Unit 3: Randomization Methods for Correlations (download the data at http://web.me.com/bradthiessen/data/twitter.csv)

Throughout MATH 300 and MATH 301, we've learned how to use randomization methods to test for statistical significance.

Randomization methods follow three steps:

- Compute the measure from the sample data (compute a mean, difference among medians, or any other measure)
- Randomize the data many times and re-compute the measure each time (generating a distribution of your measure)
- Compare the observed sample measure to the distribution of the measure generated from the randomizations
  (We reject the null hypothesis if the observed sample measure lies in the tails of the randomization distribution)

Let's apply randomization methods to analyze correlations.

Situation: On Twitter, what's the relationship between the number of people you follow, the number of people who follow you, and the number of tweets you send? To investigate this, I collected this data for the top 100 Twitter users* as of March 29, 2011: (source: http://twittercounter.com/)

@name	followers	following	tweets
ladygaga	9097060	144073	654
justinbieber	8465937	110915	8562
britneyspears	7263054	416316	671
barackobama	7198417	700748	1299
kimkardashian	6944340	129	7065
aplusk	6501379	622	6566
katyperry	6445804	69	2713
theellenshow	6289686	48784	4111
taylorswift13	5781560	54	885
oprah	5455368	30	278
shakira	5059823	38	934
twitter	4724860	456	1068
selenagomez	4584413	480	1501
twitter_es	4400375	20	399
50cent	4314987	2	3635
jtimberlake	4229893	20	473
ryanseacrest	4221722	377	4544
rihanna	4212414	313	1504
ashleytisdale	4114310	105	1540
cnnbrk	4101761	40	8845
mariahcarey	4085304	47	2114
parishilton	3683301	1735	7018
the_real_shaq	3660119	627	3575
eminem	3650073	0	118
jessicasimpson	3602015	79	693
coldplay	3588165	2466	798
pink	3570141	106	2496
iamdiddy	3522100	1189	12087
mrskutcher	3446428	205	4464
aliciakeys	3396158	163	1250
jimmyfallon	3361748	3274	3502
twitpic	3342059	28143	463
charliesheen	3293980	36	143
chelseahandler	3277143	45	2312

@name	followers	following	tweets
kaka	3229966	314	1915
perezhilton	3138750	296	32697
nytimes	3055071	435	61107
snoopdogg	2985663	1339	5381
nickiminaj	2938585	627	7625
google	2918918	361	2299
kanyewest	2880679	0	1402
huckluciano	2869924	250	4492
tyrabanks	2859143	598	1460
ubersoc	2845950	1926	8276
jimcarrey	2832556	1	2301
ddlovato	2803033	113	4648
eonline	2798585	32893	31292
lancearmstrong	2780985	221	6782
conanobrien	2771171	1	422
lilyroseallen	2760701	132	2713
khloekardashian	2750847	110	16088
soujaboy	2706201	495	26248
theonion	2704976	2	7692
jonasbrothers	2698367	2109	1172
ricky_martin	2628854	145	2262
peoplemag	2460670	606	5995
time	2432533	308	21494
ashsimpsonwentz	2411791	84	175
nba	2408046	842	16081
stephenfry	2396700	52764	8127
billgates	2392666	63	270
nickjonas	2390228	133	1425
nickcannon	2363713	504	4246
kourtneykardash	2357577	43	2012
tonyhawk	2354347	271	4419
breakingnews	2336977	360	49241
johnlegend	2269666	191	2696
rainnwilson	2252854	198	4688

@name	followers	following	tweets
themandymoore	2249060	36	696
ivetesangalo	2245944	236	17837
mashable	2241026	2231	33497
algore	2227440	9	309
petewentz	2226744	232	6847
cristiano	2207766	52	521
lennykravitz	2203660	1896	569
youtube	2195717	235	3451
noaheverett	2192578	1075	3852
marthastewart	2170806	8216	2395
chrisbrown	2169497	574	1193
106andpark	2157661	220	3274
s***mydadsays	2153407	1	138
drdrew	2142125	263	2274
stephenathome	2142015	0	1431
giulianarancic	2115177	308	3434
juanes	2106882	556	4046
joelmchale	2088173	130	1350
rustyrockets	2088018	58	1838
mchammer	2068458	37088	16629
drakkardnoir	2058271	255	384
whitehouse	2027599	120	2298
nfl	2006069	135	4303
sarabareilles	1999575	65	803
schwarzenegger	1987976	109843	1333
serenawilliams	1986802	107	8077
alyankovic	1984377	197	1163
nellyfurtado	1978102	241	1571
revrunwisdom	1975057	0	8512
denise_richars	1966927	141	6072
funnyordie	1966074	2950	4210
instyle	1941157	2078	12085
robb_fisher	167	99	411
Thiessen	35	60	105

* Ok, so @Thiessen and @Robb_Fisher aren't in the Top 100. You should follow us anyway.

