Project Log: Multiple Outcomes

Contents

Back to Romano vs. Westfall

Romano's crit values: 95th percentile of biggest CENTERED test stat (N(0,1)) among the subset of hypotheses that 's being considered, where the original test stats are computed not under null (pg 1384)

Westfall's crit values: adjust p-values using the p-values under the strong null

From examples in Romano Wolf JASA paper, seems like subset pivotality basically fails when, e.g., the SE or correlation structure of some of the test stats depends on parameters used in *other* tests. For example (Romano 2003 JASA, Example 4.1), say you're testing all pairs of correlations for three variables. X and Y are truly uncorrelated, and X and Z are uncorrelated (nulls are true). But Y and Z are correlated. Then the joint distribution of the X-Y and X-Z correlations depends on the Y-Z correlation, which makes sense intuitively. So subset pivotality fails. In this case, Westfall's approach would be to resample under the strong null, so all three correlations are null and the X-Y and X-Z correlations themselves are uncorrelated. But this does not reflect all possible distributions under which the X-Y and X-Z correlations are null since their distribution depends on the Y-Z correlation. Romano's approach would be to resample under the original distribution, then center the test stats. So then we would have correlations that are all centered at 0, but the X-Y and X-Z correlations are correlated if, in reality, the Y-Z correlation is non-null.

But I think this also means that if you have test stats whose distribution depends on the parameter being

tested (not on the other parameters, so not talking about subset pivotality), like $\widehat{p}$ for a Bernoulli RV (whose SE depends on p), then Westfall's method would have the right SEs (reflecting the null), but Romano's would not. Right?

Maybe it doesn't matter. The key difference is that in first paragraph, we're saying null is true for some set of hypotheses, but not necessarily the others. In the second paragraph, the problem only arises if the null is false EVEN FOR the set we're considering, which is different.

(e.g. pairwise correlations)

Question: When does centering test stats differ from generating data under null?

Stream of consciousness

MICE thing doesn't work because we're regressing Y_2 on Y_1, and if data are generated under H_A, these are correlated because both outcomes are correlated with X. This remains true even if we leave X out of the imputation model.

FCR-center works for OLS but not logistic regression because for the latter, SE depends on value of parameter. (Is this subset pivotality?)

Marginally vs. conditionally: See simulation writeups.

Bootstrapping brainteaser: Observations really are independent in FCR resamples. I think the previous issues with this were actually not due to failed assumptions. https://stats.stackexchange.com/questions/339237/are-observations-independent-in-bootstrapped-resamples

Our two metrics

Metric #1: 95% CI for number of hits above expected value that we see in observed data (resample under original distribution, using first half of my theory)

Metric #2: 95% CI under the null for number of hits (resample under null, using second half of my theory)

What about other link functions or hypothesis tests not based on regression? How would we do residual resampling? Davison & Hinkley talk about other GLMs. I think we'd end up with either some form of residual or Y itself for each outcome, and then we'd jointly resample these things as planned.

Possible resampling algorithms

Conclusion: Since we are bootstrapping and then fitting parametric model (OLS), I think we should use Fox's or Westfall's residual resampling approach (and indeed, this seemed to work in the last simulation). Full-case resampling violates the iid error assumption in OLS regression, meaning the probability of rejecting the null within each resample is not α. Our application of bootstrapping is unusual in that we need to resample in a way that is compatible with parametric assumptions of the test we're going to apply for each outcome.

Note that all of these methods assume $Var(Y^*|X^* = 0) = Var(Y^*|X^* = 1)$; in the resampled data, this is the average of the original within-stratum variances. I think this is fine since we are *not* avoiding parametric inference for OLS (unlike in other bootstrapping applications).

Basic "fix X, then resample Y"

1. Fix X and resample only Y: $(Y_i^*, X_i^*) = (\text{sample}(Y_i), X_i)$. So Y_i^* ⫫ X_i^*.
2. Regress: $Y_i^* = \beta_0 + \beta X_i^* + \epsilon$.
3. Test $H_0 : \beta = 0$.

Has multimodality problem. No guarantee that residuals are normal, so OLS inference will be wrong anytime the original data are generated under H_A.

Figure 1: Original data.

