* **To do before TVW meeting**
  + Check theory
  + In Example, say that OLS coefficient is continuous function because matrix inverse is continuous
  + Compare my proofs to standard bootstrapping proofs
* **To read about multiple testing**
  + Fisher’s combined probability test
    - https://en.wikipedia.org/wiki/Fisher%27s\_method
* **ECDF approach**
  + Is ECDF valid when the variables are correlated?
    - I think is the same “case-by-case weak convergence” that one of the online things was talking about.
    - Maybe use WLLN for correlated observations?
  + If we bootstrap the ECDF under the null, how do we average them together, and how to we get quantiles?
    - Idea
      * Take a vector of quantiles; e.g. seq( 0, 1, 0.05 )
      * For average ECDF, just take columnise average of sample values at each of these chosen quantiles
      * For lower CI bound, take the entire sample vector of values such that *every one* of the quantiles is above its respective 2.5th percentile
* \* = should probably read in more detail
* **~~Fox chapter on bootstrapping regressions~~**
  + ~~In section on bootstrap hypothesis tests, uses this residual resampling algorithm:~~
    - ~~Set Y\*\_i = Yhat\_i + sample(Y - Yhat)~~
    - ~~Test H0: beta = beta.hat (rather than beta = 0)~~
* **~~Bickel~~**
  + ~~Classic paper that goes over basic asymptotics of bootstrap~~
  + ~~Uses SLLN~~
* **~~Chernick textbook~~**
  + ~~Great table comparing different CI methods with their assumptions (Ch 3, page 8)~~
* **~~Hall & Wilson “Two guidelines”~~**
  + ~~The point of using asymptotically pivotal statistics (centered and scaled) is that then the asymptotic distribution of the statistic does not depend on any unknowns.~~
  + ~~When you can’t estimate variance of estimator in order to create pivot, suggests using accelerated bias correction bootstrap~~
* **~~Bretz, Hothorn, & Westfall (2016) R paper~~**
  + ~~They are fixing the Y and permuting the Xs (in their application, these are just the group labels), so are keeping the correlation structure of the Ys as we are (pg. 130)~~
  + ~~Bootstrapping vs. permutation pros and cons (pg. 140)~~
* **~~Westfall & Young (1993) book~~**
  + ~~Page 39-41: Analyzes P( P\* < alpha ) in bootstrap iterates~~
  + ~~Page 92-93: Re-randomization~~
  + ~~106-109: OLS; example of parametric resampling under the null~~
  + ~~210-216~~
  + ~~Bootstraps of pivot statistics converge faster (Section 2.2.2)~~
  + ~~Pg. 133: How to resample for OLS regression using residuals~~
    - ~~Set Y\*\_i = sample(Y - Yhat)~~
    - ~~Test H0: beta = 0~~
* **~~Westfall & Troendle (2008)~~**
  + ~~Shows that “permutation methods are distribution-free under an appropriate exchangeability assumption and FWER follows mathematically, regardless of sample size”~~
  + ~~Setting: You have multiple outcomes measured on multiple categorical groups and are interested in full exchangeability on the outcome across the groups (not just different means, for example).~~
  + ~~Implemented in SAS’ PROC MULTTEST~~
  + ~~According to SAS documentation, shows that: “when subset pivotality holds, the joint distribution of p-values under the subset is identical to that under the complete null”~~
* **~~\*Troendle (2004) “Slow convergence”~~**
  + ~~Talks about multiple 2-sample t-tests~~
  + ~~2 approaches:~~
    - ~~Permute a subject’s outcome variables while assigning group labels (with~~ *~~uncentered~~* ~~variables)~~
    - ~~Bootstrap by resampling a subject’s outcome variables while assigning group labels (with~~ *~~centered~~* ~~variables)~~
  + ~~Page 3 gives very clear overview of the actual permutation and bootstrapping methods~~
  + ~~In general, Westfall and Troendle are interested in categorical-X problems, so they resample by centered observations within each group in order to fix the sample sizes~~
* **~~Notes on resampling-based MT literature~~**
* ~~Basic bootstrapping theory~~
* ~~https://stats.stackexchange.com/questions/9664/what-are-examples-where-a-naive-bootstrap-fails/9722#9722~~
* ~~https://stats.stackexchange.com/questions/11210/assumptions-regarding-bootstrap-estimates-of-uncertainty?noredirect=1&lq=1~~
* ~~Assumptions~~
* ~~https://stats.stackexchange.com/questions/11210/assumptions-regarding-bootstrap-estimates-of-uncertainty~~
* ~~Troendle (2004) and Westfall (2010) show that “the bootstrap often does not control the FWER adequately”~~
* ~~Application of bootstrap to arbitrary smooth function:~~
* ~~https://stats.stackexchange.com/questions/246632/smoothness-of-a-statistic-for-bootstrapping~~

