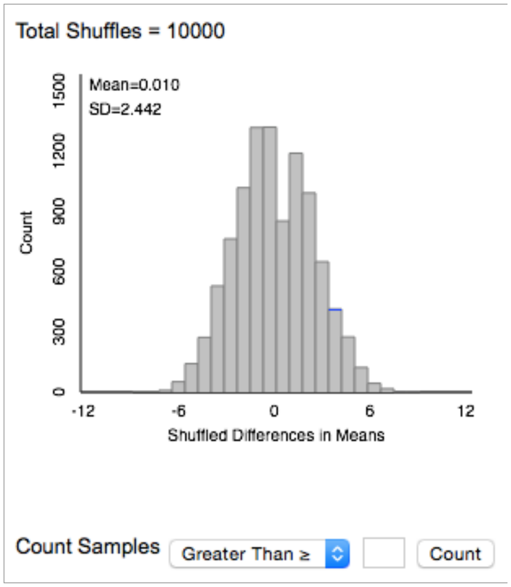


☑ 95% CI(s) for difference in means average, - obese,: (1.99, 11.26)*

f) Check the **SHOW SHUFFLE OPTIONS** box on the top-right and click **SHUFFLE RESPONSES**

g) Change the number of shuffles to **9999**, click **SHUFFLE RESPONSES**, and run all 10,000 randomizations

h) You now see the randomization distribution (similar to what's pasted below).



Total Shuffles = 10000

Mean=0.010

SD=2.442

i) Look at the center of this distribution. Is the center where you'd expect it to be? Explain.

ii) Is the shape of this distribution what you'd expect? Explain.

g) Explain what we're going to do to estimate a p-value. Remember, a p-value tells us the likelihood of observing results as or more extreme than what we observed if, in fact, the null hypothesis were true.

p-value = _______________

13) If you prefer, you can use a different applet to run this randomization-based test.
a) Copy the data from: http://www.bradthiessen.com/html5/data/doctors.csv
b) Applet: http://lock5stat.com/statkey/randomization_1_quant_1_cat/randomization_1_quant_1_cat.html
h) Click the **EDIT DATA** button, paste the data into the box, and click **OK**
i) The randomization method we're using is **REALLOCATE GROUPS**
j) Click the **GENERATE 1000 SAMPLES** button 10 times to get 10,000 randomizations
k) To estimate a p-value, check the **RIGHT-TAIL** box, click the number on the x-axis, and change it to 6.627
l) Record the p-value below. Should it match the p-value from the previous applet? Explain.

p-value = _______________

14) Based on our estimated p-value(s), what conclusions can we make from this study?

15) Now, let's finally conduct an independent samples t-test to compare our group means.
Remember our null hypothesis is: H_0: $\mu_{average} = \mu_{obese.}$

Assuming this null hypothesis is true (and our assumptions of normality, equal variances, and independence hold), sketch the sampling distribution we would get if we repeatedly took samples of size 33 and 38 and calculated the difference between the means of those samples. Label the mean and standard error of this distribution. Then, identify the critical value and shade-in the rejection region. Finally, estimate the p-value.

To get the p-value, you may wish to use: http://lock5stat.com/statkey/theoretical_distribution/theoretical_distribution.html#t

```
  weight |   N      mean         sd
---------+-------------------------
non-obese|  33  31.36364   9.864134
   obese |  38  24.73684   9.652571
---------+-------------------------
   Total |  71  27.81690  10.237620
---------+-------------------------
    Diff |       6.62680
```

16) Assuming doctors actually spend 5 minutes less with obese patients, estimate the power of our t-test.

17) I conducted this independent samples t-test on a computer (using a program called *Stata*). Interpret this output. Does it match your calculations from the previous question?

```
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
---------+--------------------------------------------------------------------
    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
```

18) Try conducting an independent samples t-test on the following applet. Why do we get different results?

a) Copy the data from: http://www.bradthiessen.com/html5/data/doctors.csv
b) Open the applet: http://www.rossmanchance.com/applets/TBIA.html

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 col

a "felt" smile

a false smile

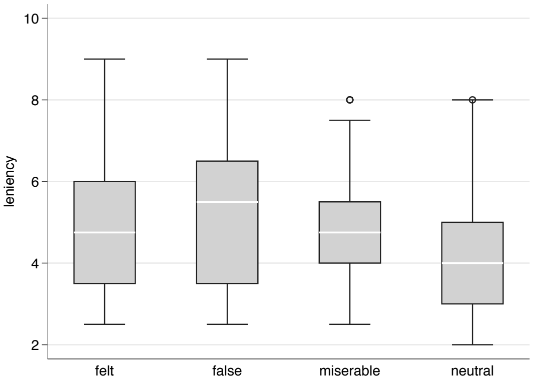
a miserable smile

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)

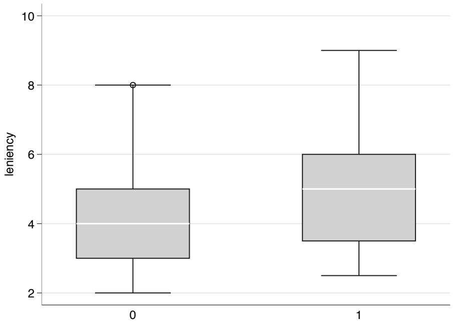
The following data were obtained:

```
          |         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

```
   smile |         N      mean        sd
---------+------------------------------
  0 (no) |        34  4.117647   1.52285
 1 (yes) |       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.