1) Let's start by examining the correlations among our variables. Using the Pearson's correlations listed below, explain the nature of the relationship between each variable. How much of the variance in number of followers is explained by the number of people you follow?

	followers	following	tweets
followers	1.0000		
following	0.4663	1.0000	
tweets	-0.1021	-0.0658	1.0000

2) Those numbers represent the observed correlations -- the degree to which the relationship between our variables is linear *for our sample.* We see, for example, that the correlation between the number of followers and the number of people followed is 0.4663. Is this correlation statistically significant? If, for the population of twitterers, there is no relationship between the number of followers and number of people being followed, is it *possible* that we could observe a correlation as high as 0.4663 just by random chance with our sample? The answer to this, of course, is *yes.* The real question of interest is, though, *how likely were we to observe a correlation as high as 0.4663 if, in fact, these variables have no relationship?*

To investigate this question, we can conduct a hypothesis test. As we'll learn in this unit, we can conduct a t-test or F-test to determine if an observed correlation is significantly different from zero. I had Stata conduct this test and display the p-values. Assuming the null hypothesis is that the two variables have no correlation, interpret these p-values.

Correlation between followers and following = 0.4663; p-value < 0.0001
Correlation between followers and tweets = -0.1021; p-value = 0.3096
Correlation between following and tweets = -0.0658; p-value = 0.5136

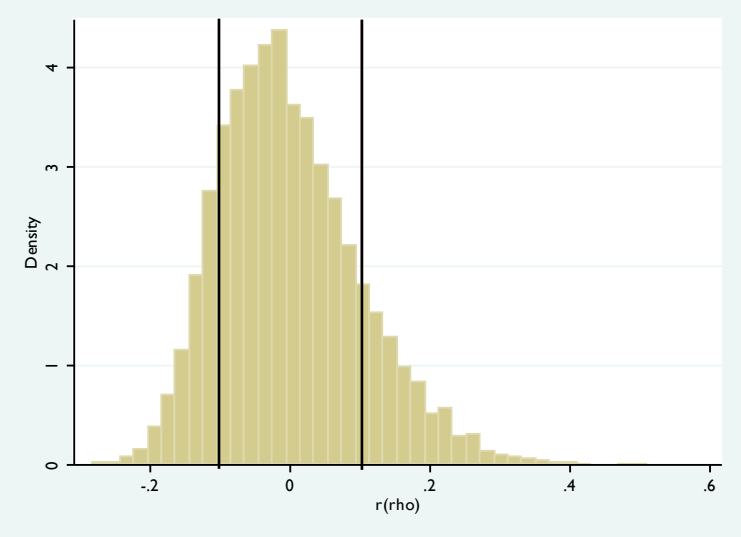
3) Rather than conducting a t-test (or F-test) that have certain parametric assumptions, we could investigate the significance of these correlations through *randomization methods*. As an example, let's focus on the correlation between the number of followers and the number of tweets (r=-0.1021). If our null hypothesis is true and there is no relationship between these variables, then it doesn't matter which number of followers corresponds to which number of tweets. Therefore, we can randomly scramble the number of tweets and calculate the correlation of this randomized dataset. We can repeat this and continue to calculate correlations. The following table shows how this process works:

Original Data			1st randomization (tweets are scrambled)			2nd randomization (tweets are scrambled)		
@name	followers	tweets	@name	followers	tweets	@name	followers	tweets
ladygaga	9097060	654	ladygaga	9097060	671	ladygaga	9097060	654
justinbieber	8465937	8562	justinbieber	8465937	12085	justinbieber	8465937	105
britneyspears	7263054	671	britneyspears	7263054	1299	britneyspears	7263054	8562
barackobama	7198417	1299	barackobama	7198417	411	barackobama	7198417	12085
...	...	...	...	...	...	...	...	...
instyle	1941157	12085	instyle	1941157	654	instyle	1941157	411
robb_fisher	167	411	robb_fisher	167	105	robb_fisher	167	671
Thiessen	35	105	Thiessen	35	8562	Thiessen	35	1299
Correlation = -0.2143			Correlation = 0.0512			Correlation = 0.3163		

(correlations are calculated from the entire dataset (N = 102) after tweets have been scrambled)

4) We repeat this randomization process a large number of times to obtain a distribution of possible correlations (under the assumption that the variables have no relationship). This gives us an idea of what correlations we could expect from this dataset.

