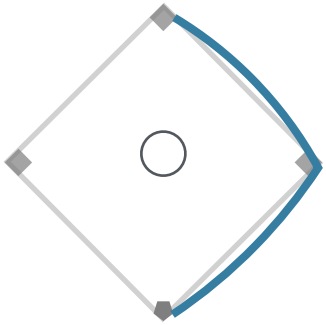
Activity #19: Group comparisons continued Name: _____

Dependent / matched-pairs designs, introduction to ANOVA, NHST limitations, effect sizes, Bayesian methods

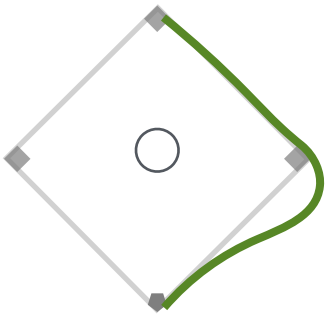
In the last activity, we learned how to use theory-based methods (t-test) to test for a difference between two independent group means. We also learned how to use randomization-based methods to compare independent group means and medians. In this activity, we'll learn more about group comparisons, including:

- How can we compare means if the two groups are dependent?
- How can we compare the means of more than two groups?
- What are some limitations with this null hypothesis significance testing?
- How can we estimate the size of the difference between two group means?
- How can we estimate the probability that two group means differ?

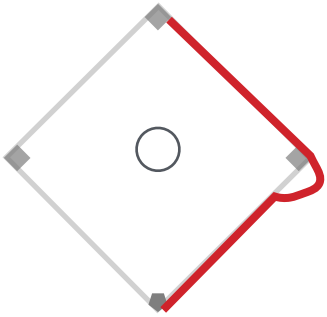
Scenario: In baseball, does the path you take to round first base influence the time it takes to reach second base? In 1970, W. F. Woodward timed how long it took 22 baseball players to run from home plate to second base. Each player ran a total of six times. Using a randomized order, these six trials per player were evenly divided (two each) among three methods:



Narrow-angle method



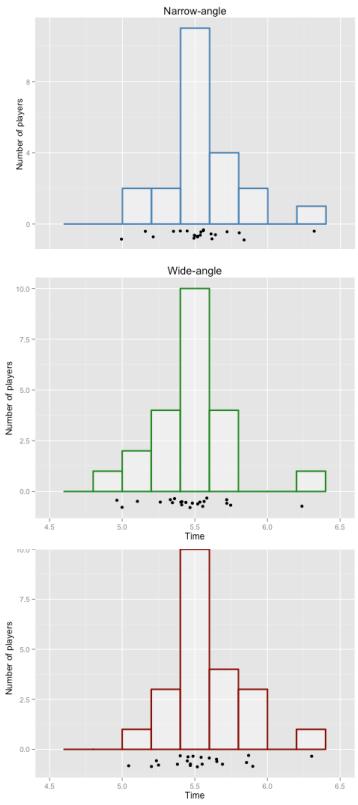
Wide-angle method



Round-out method

The table below displays the average times of the two runs per method.

Player	Narrow	Wide	Round-out
1	5.50	5.55	5.40
2	5.70	5.75	5.85
3	5.60	5.50	5.20
4	5.50	5.40	5.55
5	5.85	5.70	5.90
6	5.55	5.60	5.45
7	5.40	5.35	5.40
8	5.50	5.35	5.45
9	5.15	5.00	5.25
10	5.80	5.70	5.85
11	5.20	5.10	5.25
12	5.55	5.45	5.65
13	5.35	5.45	5.60
14	5.00	4.95	5.05
15	5.50	5.40	5.50
16	5.55	5.50	5.45
17	5.55	5.35	5.55
18	5.50	5.55	5.45
19	5.45	5.25	5.50
20	5.60	5.40	5.65
21	5.65	5.55	5.70
22	6.30	6.25	6.30
Avg	5.5341	5.4591	5.5432
Std. Dev	0.25976	0.27283	0.27181



Source: Woodward, W. F. (1970). A comparison of base running methods in baseball. Master's thesis, Florida State University.

1. In this study, we have 3 treatment groups: narrow-angle, wide-angle, and round-out. Suppose we compared each possible pair of group means. If we did that, we'd need to run 3 tests (theory- or randomization-based):

   Narrow-angle vs. Wide-angle Narrow-angle vs. Round-out Wide-angle vs Round-out.

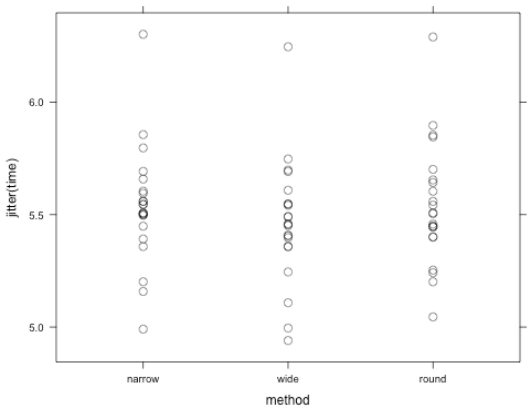
   Suppose we conducted each test using $\alpha = 0.05$. Over all three tests, what would be the probability that we would make at least one Type I error?