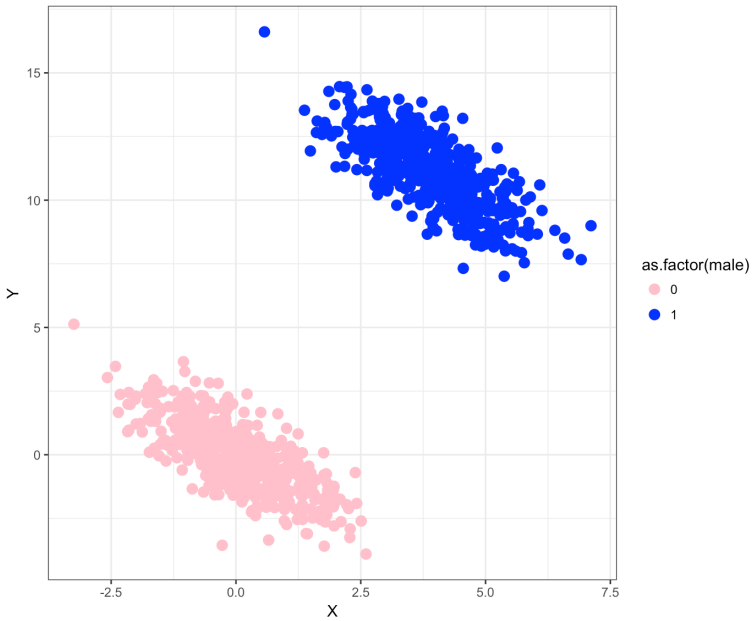
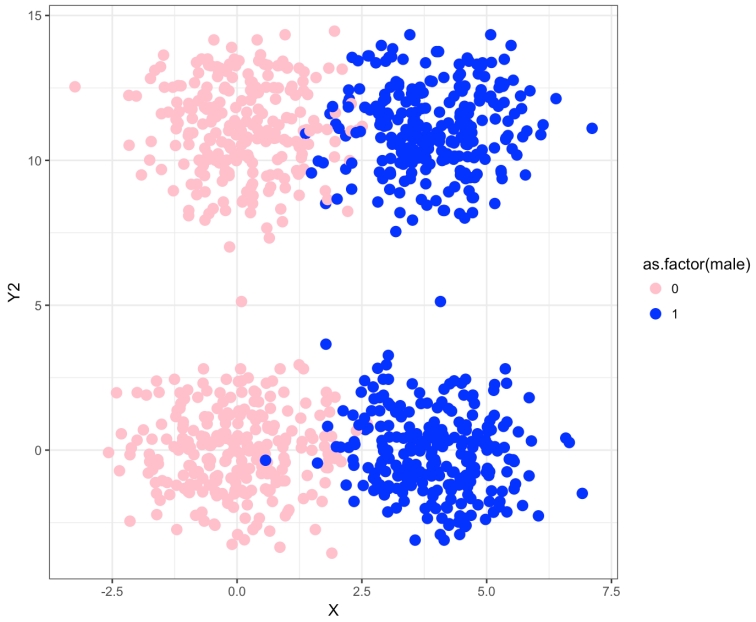


Figure 2: Resampled data under basic algorithm. Loses association between sex and Y, so regression residuals are bimodal.

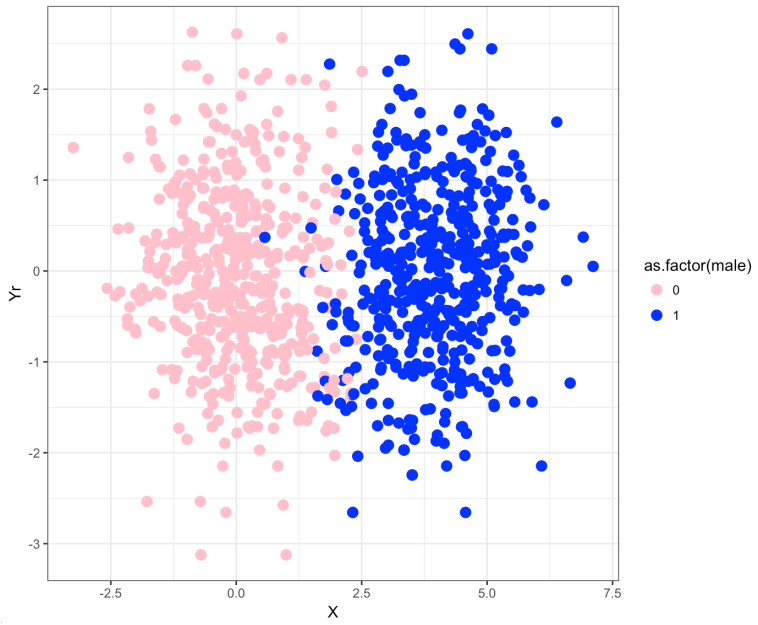


Westfall's "center, then resample Y"

1. Fix X and set Y_i^* = sample $\left(Y - \widehat{Y}\right)$. So Y_i^* ⫫ X_i^*.
2. Regress: $Y_i^* = \beta_0 + \beta X_i^* + \epsilon$.
3. Test $H_0 : \beta = 0$.

Avoids multimodality problem, but could still have non-normal residual distribution, as in the above. So OLS inference could be wrong, but only if original data already violate homoskedasticity. Works well in my simulations.

Figure 3: Resampled data under Westfall's algorithm. Residuals are normal.