19) 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: ____

20) 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?

21) Using both parametric and bootstrap methods, construct a 99% confidence interval for the difference in means between the smile and neutral groups. What conclusions can you make?

Data: http://www.bradthiessen.com/html5/data/smiles.csv
Bootstrap applet: http://lock5stat.com/statkey/bootstrap_1_quant_1_cat/bootstrap_1_quant_1_cat.html

Parametric CI:

Bootstrap CI:

22) Use one of the following applets to conduct a randomization-based test for the difference in means. Record the p-value and write out any conclusions you can make.

Applet 1: http://www.rossmanchance.com/applets/AnovaShuffle.htm?hideExtras=2
Applet 2: http://lock5stat.com/statkey/randomization_1_quant_1_cat/randomization_1_quant_1_cat.html

p-value = _______________ Conclusions:

23) 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.

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

25) 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
```

26) Suppose we wanted to test the difference between means of the “felt” and “false” smile groups. Look at the output pasted 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
```

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

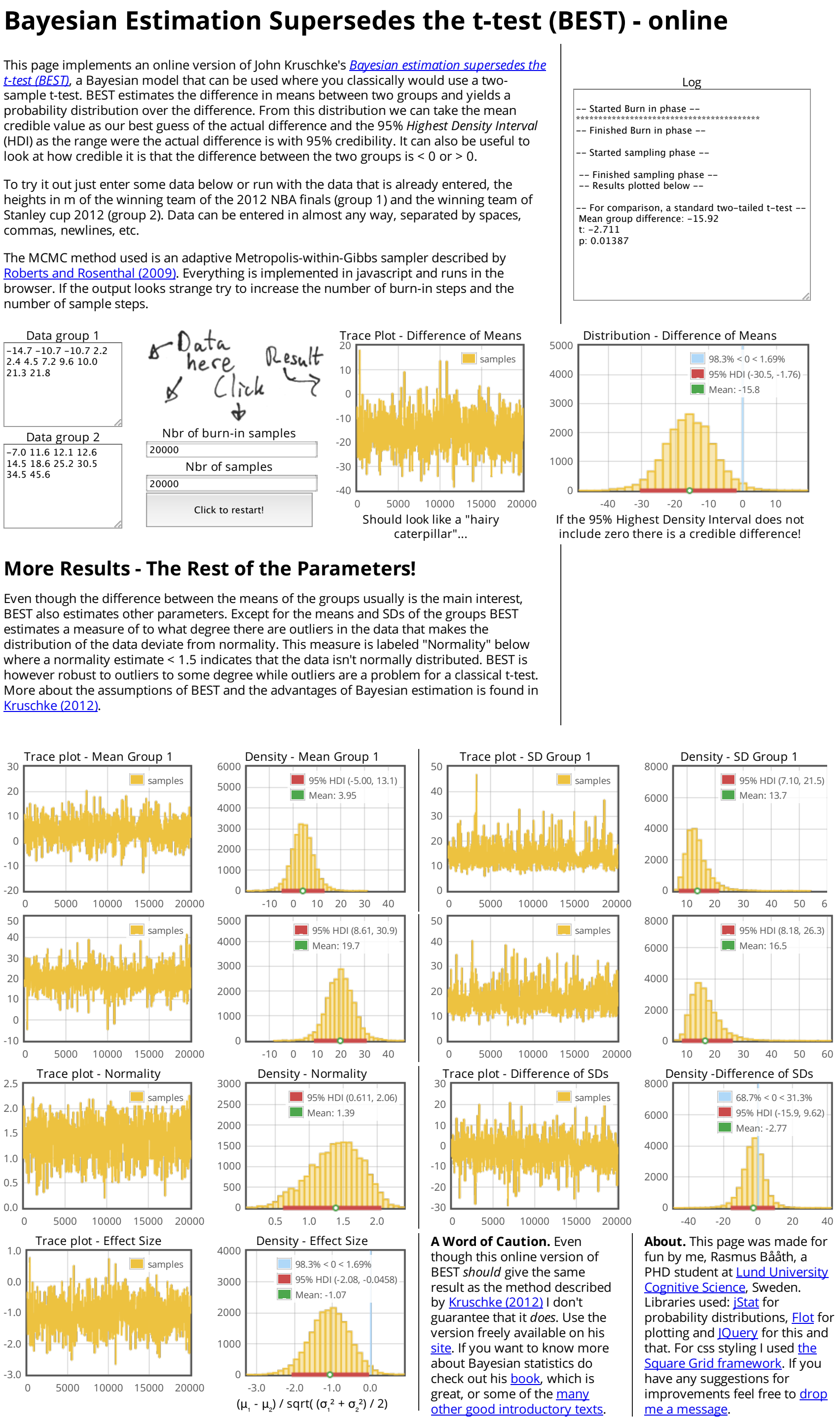
28) Is there any way we could use our independent samples t-test to compare all 4 groups in this study? How many t-tests would we need to conduct to test all pairs of group means?

29) Suppose we set α = 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?

30) 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?

31) How could we compare the means from 3+ groups using randomization- or theory-based methods?

To the right, 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
```

Nbr of burn-in samples: 20000

Nbr of samples: 20000

Click to restart!

Should look like a "hairy caterpillar"...

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).

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.