   P(at least one alpha error) = _____

2. Suppose we wanted to conduct all 3 tests yet keep a 0.05 chance of making at least one Type I error. At what level would we need to set α in each test in order to ensure a *family-wise α-error rate* of 0.05?

   Set α in each test equal to: _____

3. Let's try to compare all 3 group means in a single test. If you take MATH/STAT 301, you'll learn about ANOVA (analysis of variance). For now, we'll figure out a randomization-based method.

   To do this, we need to calculate a single test statistic that, ideally:
   - Compares all 3 means
   - Gets larger as the means differ more from each other

   One value that meet these conditions would be the *sum of absolute deviations* (or SAD):

   SAD = |mean1 – mean2| + |mean1 – mean3| + |mean2 – mean3|

   In our scenario:

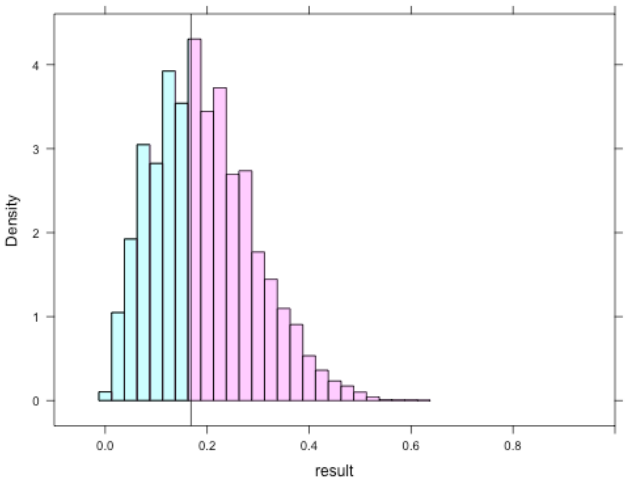
   SAD = |5.5341 – 5.4591| + |5.5341 – 5.5432| + |5.4591 – 5.5432| = 0.1682 (or MAD = .1682 / 3 = 0.056)

   Under a null hypothesis, we would expect the value of SAD to be: _____. Larger values of SAD would represent situations where the group means were farther apart from one another.

4. Now that we have our test statistic, we can use the randomization-based methods we've used since the first day of class to compare the methods. Under our null hypothesis, the running methods do not matter. If that's true, we can simulate what would happen if we went back in time and randomly assigned runners to different groups.

   Note that in this example, we're assuming we have 22 <u>independent</u> baseball players in each group. This is **<u>not</u>** true!

   I had a computer take the 66 values in our dataset and randomly assign 22 of them to each group. Then, I had the computer calculate the SAD. The 10,000 values of the SAD are displayed on the next page.



What does this distribution represent?

Why are values greater than 0.1682 shaded-in?

The proportion of SAD values > 0.1682 is **0.5731**. From this, what conclusions can we make?

Data: http://bradthiessen.com/html5/stats/m300/bball.txt
Applet: http://www.rossmanchance.com/applets/AnovaShuffle.htm?hideExtras=2

5. So that was one possible way of comparing all 3 group means in a single test. As mentioned earlier, we could also run 3 pairwise tests to compare all 3 group means (as long as we adjust our α-level for each test). Below, I've pasted the output from 3 independent samples t-tests.

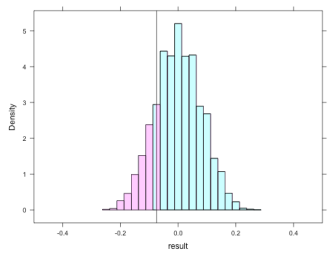
Narrow-angle vs. Wide-angle	Narrow-angle vs. Round-out	Wide-angle vs Round-out
H_1: narrow > wide	H_1: narrow < round	H_1: wide < round
t = 0.9338	t = –0.1134	t = –1.0242
df = 42	df = 42	df = 42
p-value = 0.1779	p-value = 0.4551	p-value = 0.1558
90% CI: (–0.06, 0.21)	90% CI: (–0.14, 0.13)	90% CI: (–0.22, 0.05)

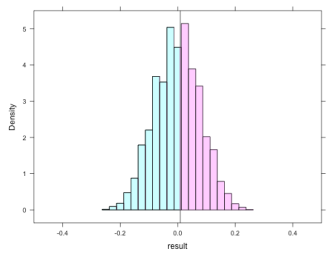
Verify the df values, interpret the output, and, using your adjusted α-level, draw a conclusion from each test.

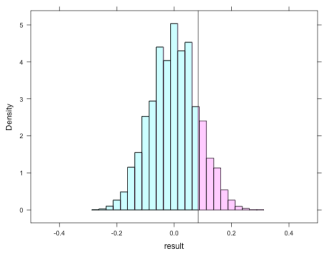
Narrow-angle vs. Wide-angle	Narrow-angle vs. Round-out	Wide-angle vs Round-out

6. We could have also conducted these tests using randomization-based methods. Interpret the following output. Do any of your conclusions change?