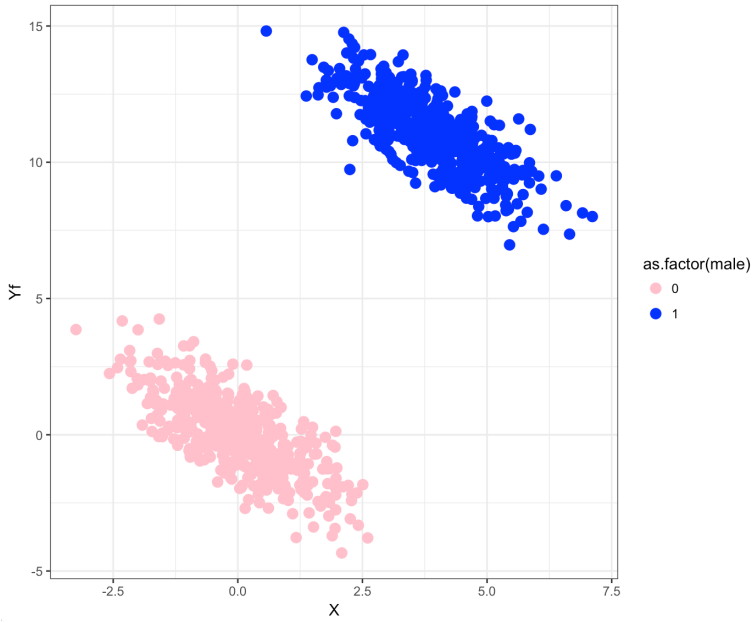
Fox/Davison's residual resampling

See Davison Algorithm 6.1 in "Linear Regression" chapter.

1. Fix X and set $Y_i^* = \widehat{Y}_i +$ sample $\left(Y - \widehat{Y}\right)$. So $Y_i^* \not\perp\!\!\!\perp X_i^*$.
2. Regress: $Y_i^* = \beta_0 + \beta X_i + \epsilon$.
3. Test $H_0 : \beta = \widehat{\beta}$ (i.e., the fitted coefficient from original sample).

Avoids multimodality problem residuals will be normal as long as they are normal within strata of X, so OLS inference should be correct. $Var(Y^*|X^* = 0) = Var(Y^*|X^* = 0)$ and is an average of the original conditional variances.

Figure 4: Resampled data under Fox's algorithm. Residuals are normal.



Other online resources

⋆ CV: Permutation tests vs. bootstrap tests

What we are doing is similar to a permutation test because we're not centering the observations. In this answer, Snow says "permutation tests test a specific null hypothesis of exchangeability, i.e. that only the random sampling/randomization explains the difference seen". Also, permutation tests are more powerful.

"Resample Y while fixing X" strategy is "similar in spirit" to permutation tests and assumes that the distribution of Y in each group of X is identical. In contrast, "mean-center, then resample within each group of X" does not make this assumption.

https://stats.stackexchange.com/questions/20217/bootstrap-vs-permutation-hypotheis-testing

⋆ CV: Center first or resample first?

https://stats.stackexchange.com/questions/136661/using-bootstrap-under-h0-to-perform-a-test-for-the-difference-of-two-r
187630

Goes over difference between "resample, then center" (OP's first bootstrap) and "center, then resample" (OP's second bootstrap).

Accepted answer says that first approach is actually testing whether the distribution of x and y are identical.
I DON'T REALLY GET THIS.

CV: Cases when naive bootstrap fails

https://stats.stackexchange.com/questions/9664/what-are-examples-where-a-naive-bootstrap-fails/9722#9722

CV: Applying bootstrap to arbitrary smooth function of the data

https://stats.stackexchange.com/questions/246632/smoothness-of-a-statistic-for-bootstrapping

Notes on past literature

Other "global tests" (i.e., of joint null)

- Fisher's Combination Test aggregates all the p-values instead of just using the smallest one; better than Bonferroni joint test if many small effects, but worse if there are a few large effects. Assumes independent tests.

Frane (per-family error rates)

- Per-family error rate (PFER): Number of expected false positives
- Controlling PFER implies controlling FWER (so PFER is more stringent)
- Bonferroni also controls PFER to α, which is partly why it's more "conservative"

Talks about *another interesting interpretation of Bonferroni*, which is that (unlike other FWER-control procedures), it also controls the more stringent "per-family error rate" (PFER), which is the expected number of false positives.

With other FWER procedures that don't control PFER, which is most of them, there is a <5% probability of at least one false positive, but that could be 1 false positive or it could be 100. With Bonferroni, if there are false positives (<5% chance), it's also the case that the expected number of them is $< 0.05\times$ (# of tests) with <5% probability.