**Open questions**

**To read:**

**https://stats.stackexchange.com/questions/91631/why-does-my-bootstrap-interval-have-terrible-coverage**

**http://onlinelibrary.wiley.com/doi/10.1002/cjs.5550340103/abstract**

**Q:** If creating a bootstrap CI instead of hypothesis test, should we still use pivotal thing?

* **Q:** Why is it more accurate to bootstrap a pivot (e.g., studentized bootstrap) than a non-pivot (e.g. statistic itself)?
  + Read this: Hall ( 1986a , 1988b ) wrote two key papers which demonstrate the value
  + of asymptotically pivotal quantities in the accuracy of bootstrap confi dence
  + intervals.
* **Q:** Do we need to worry about making N pivotal or is it enough for the test stat itself to be pivotal?

**Q:** N-hat is highly right-skewed and is definitely not asymptotically normal since k is fixed. Are there bootstrap methods that help accommodate this?

**Q:** Does centering and scaling a statistic automatically make it pivotal regardless of whether it’s asymptotically normal?

* + See Hall & Wilson pg 759 (highlighted)

**Q:** What are the differences in assumptions between bootstrapping and permutation? Should we consider permutation?

* **Q:** Will ours need to assume no logical constraints among the hypotheses?
* **Q:** Instead of bootstrapping, can we use nonparametric delta method to derive the SE of the number of rejections? Maybe using the empirical influence function? (See Davison & Hinkley, 2.7.2)
* **~~Resolved questions~~**
* **~~Q:~~** ~~Why does Blakesley/Westfall’s minP bootstrapping work since it is an extreme order statistic?~~
* **~~A:~~** ~~Probably because minP is not an order statistic of the data themselves.~~
* **~~Q:~~** ~~If we go the confidence interval route, why not just treat the p-values as our data and resample directly from them to produce CI?~~
* **~~A:~~** ~~Because there are only a fixed number of p-values (k), and this is not necessarily asymptotically large.~~
* **~~Q:~~** ~~Why does bootstrapping work with small samples since we need the ECDF in sample to go to the true CDF? Or does it not work in small samples?~~
* **~~A:~~** ~~Correct. It does not necessarily work in small samples.~~
* **~~Highlighted = ones I haven’t looked at in detail yet~~**
* **~~Summary of previous two~~**
  + ~~Plot1 (CIs)~~
    - ~~Shows that the residual resampling is working correctly (since results are same regardless of whether we generate under H0 or under HA)~~
    - ~~As expected, more variable when Y’s are more correlated~~
* **~~2017-8-1 (generate under alternative)~~**
  + ~~Same as 2017-7-25, but now generating under weak alternative (rho.XY = 0.03)~~
  + ~~Expect same exact results for CIs (since they only use results from resampling under H0), but higher rejection rates~~
* **~~2017-7-25 (generated under null)~~**
  + ~~CIs are 95% two-sided ones for both values of alpha~~
  + ~~Hypothesis tests are one-sided and have sample alpha as the individual tests~~
  + ~~N=1,000 and B=2,000 (j=10 per simulation)~~
  + ~~Took maybe 10 hours total~~
  + **~~Figure out why file 125 has extra rows (try running it again)~~**
  + **~~The apparently too-conservative behavior for the uncorrelated case is an artifact of N-hat’s being discrete.~~** 
    - ~~It occurs because when the Y’s are uncorrelated, the distribution of the number of rejections is less right-skewed, so it drops off more quickly in the tails. Since N-hat is discrete, we are forced to pick a quantile that doesn’t have exactly 5% of the mass above it. By default, R inverts the ECDF, so uses the quantile with LESS than 5% above it. Because the distribution of N-hat drops off quickly around the chosen quantile, we end up with conservative performance (see artifact\_plot1 and artifact\_plot2, where I shade in red the proportion that are above the variable’s own critical value).~~
    - ~~Artifact might improve a little with more simulation reps, but only to a point.~~
    - ~~For this case, we can benchmark against the truth, since N-hat is binomial. Indeed, the empirical quantiles of both n.rej and the bootstrap estimates are exactly correct. ☺~~
* **~~Summary so far~~**
  + ~~Seems like N=1,000 and B=2,000 is enough for good asymptotics~~
  + ~~When data are generated under null, resampling the Y’s or the residuals seems to works. But former theoretically should work only in more limited cases.~~
* **~~2017-7-24 “Sherlock sim”~~**
  + ~~Huge with N=10,000, B=10,000~~
  + ~~Resampling under nulls (Y’s rather than residuals)~~
  + ~~This seemed to work based on interim results (though I had trouble with existing stitching script, so stitched results in overall\_stitched folder might be wrong)~~
* **~~2017-7-24 “Test smaller sim”~~**
  + ~~N=1,000, B=2,000~~
  + ~~Switched to resampling residuals instead of Ys themselves for the first time~~
  + **~~Seems to work!~~** ~~Going to try on cluster to have larger simulation~~
* **~~2017-7-22~~**
  + ~~N=100, B=1,000~~
  + ~~Looked good at alpha=0.05 (4.4% rejections) but maybe too conservative for alpha=0.01? (3.6% rejections)~~
  + ~~Took about 1.5 hours locally~~
  + **~~Seems too small~~**
* **~~2017-7-21~~**
  + ~~With N=5,000, B=2,000, n.sims = 250, looks good~~
  + ~~5.5% rejection rate on joint null for alpha = 0.05~~
  + ~~6% rejection rate for alpha = 0.01~~
  + ~~Took about 30 hours locally~~
  + **~~Seems good~~**
* **~~Summary so far~~**
  + ~~Resample under H0 by drawing Y’s separately; reject by inverted CI using its percentiles => WORKS~~
    - ~~Sherlock sim~~
  + ~~Resample from original with single Y1; look at variance of mean => WORKS~~
  + ~~Bootstrap from original sample with 100 independent std. normal; reject using z-test that’s only a function of mean => DOESN’T WORK~~
    - ~~The bootstrapped distribution of the absolute differences is too variable compared to true distribution~~
    - ~~Hence rejects only 1.8% of time~~
  + *~~Conjecture~~*~~: Hall & Wilson approach seems to rely on the idea that the standardized estimator is pivotal, i.e., that it comes from a location-scale family. This is why we can get away with only sampling from the original distribution. Since our estimator definitely isn’t from a location-scale distribution, maybe that’s why it does not work.~~