Narrow-angle vs. Wide-angle	Narrow-angle vs. Round-out	Wide-angle vs Round-out
Observed mean diff. = -0.0750	Observed mean diff. = 0.0091	Observed mean diff. = 0.0841
10,000 randomizations of method	10,000 randomizations of method	10,000 randomizations of method
P(diffs ≤ -0.0750) = 0.1854	P(diffs ≥ 0.0091) = 0.4641	P(diffs ≥ 0.0841) = 0.1591
Bootstrap 90% CI: (–0.21, 0.06)	Bootstrap 90% CI: (–0.12, 0.14)	Bootstrap 90% CI: (–0.05, 0.22)







7. Evaluate the assumptions we made in conducting the t-tests and the randomization-based comparisons. Are the assumptions reasonably satisfied?

   Assumptions for t-tests: _____

   Assumptions for randomization-based methods: _____

   Are either set of assumptions satisfied? _____

8. All the analyses we've conducted in this activity have been based on incorrect assumptions, therefore, the conclusions we've made are suspect. These methods would have been fine if the study had randomly assigned 66 (independent) baseball players to the 3 running methods. With the same 22 players across all 3 methods, we did not have independent groups. We cannot conduct an independent-samples t-test or this randomization-based method if we do not have independent groups.

   So how can we compare two group means if the groups are dependent (or, in this case, *matched-pairs*)? Let's take a look at the narrow-angle method and the wide-angle method:

Player	Narrow	Wide	Differences
1	5.50	5.55	-0.05
2	5.70	5.75	-0.05
3	5.60	5.50	0.10
4	5.50	5.40	0.10
5	5.85	5.70	0.15
6	5.55	5.60	-0.05
7	5.40	5.35	0.05
8	5.50	5.35	0.15
9	5.15	5.00	0.15
10	5.80	5.70	0.10
11	5.20	5.10	0.10
12	5.55	5.45	0.1
13	5.35	5.45	-0.1
14	5.00	4.95	0.05
15	5.50	5.40	0.10
16	5.55	5.50	0.05
17	5.55	5.35	0.2
18	5.50	5.55	-0.05
19	5.45	5.25	0.2
20	5.60	5.40	0.20
21	5.65	5.55	0.1
22	6.30	6.25	0.05
Avg	5.5341	5.4591	0.075
Std. Dev	0.25976	0.27283	0.0883

The 22 subjects within each group are independent, but the subjects across groups are not independent (they are the same subjects).

Therefore, we have 22 independent observations; not 44. We can calculate the differences in running times for each subject and treat those 22 differences as independent.

The last column of the table to the left displays the differences in running times for each subject.

Number of differences, n = 22
Average difference = 0.075
Std. deviation of differences = 0.0883.

Write out the null hypothesis for the average of these 22 differences: _____

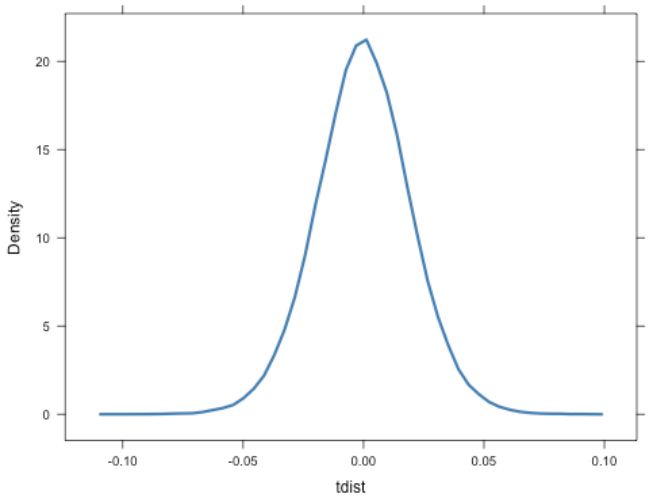
9. With our null hypothesis and our single sample of size n=22, we can conduct a single-sample t-test (like we did in activity #16).

   Below, I've sketched the distribution of sample means we would get if we repeatedly sampled 22 baseball players and calculated the difference in running times (narrow-angle minus wide-angle) for each player. Fill-in the blanks to describe this distribution:

   degrees of freedom = _____

   mean of sampling distribution = _____

   standard error = _____



10. The observed mean difference between the narrow- and wide-angle methods is 0.075. Locate this value on the above sampling distribution. The p-value, which you should verify, was found to be p = 0.00034. From this, what would you conclude?

11. Suppose the actual difference in means is 0.07 (in other words, the wide-angle method is 0.07 seconds faster). From this, we can estimate the power of the tests we've conducted:

    Independent samples t-test (from question #5), assuming:

    - Alpha = 0.05; difference = 0.07; n = 22: **Power = 0.217**
    - Alpha = 0.10; difference = 0.07; n = 22: **Power = 0.339**
    - Alpha = 0.05; difference = 0.10; n = 22: **Power = 0.340**
    - Alpha = 0.05; difference = 0.07; n = 44: **Power = 0.339**

    Matched-pairs t-test (from question #9-10), assuming:

    - Alpha = 0.05; difference = 0.07; n = 22: **Power = 0.974**
    - Alpha = 0.10; difference = 0.07; n = 22: **Power = 0.991**
    - Alpha = 0.05; difference = 0.10; n = 22: **Power = 0.9998**
    - Alpha = 0.05; difference = 0.07; n = 44: **Power = 0.9998**

    As we increase our alpha-level, the power of our test……..…… INCREASES DECREASES
    As we increase the difference between our group means, the power……INCREASES DECREASES
    As we increase our sample size, the power of our test……..…… INCREASES DECREASES

    A matched-pairs test has HIGHER LOWER power than an independent samples t-test

12. Before we move on to another example, let's try another couple of analyses on this data comparing narrow-angle to wide-angle running.