So to build upon Tyler's earlier statement, I believe Bonferroni lets us say:

"*We did 100 tests and rejected 20. With >95% confidence, all 20 are true effects (FWER). Even if we are in the 5% of instances where we actually do have false positives, there aren't very many of them: only 5 or fewer with 95% confidence (PFER).*"

Yekutieli & Benjamini

A parametric bootstrap correction for LMM.

Yekutieli & Benjamini

A resampling-based procedure for FDR control. Very readable!

- Pg. 7: Outlines residual resampling under H_0, exactly as we will be doing
- Talks a lot about Westfall and also uses subset pivotality
- Pg 7 (subset pivotality): When we resample under the strong null, the distribution of p-values for the null hypotheses that are actually *false* is different from what it is in real life. That's because we are resampling such that *all* nulls are true, whereas in real life some might be false. The goal is to use the resampled distribution of p-values *for the nulls that are true* to approximate its true distribution. This only holds if the distribution of that subset of p-values isn't affected by the truth or falsehood of the remaining hypotheses.
- Seems to be very closely related to Westfall, but cares about FDR instead of FWER.

van der Laan, Dudoit, Pollard (2004) - not saved

A resampling-based multiple testing control procedure.

- $k-$FWER control procedure where you first reject hypotheses based on a procedure that controls regular FWER, but then also "augment" by rejecting an additional $k = 1$ hypotheses (summarized by Romano & Wolf, pg 1396)

Romano & Wolf

A resampling-based multiple testing control procedure for FWER, k-FWER, or FDP.

- $k-$FWER control: With 95% confidence, the number of false positives is $< k$
- FDP control: With 95% confidence, the proportion of rejections that are false positives is $< \gamma$
- FDR control: with 95% confidence, the EXPECTED proportion of rejections that are false positives is $< \gamma$
- They bootstrap under the original data distribution and use this to get quantiles of the distribution of the k^{th}-largest test statistic for various subsets of the test statistics, allowing control of the $k-$FWER through a step-down procedure.
- Why resample not under the null? Apparently resampling not under the null avoids subset pivotality (pg 1388). Resampling under H_A "exploits the duality between CIs and hypothesis tests" (pg 1388). Even though we're resampling under H_A, the *critical value* is still under the null distribution. (See

Example 2.1, where the critical value is clearly for the distribution when $\mu = 0$. Also note that the A subscript on the critical values, for example on page 1384, indicates that it's the critical value when all s nulls are true.) Thus, I think he must be centering the bootstrapped test stats by the observed one in order to get the null distribution of the critical values. This is corroborated by an email exchange I had with author of StepwiseTest package (see file "Romano Wolf StepwiseTest package question email thread").

- Page 1396: Talks about other resampling methods in some detail, including Westfall
- Does not use subset pivotality, unlike Westfall

Davison & Hinkley text

- Linear regression chapter has great explanation of case resampling (Tyler's approach) vs. residual resampling (Fox, sort of Westfall).
- Pg 12 in "Linear Regression" chapter: Says that with case resampling, "the null hypothesis being tested is stronger than just zero slope"

Good (2005) book chapter

- Clear explanation of difference in assumptions between nonparametric bootstrap (equal parameters) and permutation test (exchangeability).
- I think the "resample, then center" strategy is closer to a permutation test since it forces equal distributions, while the "center, then resample" or "just resample" strategies are more like bootstrapping.

Fox chapter on bootstrapping regressions

- In section on bootstrap hypothesis testing, uses this residual-resampling algorithm:
  1. Fix X and set $Y_i^* = \widehat{Y}_i + \text{sample}\left(Y - \widehat{Y}\right)$. So $Y_i^* \not\perp\!\!\!\perp X_i^*$.
  2. Regress: $Y_i^* = \beta_0 + \beta X_i + \epsilon$.
  3. Test $H_0 : \beta = \widehat{\beta}$ (i.e., the fitted coefficient from original sample).

Hall & Wilson "Two guidelines"

- "Sample, then center" algorithm (here, I am extrapolating what they present to the regression setting):
  1. Resample entire vector $(Y_i^*, X_i^*) = \text{sample}\left((Y_i, X_i)\right)$. So $Y_i^* \not\perp\!\!\!\perp X_i^*$.
  2. Regress: $Y_i^* = \beta_0 + \beta X_i^* + \epsilon$.
  3. Test $H_0 : \beta = \widehat{\beta}$ (i.e., the fitted coefficient from original sample).
- 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.