I had Stata run 10,000 randomizations and calculate a correlation each time. The following histogram shows the 10,000 correlations obtained. A vertical line has been drawn corresponding to our observed correlation of -0.1021. I've also drawn a line corresponding to +0.1021, since we're interested in the magnitude of the correlation; not just its sign.



Look at where our observed correlation would be located in this distribution. Remember, this distribution represents 10,000 correlations we could have gotten if our variables had no relationship. Do you think it was *likely* or *unlikely* that our observed correlation of -0.1021 could have come from this distribution? Does that mean the correlation was statistically significant?

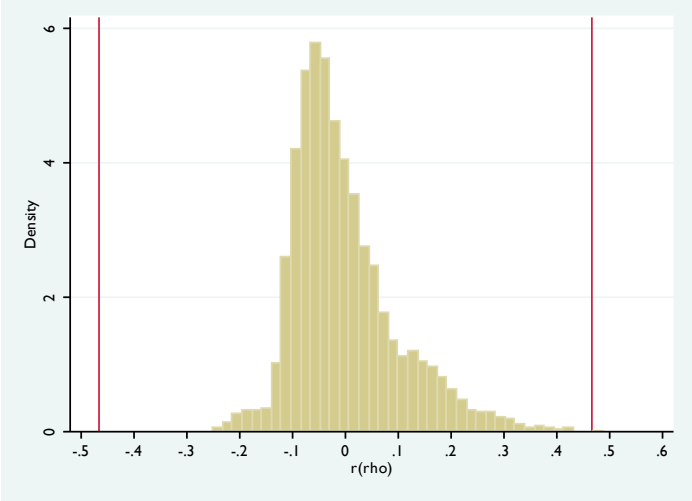
5) Remember that a p-value, in this situation, represents the probability of observing a correlation as or more extreme than +/- 0.1021. Once we generate the above histogram, it's easy to calculate the p-value. We simply need to count how many of our randomized correlations were more extreme than 0.1021.

Here's the syntax and output from Stata. T(obs) represents our observed correlation; c represents the number of randomizations with correlations more extreme than T(obs); n represents the number of randomizations generated. Based on the output, what is our p-value? What do we conclude about the correlation between number of followers and number of tweets?

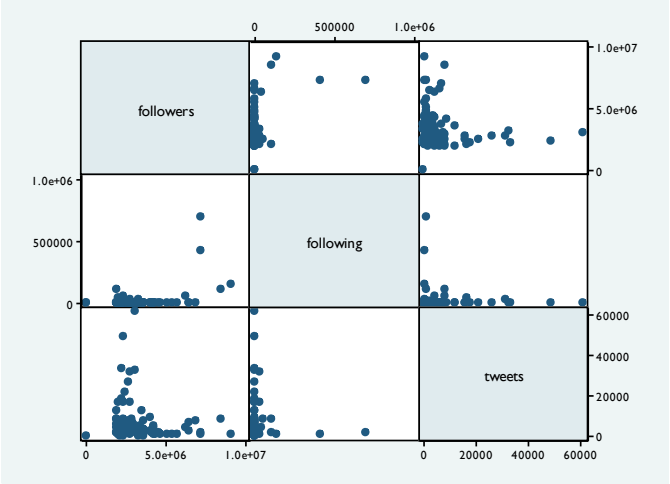
```
. permute tweets rho=r(rho),reps(10000) nodots nowarn: correlate followers tweets

Monte Carlo permutation results                 Number of obs    =       102
------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rho |  -.1021066    3004   10000  0.3004  0.0046   .2914241   .3094923
------------------------------------------------------------------------------
Note:  confidence interval is with respect to p=c/n.
Note:  c = #{|T| >= |T(obs)|}
```

6) Let's quickly conduct one more example. The observed correlation between number of followers and the number of people followed was 0.4663. Let's generate 10,000 randomizations, display the histogram, and examine the Stata output. What conclusions can you make?

```
. permute following rho=r(rho),reps(10000) nodots nowarn: correlate followers following

Monte Carlo permutation results                 Number of obs    =       102
------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rho |    .460522       2   10000  0.0002  0.0001   .0000242   .0007223
------------------------------------------------------------------------------
Note:  confidence interval is with respect to p=c/n.
Note:  c = #{|T| >= |T(obs)|}
```



7) Now that we think we understand the relationship among our variables, let's look at some scatterplots.