    Remember that with a matched-pairs design, we only have 22 independent observations. In questions #9-10, we conducted a dependent-samples t-test. We could have also chosen to use randomization-based methods, a sign test, or a signed-ranks test.

    To use the randomization-based method, we once again assume our null hypothesis is true. If it's true, then the time it took to run to second base was not at all affected by the running method.

    If the base running strategy doesn't make a difference, then the two times for each runner were going to be the same regardless of which strategy was being used. Any difference in time was just by chance, perhaps which one they ran first. So it doesn't matter which one we call the "wide angle" time and which we call the "narrow angle" time – the two times are completely interchangeable. In other words, we can consider the pair of measurements for each subject to be "swappable."

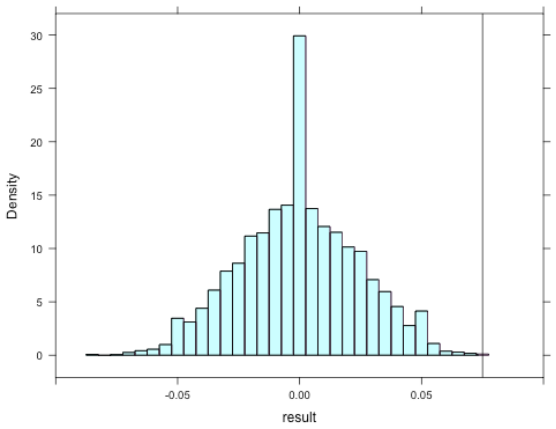
    To simulate this, we'll use the same two times for each runner but we'll randomly assign which time is associated with each method. For example, we could get the following random assignments:

Player	Narrow	Wide	Difference
1	5.50	5.55	-0.05
2	5.70	5.75	-0.05
3	5.60	5.50	0.10
4	5.50	5.40	0.10
5	5.85	5.70	0.15

Player	Narrow	Wide	Difference
1	5.50	5.55	-0.05
2	5.75	5.70	0.05
3	5.50	5.60	-0.10
4	5.50	5.40	0.10
5	5.85	5.70	0.15

Player	Narrow	Wide	Difference
1	5.50	5.55	-0.05
2	5.75	5.70	0.05
3	5.60	5.50	0.10
4	5.40	5.50	-0.10
5	5.70	5.85	-0.15

For each of these randomizations, we can then calculate the average difference between the narrow and wide methods. Using a computer, I simulated 10,000 randomizations and created a histogram of the mean differences:



The distribution is centered near zero, which is what we should expect if the running method doesn't matter.

Our actual observed difference was 0.075 (as indicated by the vertical line in the graph). The p-value was estimated to be p = 0.0004.

This p-value is similar to what we obtained from the dependent-samples t-test. Which method do you prefer? Identify advantages and disadvantages of each method.

Data: http://bradthiessen.com/html5/stats/m300/bball2.txt
Applet: http://www.rossmanchance.com/applets/MatchedPairs/MatchedPairs.htm

13. We could also choose to conduct a sign test on this data. We conducted sign-tests way back in activity #7 (blindfolded homing pigeons). In a sign test, we convert all our differences to either + or – (as displayed to the right).

    Under a null hypothesis, what proportion of our observed differences should be plusses or minuses?

    P(observing a +) = _____

    Using this value, calculate the likelihood of observing data as or more extreme than what we observed. In other words, calculate:

    P(X ≥ 17 | binomial, p=0.50, n=22) = _____

    Identify advantages and disadvantages of the sign test (compared to the dependent samples t-test and the randomization-based method).

Player	Narrow	Wide	Differences
1	5.50	5.55	–
2	5.70	5.75	–
3	5.60	5.50	+
4	5.50	5.40	+
5	5.85	5.70	+
6	5.55	5.60	–
7	5.40	5.35	+
8	5.50	5.35	+
9	5.15	5.00	+
10	5.80	5.70	+
11	5.20	5.10	+
12	5.55	5.45	+
13	5.35	5.45	–
14	5.00	4.95	+
15	5.50	5.40	+
16	5.55	5.50	+
17	5.55	5.35	+
18	5.50	5.55	–
19	5.45	5.25	+
20	5.60	5.40	+
21	5.65	5.55	+
22	6.30	6.25	+
Avg	5.5341	5.4591	17 +
Std. Dev	0.25976	0.27283	5 –

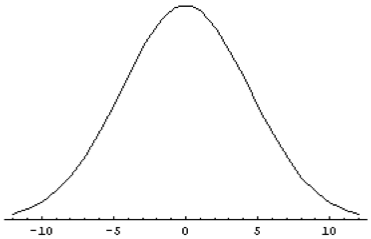
14. You hypothesize that men tend to marry women younger women. To test this hypothesis, you somehow collect the ages of a random sample of 15 couples (displayed to the right).