⋆ Troendle 2004 "Slow convergence"

- Talks about multiple 2-sample t-tests
- Two approaches (see page 3 for very clear overview):
  1. Permute a subject's outcome variables while assigning group labels (with *uncentered* variables)
  2. Bootstrap by resampling a subject's entire row (with *centered* outcome variables)
- "Center, then sample" bootstrapping algorithm (here, I am extrapolating what they present to the regression setting):
  1. Center outcome by setting $\widetilde{Y}_i = Y_i - \widehat{Y}$
  2. Resample entire vector: $(Y_i^*, X_i^*) = \text{sample}\left((\widetilde{Y}_i, X_i)\right)$. So $Y_i^* \amalg X_i^*$.
  3. Regress: $Y_i^* = \beta_0 + \beta X_i^* + \epsilon$.
  4. Test $H_0 : \beta = 0$.

Westfall & Young textbook

Resampling-based procedures for FWER control.

- "Center, then resample" strategy
  1. Center outcome by setting $\widetilde{Y}_i = Y_i - \widehat{Y}$

2. Fix X and resample only Y: $(Y_i^*, X_i^*) = \left(\text{sample}(\widetilde{Y}_i), X_i\right)$. So $Y_i^* \amalg X_i^*$.
3. Regress: $Y_i^* = \beta_0 + \beta X_i^* + \epsilon$.
4. Test $H_0 : \beta = 0$.

- Single-step adjusted p-value, aka "minP" method (page 46, 51): For each test j, set its adjusted p-value equal to the probability in the resamples (generated under the strong null) that the minimum p-value is less than the observed p-value for test j. So, it's like a p-value of the p-value (probability of observing a p-value at this small under the strong null)
- Free step-down resampling method (pg 66): Not very intuitive...
- Strong FWER control
- 106-109: More on OLS
- Page 79: "Resampling from the residuals, rather than the actual data values, is important. If the actual data are pooled and resampled, then the underlying residual distribution will be estimated using a multi-modal empirical distribution function when the means differ. Such resampling violates the first principle of Hall & Wilson (1991): '...even if the data might be drawn in a way that fails to satisfy H_0, resampling should be done in a way that reflects H_0'."
- Page 39-41: Analyzes $P(P^* < \alpha)$ in bootstrap iterates
- Page 92-93: Re-randomization
- Bootstraps of pivotal statistics converge faster (Section 2.2.2)

Westfall & Troendle (2008)

- Subset pivotality assumption: Take any subset of the hypotheses to be tested. The distribution of the max test statistic within this group, given that all the nulls in that subset hold, is the same as its distribution given the strong null (all the nulls for ALL the hypothesis tests hold). In other words, once you know that all nulls within the subset are true, knowing that all the other nulls are true doesn't change anything about distribution of max test statistic in that subset.
- Reason for using subset pivotality: To avoid having to resample under every intersection null (i.e., subsets of nulls being true) so that we can just resample under the joint null.

- 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).
- Shows that "permutation methods are distribution-free under an appropriate exchangeability assumption and FWER follows mathematically, regardless of sample size"
- Implemented in SAS' PROC MULTTEST
- According to SAS' documentation, this(?) paper shows that: "when subset pivotality holds, the joint distribution of p-values under the subset is identical to that under the complete null"

Bretz, Hothorn, & Westfall (2008)'s 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)"

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)

Resolved questions

What does Westfall do for logistic regression?

Westfall's algorithm (pg 215) doesn't seem to work. He just regresses each outcome on the Xs and then generates from a binomial using the fitted parameters. But this doesn't preserve the correlation between the Ys unless you condition on all covariates such that the Ys are independent (see `westfall_logistic_regression.R`).

Should we compare our joint test to joint tests based on non-FWER procedures?

If we reject joint null based on FWER control method, then we can conclude with 95% confidence that this rejection is real, which implies that the joint null is false.

With k-FWER control, we can conclude with 95% confidence that there are fewer than k false positives, so $R - k + 1$ of the observed rejections are real (don't know which ones), where R is the number of adjusted rejections. So, as long as $R > k - 1$, we can reject the joint null with 95% confidence. PROBLEM: WOULD HAVE TO TRY EVERY k??

With FDP control, we can reject the joint null if the proportion of adjusted rejections exceeds the allowed amount (γ). Again, problem because we'd have to try every γ?

FDR does not work for a joint test because it uses expected proportion of rejections.

The problem is that k-FWER and FDP would probably lead to more powerful joint test than the FWER control method, but not clear which parameters to choose. So maybe we should not worry about them?

What we can learn from existing FWER procedures?

Say we observe 18 rejections at the 0.05 level. Then, with increasing stringency, we can say:

- Our null CI is $[0, 10]$. "We observed $18 - 10 = 8$ more hits than we would expect in 95% of samples taken under the strong null (every null holds)."
- $8-$FWER controlling procedure (Romano) rejects 11. "There are fewer than 8 false positives, so of these 14 rejections, at least $14 - 7 = 7$ are real effects, but we don't know which ones." (How should we decide which k to use here?)

- FWER-controlling procedure (Bonferroni or preferably Romano/Westfall/another resampling method) rejects 3. "All 3 of these rejections are real effects with 95% confidence", or alternatively: "We reject the strong null with >95% confidence."