~~Also, based on the different results between currently running one and very first one, I think B=500 is not enough, and B = 10,000 is enough. Not sure about intermediate ones.~~

* **2017-7-16**
  + Local simulations to look at Hall & Wilson performance (sandbox)
  + “hall\_wilson\_1.csv”: Basic test of H&W idea when bootstrapping OLS coefficient (not number rejections) and with data generated under null. Works (achieves nominal alpha).
  + “hall\_wilson\_2.csv”: Same as above, but data generated under Ha. Works in that the bootstrap test rejects much more often than 5%.
  + Other scenarios (not saved)
    - Generate 1 std. normal, bootstrap its mean. This WORKS with H&W approach (so problem is not that the means are too variable).
* **2017-7-10**
  + Trying 0 correlation among Ys to see if we get expected (binomial) results
* **2017-7-9**
  + Same as below, but with B=1,000
  + Did not help
  + From running simulations 1 at a time, it seems like the resampling approach of the last few simulations (resampling X’s and Y’s together) does not adequately recover the distribution of N-hat. Although the average correlation of X with the Y’s is the same in the bootstrap sample as in the original, its SD is about 30% (?) larger. I assume this is because when you are duplicating individuals, you could accidentally get samples with many duplicates of one person that drive a spurious correlation between X and some of the Ys. Resampling as Tyler said (Y’s only while fixing X) results in bootstrap samples where the correlations of X with Y’s have the same variability (SD) as in the original. **Need to think about how this relates to the theory I already have…**
* **2017-7-8**
  + Same as below, but now returning all output (in order to help with diagnosing CIs)
  + Each sim rep took about 3 min
* **2017-7-7**
  + Each sim rep took about 3 min
* **2017-7-2**
  + Re-running 2017-6-26 simulation, but **now using Hall & Wilson bootstrap** for the hypothesis test to see if it rejects closer to 5% of the time
  + Only running 1 point on the graph
  + Did even worse
* **2017-7-2** 
  + Same as below, except increasing n to 50,000
  + Same performance as below, so probably n was large enough at 10,000
  + Each sim rep took ~ ½ hr
* **2017-6-26**
  + B = 500 still
  + This is an exact replication of the alpha = 0.05 line in the upper left panel of the 2017-6-25 results, except now using n=10,000 instead of n=50.
  + Performance is much closer to nominal than with n=50, but still not hitting 0.05.
  + Each sim rep took ~ 7 min
* **2017-6-25**
  + Ran all scenarios with n=50, B = 500
  + The way CI width changed in response to all the scenario parameters made sense, but hypothesis tests were too conservative