    Below, I've pasted the results from an <u>independent</u>-samples t-test:

INDEPENDENT SAMPLES T-TEST (using a 0.05 level of significance)

Hypotheses: $H_0: \mu_H - \mu_W = 0$
$H_1: \mu_H - \mu_W > 0$

Sampling distribution will be centered at zero with a standard error of: $\sqrt{\frac{1}{15}+\frac{1}{15}}\sqrt{\frac{14(13.6)^2+14(10.7)^2}{(14)+(14)}} = 4.47$



Our critical value, from the table in our textbook, is $t_{28,.05} = 1.701$

We can convert this to the $\overline{X}$-scale: $(1.701)(4.47) + 0 = 7.603$

Our observed value is: $\overline{X}_H - \overline{X}_W = 34.47 - 31.93 = 2.54$

We can convert this to the t-scale: $t_{28} = \frac{2.54 - 0}{4.47} = 0.57$

Our decision is to <u>retain the null hypothesis</u>. The p-value is calculated to be approximately 0.30

	Husband	Wife	Diff
1	25	22	3
2	25	32	-7
3	51	50	1
4	25	25	0
5	38	33	5
6	30	27	3
7	60	45	15
8	54	47	7
9	31	30	1
10	54	44	10
11	23	23	0
12	34	39	-5
13	25	24	1
14	23	22	1
15	19	16	3
Avg	34.467	31.933	2.533
SD	13.611	10.660	5.397
			(13 +, 2 –)

Conclusions?

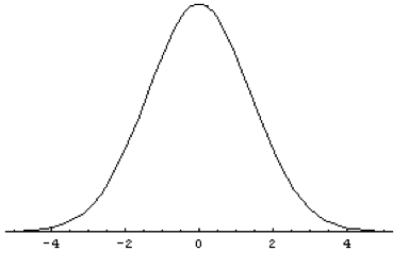
15. The independence assumption is obviously violated. Below, I've pasted the results from a dependent samples t-test. What conclusions could you make from this analysis? Why does the dependent-samples t-test yield a much lower p-value than the independent samples t-test?

DEPENDENT SAMPLES T-TEST (using a 0.05 level of significance)

Hypotheses: $H_0 : \mu_D = 0$
$H_1 : \mu_D > 0$

Statistics calculated from the difference scores: $\overline{X}_d = \overline{d} = 2.53$ and $s_d = 5.4$

Sampling distribution will be centered at zero with a standard error of: $SE = \frac{SD}{\sqrt{n}} = \frac{5.4}{\sqrt{15}} = 1.4$



Our critical value, from the table in our textbook, is $t_{14,.05} = 1.761$

We can convert this to the $\overline{X}$-scale: $(1.761)(1.4) + 0 = 2.465$

Our observed value is: $\overline{X}_D = 2.53$

We can convert this to the t-scale: $t_{14} = \frac{2.53 - 0}{1.4} = 1.807$

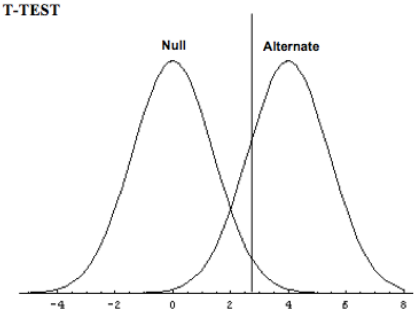
Our decision is to reject the null hypothesis. The p-value is calculated to be approximately 0.0455

16. To the right, I've shown how to estimate the power of this study (assuming the true difference in ages is 4).

    The dependent-samples t-test is obviously more powerful than the independent-samples t-test. Why? Will this always be the case?

Power Calculations (assuming the true difference between husband and wife ages is 4 years)

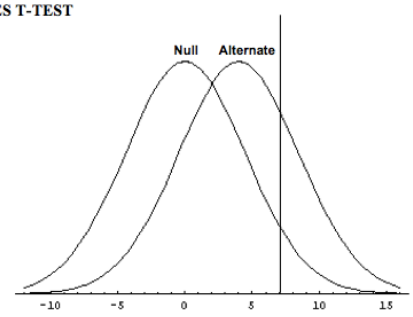
DEPENDENT SAMPLES T-TEST



Power: $P(\overline{X}_d > 2.465) = P(t_{14} > \frac{2.465 - 4}{1.4}) = P(t_{14} > -1.1) \approx 0.85$

17. Conduct a sign test and report/interpret the p-value.

INDEPENDENT SAMPLES T-TEST



18. Conduct a randomization-based test using
    www.rossmanchance.com/applets/MatchedPairs/MatchedPairs.htm
    http://bradthiessen.com/html5/stats/m300/marriage.txt

Power: $P(\overline{X}_H - \overline{X}_W > 7.603) = P(t_{28} > \frac{7.603 - 4}{4.41}) = P(t_{28} > 0.817) \approx 0.21$

19. Suppose we conduct an independent samples t-test and find an extremely small p-value (e.g., p = 0.0000001). What does that tell us about the magnitude of the difference between the two group means?

20. Instead of worrying about rejecting or failing to reject a null hypothesis (a hypothesis we already know cannot be true), we may be interested in estimating the *effect size* (the magnitude of the difference between the groups).

    A general form of an effect size for the magnitude of difference between two group means is: $\text{E.S.} = \frac{\mu_1 - \mu_2}{\sigma}$, where σ represents a standard deviation based on one or two of the groups.