In practice, you could get the first two from the same bootstrap resamples (taken parametrically under the strong null). Then you'd have to resample under the original distribution to get the third one.

Why do datasets created through full-case resampling reject the null too often?

See my Cross Validated question (https://stats.stackexchange.com/questions/323455/why-do-hypothesis-tests-on-resample
There are two separate issues:

1. You have to center the bootstrap statistics by the estimated statistic in the original data (so do the hypothesis test on $\widehat{\beta}^{(j)} - \widehat{\beta}$, not on $\widehat{\beta}^{(j)}$ itself). This is because of the bootstrap analogy principle.
2. With FCR, we violate the OLS assumption of iid errors because every repeated pair (X, Y) has exactly the same residual, so the residuals are perfectly associated with X. If we use Westfall's *null* resampling approach where we fix X and resample residuals, then even though there are repeated observations, the *residuals* of the model newly fit to the resampled data are still iid because we are breaking any association between the resampled residuals and X. This issue only matters if you need to meet parametric assumptions within each resample. Our application of bootstrapping is unusual in that we need to resample in a way that is compatible with parametric assumptions of the test we're going to apply for each outcome. This is why, e.g., Davison & Hinkley do specifically recommend FCR for OLS.

It turns out that in the OLS example I tried, only issue #1 actually matter because fixing that issue led to nominal rejections. However, issue #2 seems to still be a real concern and might matter more in models less robust to assumption violations than OLS.

Why does Blakesley/Westfall's minP bootstrapping work since it is an extreme order statistic?

Probably because minP is not an order statistic of the data themselves.

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?

Because there are only a fixed number of p-values (k), and this is not necessarily asymptotically large.

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.

Exact variance under null

Let p_w be the w^{th} observed p-value. Then, under the strong null, we have:

$$\text{Var}\left(\widehat{\theta}\right) = W\alpha\left(1-\alpha\right) + 2\sum_{1\leq i<j\leq W} P\left(p_i < \alpha, p_j < \alpha\right) - \alpha^2$$

$$\begin{aligned}
\text{Var}\left(\widehat{\theta}\right) &= \text{Var}\left(\sum_{w=1}^{W} 1\{p_w < \alpha\}\right) \\
&= \sum_{w=1}^{W} \text{Var}\left(1\{p_w < \alpha\}\right) + 2\sum_{1\leq i<j\leq W} \text{Cov}\left(1\{p_i < \alpha\}, 1\{p_j < \alpha\}\right) \\
&= W\alpha\left(1-\alpha\right) + 2\sum_{1\leq i<j\leq W} E\big[1\{p_i < \alpha, p_j < \alpha\}\big] - E\big[1\{p_i < \alpha\}\big]E\big[1\{p_j < \alpha\}\big] \\
&= W\alpha\left(1-\alpha\right) + 2\sum_{1\leq i<j\leq W} \Big[\underbrace{P\left(p_i < \alpha, p_j < \alpha\right)}_{=\alpha^2 \text{ under independence}} - \alpha^2\Big]
\end{aligned}$$

Simulation notes

2018-6-6 *more scenarios

Trying more scenarios to have smaller proportion of effects but still with observed rejections way outside CI.

- $\rho_{XY} = 0.25$ with 20% correlated: Has 100% power, so need to go down.
- $\rho_{XY} = 0.25$ with 5% correlated:

2018-5-18 *all scenarios Freedman resampling

These are in the paper.

"Marginally vs conditionally"

Goal was to test the hypothesis that if outcomes are independent conditional on *adjusted* covariate C, then the test stats for X are independent even though the outcomes are marginally correlated. However, if the outcomes are independent only conditional on *unadjusted* U (as in below simulations), then the test stats for X are correlated.

"Validate FCR-center resampling"

Since realizing that MICE algorithm wouldn't work, I tried to validate FCR-center (i.e., do FCR, but then center each resampled test statistic by its counterpart in the original data) by generating data with 1 standard normal covariate and 1 unmeasured non-confounder that induces correlation between Y_1 (normal) and Y_2 (either normal or binary depending on scenario). The empirical (Sherlock) results for both scenarios are in `stitched.csv`.

The simulation had two separate stages:

1. Locally, use `fcr_center_validation.R` to actually generate data under H_0 (i.e., override the beta for X and set it to 0) for $10,000$ simulation reps and check the correlation of the p-values. This gives us the "true" correlations under H_0.
2. On Sherlock, use `fcr_center_validation_for_sherlock.R` to generate, for each of $2,500$ simulation reps per scenario, an original dataset under H_A and then use FCR-resample to produce $B = 2,000$

resamples under H_0. Fit the analysis model to each resample. Compute the average correlation of the p-values for each simulation rep.