These scatterplots should give you reason to be concerned. Relationships between our variables aren't clear -- we have too many extreme outliers (the number of followers, for example, ranges from 35 to 9,097,060 with a mean of 3,146,499). We could transform our data by taking logarithms -- that would make the number of followers range from ln(35) = 3.55 to ln (9097060) = 16.02 with a mean of 14.70. We could also choose another type of correlation.

8) Let's use Spearman's rho. Recall that Spearman's rho is equivalent to the correlation between our variables if we convert all the data to ranks. When I have Stata calculate Spearman's rho on this dataset, I get the following output:

Correlation between number of followers and number followed

```
Number of obs =         102
Spearman's rho =        0.0673
Test of Ho: followers and following are independent
    Prob > |t| =        0.5017
```

Correlation between number of followers and number of tweets

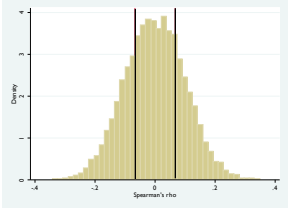
```
Number of obs =         102
Spearman's rho =       -0.0435
Test of Ho: followers and tweets are independent
    Prob > |t| =        0.6642
```

How do these correlations compare to the Pearson's correlations we calculated earlier? Which type of correlation is more appropriate for this dataset?

9) Looking at the above Stata output, you can see some p-values (and a null hypothesis being tested) under each correlation. Interpret these p-values. While there are specific procedures you can use to test these hypotheses and estimate these p-values, we can also use our randomization methods. Look at the histograms and Stata output and briefly explain if the randomization methods support the p-values listed above.

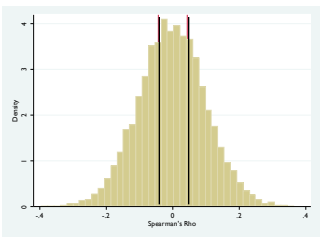
Correlation between number of followers and number followed

```
. permute following rho=r(rho),reps(10000) nodots nowarn: spearman following followers

------------------------------------------------------------------------------
T            |     T(obs)       c        n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rho |   .0672749    5066    10000  0.5066  0.0050   .4967503    .5164458
------------------------------------------------------------------------------
```



Correlation between number of followers and number of tweets

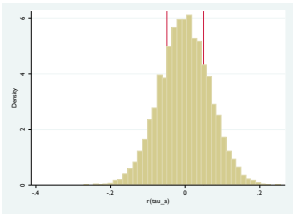
```
. permute tweets rho=r(rho),reps(10000) nodots nowarn: spearman tweets followers

------------------------------------------------------------------------------
T            |     T(obs)       c        n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rho |  -.0434944    6640    10000  0.6640  0.0047   .6546456    .6732588
------------------------------------------------------------------------------
```



9) Let's repeat this one last time using Kendall's tau. Look at the observed correlations, the p-values obtained from both traditional and randomization methods, and the randomization distributions. What conclusions can you draw?

Correlation between number of followers and number followed

```
. ktau followers following, stats(taua)

  Number of obs =     102
Kendall's tau-a =       0.0491
Kendall's tau-b =       0.0492
Kendall's score =     253
    SE of score =     345.833   (corrected for ties)

Test of Ho: followers and following are independent
     Prob > |z| =       0.4662  (continuity corrected)


. permute following taua=r(tau_a),reps(10000) nodots: ktau following followers, stats(taua)

-------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+-----------------------------------------------------------------
        taua |   .0491167    4602   10000  0.4602  0.0050   .4503941   .4700291
-------------------------------------------------------------------------------
Note:  confidence interval is with respect to p=c/n.
Note:  c = #{|T| >= |T(obs)|}
```



Correlation between number of followers and number of tweets

```
. ktau followers tweets, stats(taua)

  Number of obs =     101
Kendall's tau-a =      -0.0493
Kendall's score =    -249
    SE of score =     340.806   (corrected for ties)

Test of Ho: followers and tweets are independent
     Prob > |z| =       0.4668  (continuity corrected)


. permute tweets taua=r(tau_a),reps(10000) nodots nowarn: ktau tweets followers, stats(taua)

-------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+-----------------------------------------------------------------
        taua |  -.0493069    4736   10000  0.4736  0.0050   .4637727   .4834426
-------------------------------------------------------------------------------
Note:  confidence interval is with respect to p=c/n.
Note:  c = #{|T| >= |T(obs)|}
```