    Let's calculate and interpret effect sizes for a few analyses we've already conducted:

    Activity #18: Do doctors spend less time with obese patients?
    - Obese: n = 38 mean = 24.737 std. dev = 9.653
    - Non-obese: n = 33 mean = 31.364 std. dev = 9.864
    - Just from looking at these, we could guess the pooled standard deviation to be around 9.75.

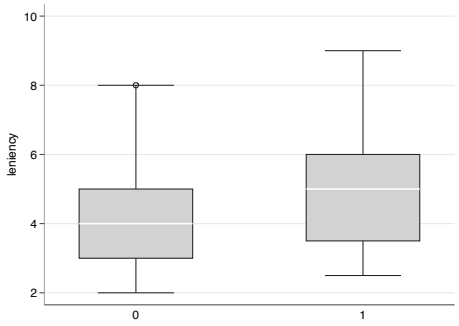
    Effect size = _____

    Activity #18: Does smiling lead to more leniency in sentencing?
    - Smile: n = 102 mean = 5.064 std. dev = 1.659
    - No-smile: n = 34 mean = 4.118 std. dev = 1.523
    - Just from looking at these, we could guess the pooled standard deviation to be around 1.6.

    Effect size = _____

21. Let's continue with the "does smiling lead to more leniency in sentencing" scenario. Recall our data was:

smile	N	mean	sd
No smile	34	4.117647	1.52285
smile	102	5.063725	1.658568
Total	136	4.827206	1.671525



When we conducted a one-tailed independent samples t-test, we obtained a p-value of p=0.0019.

Recall that the p-value represents the likelihood of obtaining data as or more extreme than what was obtained assuming the null hypothesis is true. With an equal variances assumption, we could write a p-value as:

p-value = P(observing our data | μ1 = μ2, σ1 = σ2)

Written this way, one could argue that the p-value isn't what we really want. We know we observed our data, P(observing our data = 1.0), and we also know our null hypothesis cannot be true (the means and variances cannot be exactly equal).

Don't we really want to know P(μ1, μ2, σ1, σ2 | our data)? In other words, given we have our data, we'd like to know the most probable values of our group means and standard deviations. One way to do this is with Bayesian estimation:

Posterior likelihood x prior

$$P(\mu_1,\sigma_1,\mu_2,\sigma_2,\nu \mid \text{data}) = \frac{P(\text{data} \cap \mu_1,\sigma_1,\mu_2,\sigma_2,\nu)}{P(\text{data})} = \frac{P(\text{data} \mid \mu_1,\sigma_1,\mu_2,\sigma_2,\nu)P(\mu_1,\sigma_1,\mu_2,\sigma_2,\nu)}{P(\text{data})}$$

evidence

The posterior is what we're interested in obtaining – the likelihood of a combination of population parameters given our data.

The evidence is simply the data we observed in our study.

The prior distribution represents the information we have about our population parameters. It expresses our uncertainty about the population parameters.

The likelihood represents how likely we were to observe our data given our prior distribution.

The top of the next page displays the diagram we will use to estimate our posterior distribution.

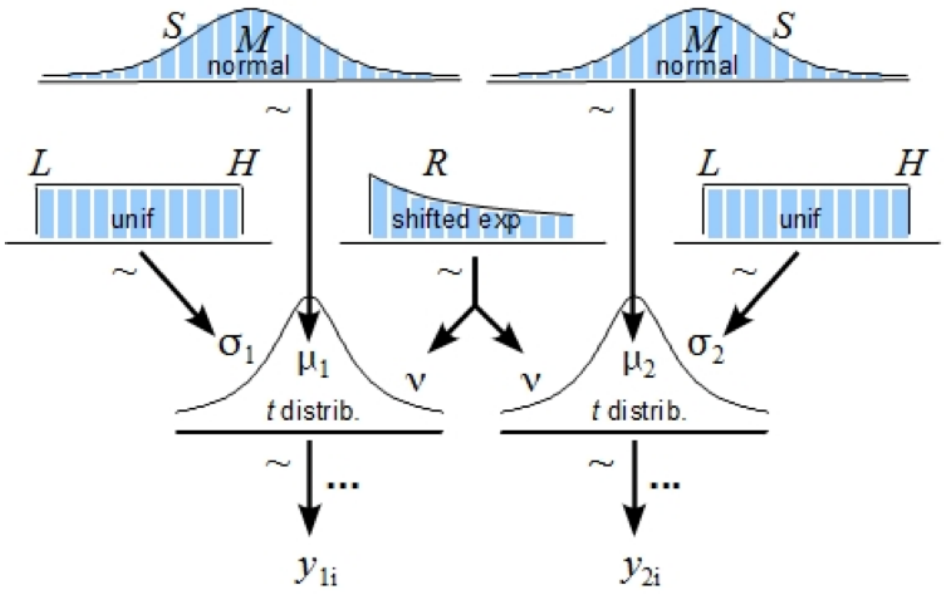


Figure 1: *Hierarchical diagram of the descriptive model for robust Bayesian estimation.*

The bottom shows the observed data from our two groups.
Each observation comes from y_1 (the smile group) or y_2 (the no-smile group).