Results: When both outcomes are normal, FCR-center exactly recovers the correct p-value correlation (true correlation: 0.70, empirical correlation: 0.70). However, when 1 outcome is normal and 1 is binary, FCR-center does not work (true correlation: 0.31, empirical correlation: 0.20).

Explanation: I think this is because CI and hypothesis test are directly equivalent for OLS, but not for logistic regression. That's because for OLS, the SE is independent of the value of the parameter β itself: it's just $\sigma^2(X'X)^{-1}$, and σ^2 is independent of β for any fixed model specification. In contrast, with logistic regression, the SE depends on β, which is why its MLEs require iteration. Basically this is the same reason that the SE for a proportion is different when computed under H_0 vs. H_A, so the CI and test might not coincide. I did a separate simulation to check this in which I avoided any bootstrapping. I just directly generated data under H_0 or H_A using `fcr_center_validation.R` and looked at whether the "true" SE of the coefficient estimates was the same for H_0 and H_A. Below, the 0.06 and 0.08 show the difference in SEs under H_0 and H_A for a binary outcome.

IS THIS THE FABLED SUBSET PIVOTALITY?

Y2 type	H_0	$SE(\widehat{\beta}_{Y_1})$	$SE(\widehat{\beta}_{Y_2})$
Binary	T	0.10	0.06
Binary	F	0.10	0.08
Normal	T	0.10	0.10
Normal	F	0.10	0.10

2018-4-7 unsaved local simulations using `mice_validation.R`

"2 normal outcomes":

Critical: If Y1 and Y2 are correlated because of U, but U is adjusted in the models for Y1 and Y2, then the test stats and p-values AREN'T correlated.

If U is NOT adjusted in analysis (implicit in our current simulations), then the test stats and p-values ARE correlated.

2018-4-3

Reran all methods to compare. This is written up in the file "2018-4-3 Multiple Outcomes update #1".

2018-3-16 ∗real Romano

Resampling parametrically under original data distribution (`ha.resid`) and centering bootstrapped test stats by observed one in order to apply Romano.

"2018-3-15 add Romano-H0"

This is the same as the previous one, except adding Holm and a version of Romano that still uses *null* resamples (`h0.parametric`). Note that this is not actually what he recommends in his paper (he resamples from the original data distribution, but then presumably centers the bootstrapped test statistics by the observed one).

For paper, could combine these results (minus Romano) with previous one in order to have 1000 simulations per method (except Holm because it wasn't in the previous simulations).

"2018-3-18 ∗with other methods"

We shelved the CI under the alternative because of the below issues (couldn't find a bootstrapping/resampling method that worked under H_A). In the present simulations, we compared our joint test of the null (using number of rejections) to other methods (Westfall's minP, Westfall's step-down, Bonferroni).

- **Realized that regenerating residuals (method `h0.parametric`) actually doesn't work because it loses the correlated structure: we were just generating separate Ys. Empirically, saw that the null CIs were the same regardless of `rho.YY`.**
- So these simulations used Westfall's reattach-residuals method (method `resid`), and all results made a lot of sense.
- CI plots shows the averaged CI limits and mean bootstrapped rejections by scenario.
- Seemed weird that both Westfall methods lead to exactly the same joint test, but this is okay. It's because the smallest p-values are the ones that change the least, so few p-values shift from being

above to below 0.05 when changing from minP to step-down (see folder "Compare Westfall minP vs step-down").

2018-1-23 unsaved local sims

Revisiting residual resampling for CI: Fix design matrix, then re-generate residuals using $\widehat{\beta}$ and $\widehat{\sigma^2}$. (Maybe resampling the residuals would also work, but should we be worried about non-independence?)

- Tried a single outcome and different correlation strengths
- Compared average rejection probability in bootstraps to average in original data
- Need large n for this! Using $n = 1000$ didn't work because regression estimates aren't precise enough. Using $n = 10,000$ seemed okay.
- Not yet assessing CIs for number of rejections because would need multiple outcomes for that

Went back to debug FCR script and found that reattaching residuals works under null.

Under alternative, though, got 25% rejections across all bootstrap resamples vs. 15% across 1000 sim reps in original datasets.

Still got 25% even when generating new residuals and adding them to the original fitted values!!!!!! Whyyyyyyyyyy???????????

2018-1-22

- Reduced number of outcomes to 30 to reduce computation time
- Now also computing CIs based on raw percentiles (below) scaled by the estimated rate
- And am trying with data generated under alternative
- Still using uncorrelated outcomes
- Rates are again almost perfectly estimated
- We can't use subsampling when data are generated under H_A because the number of rejections will be lower due to lower power. So the sampling distribution isn't just wider in the smaller samples; it doesn't even have the same expected value.