We're assuming each independent observation comes from a t-distribution (indicating that we believe the distribution is unimodal and symmetrical, but there is a chance to have some outliers).

Those t-distributions are defined by their means, standard deviations, and degrees of freedom (denoted as v).

We model the degrees of freedom, which would be a single value shared by both t-distributions, as coming from an exponential distribution (with R = the rate parameter).

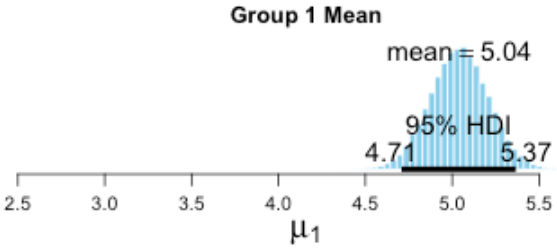
The standard deviations of each t-distribution are modeled by uniform distributions, implying that we do not have reason to favor one hypothesized set of values over any other.

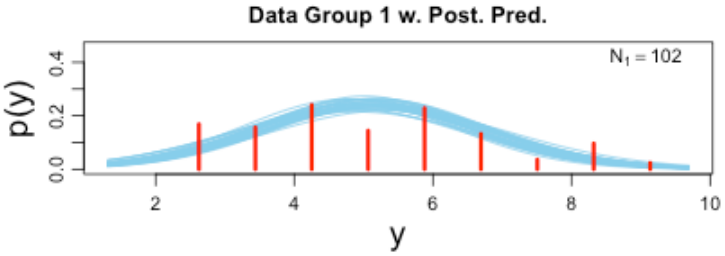
Finally, the means of each t-distribution (the parameters that we are most interested in estimating) are modeled as coming from wide normal distributions (with standard deviations equal to 1000 times the standard deviation of the pooled data). This favors values near zero, but allows a broad range of plausible values.

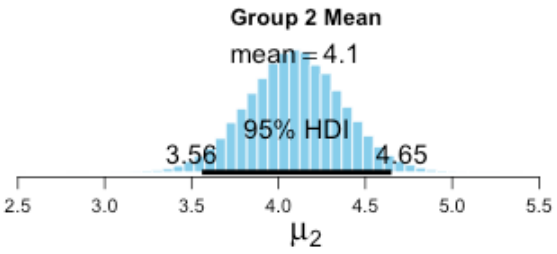
With this model, we can estimate each parameter. Interpret the following:

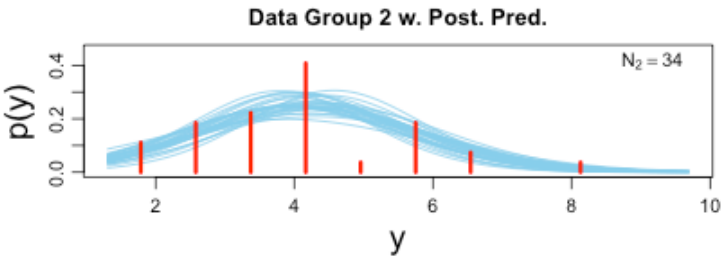
	mean	sd	median	HDIlo	HDIup	Rhat	n.eff
mu1	5.041	0.1680	5.041	4.708	5.366	1	59960
mu2	4.097	0.2770	4.096	3.559	4.651	1	60127
muDiff	0.945		0.945	0.319	1.589		
nu	48.167	32.7229	39.909	5.609	112.683	1	23152
sigma1	1.643	0.1241	1.637	1.408	1.891	1	54815
sigma2	1.549	0.2082	1.529	1.166	1.967	1	48751
sigmaDif	.095		0.107	-.398	0.548		
effSize	0.594		0.594	0.196	1.006		

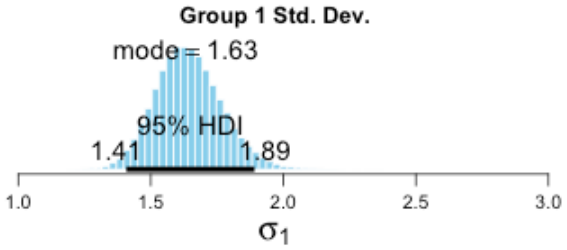
The next page shows posterior distributions for these parameters:

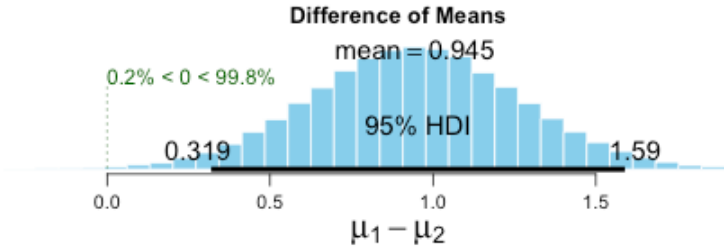

Group 1 Mean
mean = 5.04
95% HDI
4.71
5.37
2.5 3.0 3.5 4.0 4.5 5.0 5.5
μ1

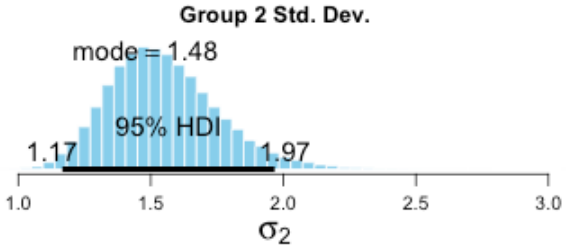

Data Group 1 w. Post. Pred.
N1 = 102
p(y)
0.0 0.2 0.4
2 4 6 8 10
y

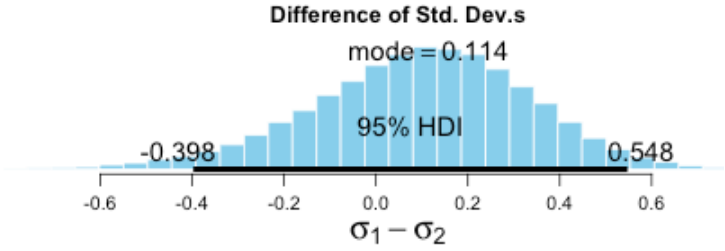

Group 2 Mean
mean = 4.1
95% HDI
3.56
4.65
2.5 3.0 3.5 4.0 4.5 5.0 5.5
μ2

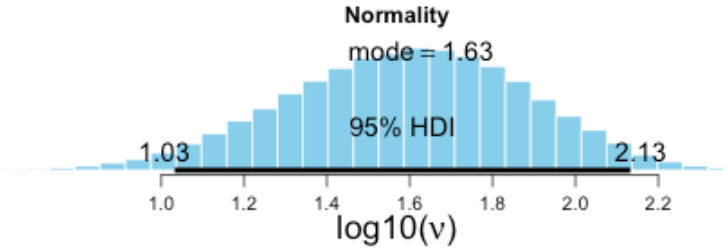

Data Group 2 w. Post. Pred.
N2 = 34
p(y)
0.0 0.2 0.4
2 4 6 8 10
y

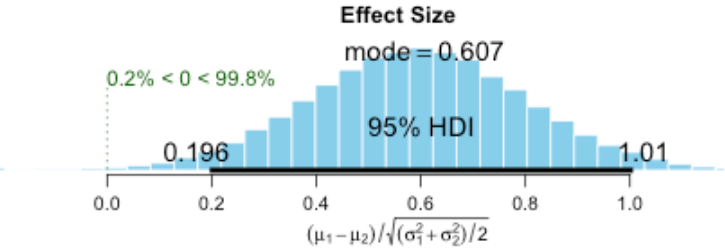

Group 1 Std. Dev.
mode = 1.63
95% HDI
1.41
1.89
1.0 1.5 2.0 2.5 3.0
σ1

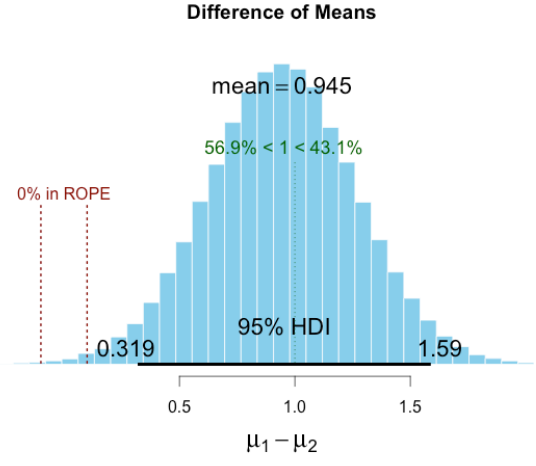

Difference of Means
mean = 0.945
0.2% < 0 < 99.8%
95% HDI
0.319
1.59
0.0 0.5 1.0 1.5
μ1 − μ2


Group 2 Std. Dev.
mode = 1.48
95% HDI
1.17
1.97
1.0 1.5 2.0 2.5 3.0
σ2


Difference of Std. Dev.s
mode = 0.114
95% HDI
-0.398
0.548
-0.6 -0.4 -0.2 0.0 0.2 0.4 0.6
σ1 − σ2


Normality
mode = 1.63
95% HDI
1.03
2.13
1.0 1.2 1.4 1.6 1.8 2.0 2.2
log10(ν)


Effect Size
mode = 0.607
0.2% < 0 < 99.8%
95% HDI
0.196
1.01
0.0 0.2 0.4 0.6 0.8 1.0
(μ1 − μ2)/√((σ1² + σ2²)/2)


Difference of Means
mean = 0.945
56.9% < 1 < 43.1%
0% in ROPE
95% HDI
0.319
1.59
0.5 1.0 1.5
μ1 − μ2

22. How much of an effect does smiling have on leniency of sentencing? Support your answer with something from the output.

23. Does an equal variances assumption seem reasonable for the data in this study? Support your answer.

24. Because we are dealing with a Bayesian posterior probability distribution, we can extract much more information than we can from an independent samples t-test:

    a) We can estimate the probability that the true difference in means is above (or below) any comparison value. In this case, we may be interested in the probability that the difference in means is at least 1.0.

    b) We can estimate the probability that the true difference in means is zero. On the bottom of the previous page, you can see the probability that the difference is near zero (in the ROPE - region of practical equivalence) is extremely small.

25. Enter your own data to analyze via Bayesian estimation: http://sumsar.net/best_online/

26. R code to replicate everything in this activity is available at: http://bradthiessen.com/html5/stats/m300/act19.R