2018-1-20 round 2 subsampling

- Running same as below, but with the real code instead of raw quantiles. "covers.correct" uses exactly rate = 0 as below; "covers" uses the rate estimate.
- Rates basically always estimated as 0 now because of larger B
- Coverage now 80
- $*$Conclusion: $B = 20,000$ instead of $B = 2,000$ results in exactly correct rate estimation, at least for data generated under H_0. Can probably find middle ground of simulation reps.
- Based on local simulations (debug_subsampling), seems like a CI based just on percentiles of the subsamples is better than the one in the lab handout, where it's centered on the point estimate.

One-sided CI idea to improve coverage a little

2018-1-20 subsampling

- Based on the exact variance I derived, when the null is true, the SE of $\widehat{\theta}$ doesn't depend on n because tests maintain α probability of rejecting regardless of n. So in this case, the rate is exactly 0 (as correctly estimated by the last round of simulations).
- So these simulations simplify things by not estimating the rate, but just taking the quantiles of the iterates for the third sample size ($n = 900$). Seems like this should work fine.
- $*$Giving it the exact rate yields 100
- Sample sizes same as below
- Rep time 90th percentile: 1700 seconds with $B = 2000$

2018-1-19 - results not saved

- Same as 2018-1-18, but increased all sample sizes by order of magnitude
- Mean rate exactly 0 with min/max $pm0.10$ and 25th, 75th percentiles ± 0.015
- Still 70
- Used $n = 100,000$ and subsample sizes $n_s = (400, 600, 900, 1350)$ to match lab handout, but I have $B = 2,000$ instead of $B = 20,000$

2018-1-18 - results not saved

- Trying subsampling with rate estimation in a null scenario and no correlation between outcomes
- Only getting 70
- Used $n = 10,000$ and subsample sizes $n_s = (40, 60, 90, 135)$ to match lab handout, but I have $B = 2,000$ instead of $B = 20,000$
- Each sbatch took 3.5 hours to run (upper limit)

2018-1-16 "under H_0 and FCR"

- Joint test plot shows power of our joint test (i.e., resample residuals under H_0 and see if observed hits are extreme in this distribution) and compares to power of a naïve Bonferroni joint test (i.e., reject the joint null if any of the 100 Bonferroni-adjusted tests rejects).
- Above results make sense: e.g., we are near the nominal joint alpha level (0.05 regardless of alpha level for individual tests) when data were generated under joint null.
- Excess hits plot shows average excess hits (hits above expectation) and CIs averaged across bootstraps. The CIs are sometimes completely above the observed excess hits. **This is because the FCR resampling approach is wrong.** See the resolved questions.

2017-8-1 (generate under H_0 and H_A)

- FCR resampling doesn't seem to work because even when data generated under the null, the resamples have more than the expected number of rejections
- Not due to lower truncation because using alpha = 0.5 doesn't help
- Seems like correlations are preserved in individual resamples?
- But I think the correlations are more variable, even if their mean is still 0, because sometimes we draw samples that have a lot of repetition

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 (generate 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)
- Figure out why file 125 has extra rows (try running it again)
- **The apparently too-conservative behavior for the uncorrelated case is a benign artifact** of $\widehat{\theta}$'s discreteness.
  - 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 $\widehat{\theta}$ 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 $\widehat{\theta}$ 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 could benchmark against the truth since $\widehat{\theta}$ 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: N=10,000, B=10,000
- Resampling under joint null (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: seems too small based on the below
- 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

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** (see 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 because SD is treated as known ⇒ **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.

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.

For earlier simulation results, see old project log in Word.

Notes on Ying's MIDUS paper

Variables

- Parental warmth (X): average of separate maternal and paternal warmth scales (so $\in [1, 4]$)
- Flourishing (main Y)
  - Continuous version: Sum of standardized subscales for emotional, psychological, and social well-being
  - Binary version: Are you in top tertile on enough subscales?
  - Count version: For how many subscales are you in top tertile? ($\in [0, 13]$?)
- Subdomains comprising flourishing: 13 continuous measures (T3)
- Bad outcomes (secondary Ys): 7 mental health problems and bad behaviors, all binary (T4)

Analyses

- Continuous flourishing and other continuous Ys: Normal GEE (cluster by siblings)
- Binary flourishing: Poisson GEE with log link (risk ratios) (Did log-binomial not converge or other reason for choosing this?)
- Count flourishing: Normal GEE (Why not, e.g., Poisson or NB?)
- Bad binary outcomes: Poisson GEE for non-rare (RRs) or logistic regression for common (ORs)

Our analyses

- Multiple imputation for main analyses?
- Randomly choose 1 sibling from each sibship to avoid clustering? Then just use OLS and regular log-linear model. Or could do mixed model, but then parametric bootstrap is more complicated.
- Maybe treat outcomes as: 3 different flourishing variables + 7 binary bad outcomes? Also addresses model specification angle.
-

References